CAMILA ASSIS

PREVISÃO DO TEOR DE LIGNINA EM CANA-DE-AÇÚCAR USANDO ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E MÉTODOS QUIMIOMÉTRICOS

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

VIÇOSA
MINAS GERAIS – BRASIL
2014

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
A848p
2014

Assis, Camila, 1988-
Previsão do teor de lignina em cana-de-açúcar usando espectroscopia no infravermelho próximo e métodos quimiométricos / Camila Assis. – Viçosa, MG, 2014.
xiv, 72f. : il. (algumas color.) ; 29 cm.

Orientador: Reinaldo Francisco Teófilo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.65-72.

1. Cana-de-açúcar. 2. Biomassa. 3. Lignina.
4. Espectroscopia de infravermelho. I. Universidade Federal de Viçosa. Departamento de Química. Programa de Pós-graduação em Agroquímica. II. Título.

CDD 22. ed. 633.61

CAMILA ASSIS

PREVISÃO DO TEOR DE LIGNINA EM CANA-DE-AÇÚCAR USANDO ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E MÉTODOS QUIMIOMÉTRICOS

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

APROVADA: 05 de abril de 2014.

Prof. Márcio Henrique Pereira Barbosa (Coorientador)	Prof. Jorge Luiz Colodette (Coorientador)
Prof. Efraim Lázaro Reis	Prof. Reinaldo Francisco Teófilo (Orientador)

Aos meus pais

Mário Lúcio de Assis e

Emerentina Pacheco de Assis

"Eu acredito no cristianismo como acredito que o sol nasce todo dia. Não apenas porque o vejo, mas porque através dele eu vejo tudo ao meu redor."

(C.S.Lewis)

AGRADECIMENTOS

Agradeço a Deus, amigo fiel, minha pérola de grande valor. "Porque dEle e por Ele, e para Ele, são todas as coisas; glória, pois, a Ele eternamente." Rm 11:36.

Aos meus pais, por toda ajuda, amor, conselhos e por, principalmente, serem meus melhores amigos. Vocês sempre me surpreendem!

Às minhas irmãs, cunhados e à Elisa. Obrigada pelo apoio, amor, risadas, amizade. Vocês enchem minha vida de alegria!

Ao Prof. Reinaldo F. Teófilo, pela orientação sempre presente. Obrigada pela confiança, amizade, paciência e por, principalmente, estar sempre disposto a ensinar e corrigir. Todos esses anos de aprendizado foram muito valiosos!

Ao Prof. Márcio Henrique Pereira Barbosa, por ter disponibilizado toda a estrutura física para que o trabalho fosse realizado. Obrigada por todo ensino e pela confiança.

Aos amigos do LBMV (Laboratório de Biotecnologia e Melhoramento Vegetal), em especial às amigas Lidiane, Karla, Rachel e Telma. Obrigada por sempre estarem dispostas a ajudar, sem medir esforços.

Aos amigos do CECA (Centro Experimental de Cana-de-açúcar), em especial ao querido amigo Volmir Kist.

À Universidade Federal de Viçosa, em especial o Departamento de Química, pela estrutura para poder realizar este curso e este trabalho.

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pela concessão da bolsa de estudo.

A todos aqueles que contribuíram direta ou indiretamente para a realização deste trabalho.

BIOGRAFIA

CAMILA ASSIS, filha de Mário Lúcio de Assis e Emerentina Pacheco de Assis, nasceu na cidade de Manhuaçu, estado de Minas Gerais, em 16 de janeiro de 1988.

Iniciou o curso de Química em maio de 2006 pela Universidade Federal de Viçosa (UFV), em Viçosa, MG, diplomando-se em bacharelado e licenciatura em janeiro 2012. No mesmo ano, no mês de agosto, iniciou o curso de pós-graduação em Agroquímica, com área de concentração em Química Analítica, em nível de Mestrado, na mesma instituição, submetendo-se à defesa de dissertação em 14 de Março de 2014.

SUMÁRIO

LISTA DE SÍMBOLOS

Abreviaturas	Termos em Inglês	Termos em Português
ASTM	American Society for Testing Materials	Sociedade Americana para Testar Materiais
bi PLS	Backward Interval Partial Least Squares	Mínimos Quadrados Parciais por Exclusão
CV	Cross Validation	Validação Cruzada
G	Guaiacyl	Guaiacil
GA	Genetic Algorithm	Algoritmo Genético
h	Number of Variable Latents	Número de Variáveis Latentes
HMF	Hhydroxymethylfurfural	Hidroximetilfurfural
i PLS	Interval Partial Least Squares	Mínimos Quadrados Parciais por Intervalo
LA	Levulinic Acid	Ácido Levulínico
LOD	Limit of Detection	Limite de Detecção
LOQ	Limit of Quantification	Limite de Quantificação
MLR	Multiple Linear Regression	Regressão Linear Múltipla
NIR	Near Infrared	Infravermelho Próximo
NREL	National Renewable Energy Laboratory	Laboratório Nacional de Energia Renovável
OPS	Ordered Predictors Selection	Seleção dos Preditores Ordenados
PCR	Principal Component Regression	Regressão por componentes Principais
PLS	Partial Least Squares	Regressão por Quadrados Mínimos Principais
R	Correlation Coefficient	Coeficiente de Correlação
RMSE	Root Mean Square ERROR	Raiz Quadrática Média do Erro
RMSECV	Root Mean Square Error of Cross Validation	Raiz Quadrática Média do Erro de Validação Cruzada
RMSEP	Root Mean Standard	Raiz Quadrada do Erro Médio de Previsão
RPD	Residual Prediction Deviation	Relação de Desempenho de Desvio
S	Syringyl	Siringil
SDV	Standard Deviation of Validation	Desvio Padrão dos Erros de Validação
SEL	Selectivity	Seletividade
SEM	Sensibility	Sensibilidade

.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

ASSIS, Camila. Universidade Federal de Viçosa, abril de 2014. **PREVISÃO DO TEOR DE LIGINA EM CANA-DE-AÇÚCAR USANDO ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO E MÉTODOS QUIMIOMÉTRICOS.** Orientador: Reinaldo Francisco Teófilo. Coorientadores: Márcio Henrique Pereira Barbosa e Jorge Luiz Colodette.

A construção de modelos de calibração multivariada usando espectroscopia de refletância na região do infravermelho próximo (NIR) e regressão por quadrados mínimos principais (PLS) para estimar teores de lignina de uma série de genótipos de cana-de-açúcar é o objetivo deste trabalho. Análises laboratoriais foram realizadas para determinar os valores de lignina, utilizando o método Klason. As variáveis independentes foram obtidas a partir de diferentes materiais: bagaço seco, bagaço seco com caldo, folha e colmo. Os espectros NIR foram obtidos na faixa de 10000 a 4000 cm^{-1}. O algoritmo Kennard-Stone foi utilizado para selecionar o conjunto calibração e previsão. Os modelos foram construídos empregando a regressão por quadrados mínimos parciais (PLS) e diferentes algoritmos para seleção de variáveis foram testados: iPLS, biPLS, algoritmo Genético (GA) e o método de seleção dos preditores ordenados (OPS). Para o bagaço seco, o melhor modelo foi obtido após seleção de 445 variáveis com o OPS, que obteve RMSEP de 0,85, Rp de 0,97, RPD de 2,87 e erro relativo médio na previsão de 2,82%; para o bagaço seco com caldo o melhor modelo foi obtido após seleção de 265 variáveis com o OPS, que obteve RMSEP de 0,65, Rp de 0,94, RPD de 2,77 e erro relativo médio na previsão de 1,94%; para a folha o melhor modelo foi obtido após seleção de 305 variáveis com o OPS, que obteve RMSEP de 0,58, Rp de 0,96, RPD de 2,56 e erro relativo médio na previsão de 2,47%; para o terço médio do colmo o melhor modelo foi obtido após seleção de 205 variáveis com o OPS, que obteve RMSEP de 0,61, Rp de 0,95, RPD de 3,24 e erro relativo médio na previsão de 1,97%; para o terço superior do colmo o melhor modelo foi obtido após seleção de 300 variáveis com o OPS, que obteve RMSEP de 0,58, Rp de 0,96, RPD de 2,34 e erro relativo médio na previsão de 1,94%; para as partes superiores, inferiores e médias do colmo, o melhor modelo foi obtido após seleção de 250 variáveis com o OPS, que obteve RMSEP de 0,80, Rp de 0,99, RPD de 2,79 e erro relativo médio na previsão de 2,90%. O algoritmo OPS selecionou um menor número de variáveis com maior capacidade preditiva.Todos os modelos mostraram-se confiáveis, com alta exatidão para previsão da lignina em cana-de-açúcar, reduzindo significativamente o tempo para realização das análises e portanto, otimizando o processo como um todo.

ABSTRACT

ASSIS, Camila. Universidade Federal de Viçosa, april 2014. **PREDICTION OF LIGNIN CONTENT IN SUGAR CANE USING NEAR INFRARED SPECTROSCOPY AND CHEMOMETRICS METHODS**. Advisor: Reinaldo Francisco Teófilo. Co-Advisors: Márcio Henrique Pereira Barbosa and Jorge Luiz Colodette.

The building of multivariate calibration models using near infrared spectroscopy (NIR) and partial least squares (PLS) to estimate the lignin content of a number of sugar cane genotypes is the goal of this work. Laboratory analyzes were performed to determine the lignin content using the Klason method. The independent variables were obtained from different materials: dry bagasse, bagasse with broth , leaf and stalk, without any pre-treatment. The NIR spectra were obtained in the range of 10000-4000 cm^{-1}. The Kennard Stone algorithm was used to select the calibration and the prediction set. The models were built using the partial least squares regression (PLS) and different algorithms for variable selection were tested: iPLS, biPLS, Genetic Algorithm (GA) and the Ordered Predictors Selection method (OPS). For dry bagasse, the best model was obtained after screening 445 variables with OPS, which obtained RMSEP of 0,85, Rp 0,97, RPD 2,87 and mean relative error of 2,82%; for bagasse with broth the best model was obtained after screening 265 variables with OPS, which obtained RMSEP of 0,65, Rp 0,94, RPD 2,77 and mean relative error of 1,94%; for leaf the best model was obtained after screening 305 variables with OPS, which obtained RMSEP of 0,58, Rp 0,96, RPD 2,56 and mean relative error of 2,47%; for the middle stalk the best model was obtained after screening 205 variables with OPS, which obtained RMSEP of 0,61, Rp 0,95, RPD 3,24 and mean relative error of 1,97%; for the top stalk the best model was obtained after screening 300 variables with OPS which obtained RMSEP of 0,58, Rp 0,96, RPD 2,34 and mean relative error of 1,94%; for foot middle and top stalk , the best model was obtained after screening 250 variables with OPS, which obtained RMSEP of 0,80, Rp 0,99, RPD 2,79 and mean relative error of 2,90%. The OPS algorithm selected fewer variables with greater predictive capacity. All models are reliable, with high accuracy for predicting lignin in sugar cane, reducing significantly the time to perform the analysis, the cost and the chemical reagents consumption, optimizing the whole process.

1. INTRODUÇÃO

Nas últimas décadas têm-se observado um aumento gradativo do interesse mundial pelo desenvolvimento dos biocombustíveis devido a maior preocupação com problemas ambientais causados pela queima de combustíveis fósseis e o interesse pelo desenvolvimento de fontes renováveis de energia. Os prejuízos ambientais causados pelo aumento da concentração dos gases responsáveis pelo efeito estufa e o esgotamento das reservas de petróleo de fácil extração têm estimulado a utilização de insumos renováveis.O objetivo é diminuir o consumo dos combustíveis de origem fóssil como petróleo, carvão e gás natural [1].

Neste contexto, a sustentabilidade econômica, social e ambiental no mundo dependerá grandemente de pesquisas por fontes alternativas de energia. Para o Brasil, pesquisas em biocombustíveis são fundamentais para garantir a atual posição de autonomia energética brasileira, além do interesse pela produção de excedentes para garantir a exportação [2].

Biocombustíveis renováveis devem ser compreendidos como um caminho para reduzir a dependência ao petróleo, diminuir as emissões de gases do efeito estufa e incentivar o desenvolvimento no setor agrário [3]. Diante dessa necessidade, a celulose tem se destacado por ser a biomassa renovável mais abundante na terra e por isso, uma alternativa energética em relação ao petróleo. A cana-de-açúcar (Saccharum spp.) apresenta cerca de 2/3 de sua massa em material lignocelulósico. Este é um recurso energético abundante e não utilizado de forma eficaz por meio das tecnologias atuais, representando, portanto, um enorme potencial para a produção de energia.O etanol da cana-de-açúcar é o biocombustível mais comum e mais promissor no Brasil, visto que seu balanço energético é geralmente positivo, i.e., a cana-de-açúcar captura mais carbono do que emite quando o etanol é queimado [4-10].

A cana-de-açúcar é uma gramínea de origem asiática e introduzida no Brasil pelos portugueses no início do século XVI, sendo hoje muito cultivada em regiões tropicais e subtropicais do país [11]. Além do caldo da cana, que é matéria prima para a produção do açúcar e etanol (anidro e hidratado), no processo de industrialização obtêm-se a biomassa constituída pelo bagaço e palha, antes classificados como resíduos da produção agrícola. Esta biomassa altamente energética tem potencial para serconvertida em biocombustíveis, líquidos ou gasosos, através de processos bioquímicos ou termoquímicos.

Quimicamente o bagaço é altamente viável para a produção de bioetanol por possuir cerca de 50% de umidade, 45% de fibras lignocelulósicas, 2 a 3% de sólidos insolúveis e 2 a 3% de sólidos solúveis. É um material altamente complexo, constituído principalmente de celulose, hemicelulose e lignina, que são os responsáveis pelo seu elevado conteúdo energético[12].

De acordo com Barbosa et al. (2000) [13], as últimas três décadas foram marcantes no sentido de contribuir com o melhoramento genético, no desenvolvimento do setor canavieiro do Brasil, com ganhos acentuados de produtividade e qualidade. Ainda que os programas de melhoramento genético da cana-de-açúcar tenham se dedicado ao desenvolvimento de variedades altamente produtivas (colmos e sacarose) destinada à produção de etanol e açúcar, nos últimos anos essa cultura também passou a ser investigada em relação à fibra. Nesse sentido, o setor carece de estudos relacionados à eficiência do processo produtivo do etanol de segunda geração [14], principalmente no que diz respeito ao elevado investimento para conexão, transmissão e distribuição.

Além do etanol de segunda geração produzido a partir da biomassa e do biodiesel produzido a partir de óleos vegetais, estudos recentes têm mostrado que o resíduo lignocelulósico pode ser usado para produção de hidrocarbonetos de cadeias longas (C9, C15, C12), tais como furfural, hidroximetilfurfural (HMF) e ácido levulínico (LA), por processo de catálise. Uma das matérias prima economicamente viável e mais abundante para produção destes hidrocarbonetos é a biomassa disponibilizada pelas usinas de cana-de-açúcar [15,16]. Em todas as possíveis aplicações da biomassa do bagaço de cana como fonte de energia, seja pela queima direta, seja pela obtenção do etanol de segunda geração, seja para a produção de hidrocarbonetos, os componentes da biomassa lignocelulósica são extremamente importantes, pois são eles os responsáveis pela energia produzida [17-20]. Portanto, clones com maior produtividade de biomassa lignocelulósica são relevantes para este fim e são conhecidos como cana energia. O desenvolvimento de clones de cana-de-açúcar com maior quantidade de biomassa lignocelulósica tem sido objeto de diversos estudos atuais [18-21]. Para obter cultivares de cana energia com as características desejadas para maior potencial de geração de energia na queima ou maior produção de etanol de segunda geração, a identificação de genótipos e a produção de clones melhorados são necessários [22, 23, 24]. Maiores quantidades de celulose, hemicelulose e lignina, além dos açúcares, são desejadas na cana para produção de energia . Portanto, as quantificações destes biopolímeros e dos açúcares na cana são necessárias para tomadas de decisão em relação à produção de clones e ao melhoramento genético [25].

Com a caracterização dos recursos genéticos em banco de germoplasma, é possível realizar estudos de diversidade genética, sugerir possíveis cruzamentos entre acessos, determinar a importância dos caracteres na avaliação da diversidade existente,determinar a relação entre os caracteres e elaborar coleção nuclear [26]. Portanto este é um instrumento de interesse para os programas de melhoramento genético de cana-de-açúcar, visando o desenvolvimento de variedades promissoras, com alta produtividade para açúcar e etanol, e que apresentem também uma biomassa vegetal interessante para produção de bioenergia, representando uma fonte de recurso renovável e levando cada vez mais a sustentabilidade do setor sucroalcooleiro.

O melhoramento da cana de açúcar com o objetivo de obter uma espécie com maior energia líquida exige o plantio de diversos clones [27, 28], que precisam ser avaliados quanto à quantidade fracionada do material lignocelulósico da biomassa. Assim, uma grande quantidade de análises químicas por via úmida são necessárias. Apesar da exatidão, precisão e robustez, estas análises não podem ser aplicadas em um cenário industrial, comercial ou em pesquisas extensivas, uma vez que são de alto custo, extremamente morosas e ambientalmente incorretas devido ao alto consumo de reagentes e descarte de produtos poluidores.O uso de métodos quimiométricos para extrair informações de dados multivariados, tais como dados espectroscópicos, permitem reduzir significativamente o tempo, o custo e o impacto ambiental das análises químicas [29-31]. Nesse sentido, o uso da espectroscopia na região do infravermelho próximo (NIR), que abrange a faixa de 12800 a 4000 cm^{-1}, tem sido aplicada com sucesso na determinação não destrutiva da composição de lignina [32, 33, celulose [34-36] e hemicelulose [37-38]. A espectroscopia NIR é uma técnica rápida (menos de um minuto para obtenção do espectro), não-invasiva, adequada para uso em linha de produção e exige preparo mínimo de amostra. Além disso, a espectroscopia NIR, em conjunto com métodos quimiométricos, fornece calibrações robustas, ou seja, os parâmetros do modelo não se alteram de maneira significativa quando novas amostras são acrescentadas ou retiradas do conjunto de calibração.

Os objetivos deste trabalho são: (1) construir e validar modelos de calibração multivariada para análise de lignina diretamente sobre a cana ou biomassa, usando espectroscopia NIR e métodos quimiométricos; (2) usar os modelos para realizar previsões das propriedades químicas em diversos clones, para selecionar os melhores cultivares para produção de energia. Os objetivos específicos são: (1) separar as amostras que serão usadas para construir o modelo de calibração; (2) determinar com exatidão e precisão a quantidade de lignina nestas amostras usando métodos padrões ou

validados; (3) nas mesmas amostras obter os espectros NIR; (4) importar os dados para softwares específicos para construir e validar os modelos de calibração multivariada empregando métodos quimiométricos; (5) usar os melhores modelos validados para prever a concentração da propriedade química de interesse para novas amostras usando apenas o espectro.

2. REVISÃO BIBLIOGRÁFICA

2.1. Cana-de-açúcar

A cana-de-açúcar (Saccharum spp.) é uma das principais culturas agrícolas do Brasil cujo cultivo vem aumentando significativamente nos últimos anos. Na safra 2010/11, foram cultivados cerca de 8 milhões de hectares, a partir dos quais foram produzidas 624.991.000t de cana-de-açúcar, correspondendo a uma produtividade média de 77,8 t ha^{-1}. Em relação à safra anterior, esses valores correspondem a aumentos de 8,4% na área cultivada e de 3,4% na produção. No cenário mundial, o Brasil aparece como o maior produtor de cana-de-açúcar, seguido por Índia e China. Em nível nacional, entre os principais produtores aparecem os estados de São Paulo, Minas Gerais, Goiás e Paraná. Juntos, estes estados são responsáveis por mais de 80% da produção nacional [39].

A cana-de-açúcar é uma planta da família Poaceae, assim como milho, sorgo, arroz e outras gramas [40]. Como características dessa família, pode-se citar: inflorescência, crescimento do caule em colmos, folhas com lâminas de sílica em suas bordas e bainha aberta. Com relação ao melhoramento vegetal, a cana-de-açúcar vem passando por uma série de modificações, resultando em várias espécies que diferenciam entre si, principalmente quanto ao conteúdo de fibras e açúcares. Na atualidade, a maior parte da cana-de-açúcar cultivada é um híbrido multiespecífico, que recebe a designação de Saccharum ssp.

A cultura é bastante influenciada pelas condições edafoclimáticas; sofrendo a influência de fatores como: a precipitação pluvial, a temperatura, a umidade relativa do ar e a insolação. Estas variáveis são condicionantes climáticos importantes na determinação da disponibilidade hídrica e térmica para a cultura da cana-de-açúcar e têm efeito sobre o comportamento fisiológico da cultura em relação ao metabolismo de crescimento e desenvolvimento dos colmos, florescimento, maturação e produtividade [41].

A cana-de-açúcar é uma cultura resistente à seca, que contém 12–17% de açúcares totais com base no peso húmido, com umidade de 68-72% (90% de sacarose e 10% de glicose ou frutose). A eficiência média de extração para produzir o caldo da cana por esmagamento é de aproximadamente 95% e o resíduo sólido restante (8-11%) é a fibra da cana (bagaço) [42]. A fibra da cana é composta de polímeros de carboidratos

heterogêneos e complexos (celulose, hemicelulose e lignina). A celulose (40-60% m/m) consiste de polímeros de glicose de alto peso molecular que são mantidos como feixes de fibras. A hemicelulose (20-40% m/m) consiste de polímeros mais curtos de vários açúcares que aglutinam os feixes de celulose. A lignina (10-30% m/m) consiste de um polímero tri-dimensional de propil-fenol, embutido e ligado à hemicelulose e proporciona rigidez à estrutura [43]. De todo este conteúdo da fibra, apenas 2,4% são cinzas.

O vegetal é composto pelo epiderme, sistema de células que recobre e protege o talo; a casca, cuja função é sustentação e proteção dos efeitos mecânicos externos, e o tecido parenquimatoso, cuja função é armazenar o suco açucarado. Imerso dentro desse tecido aparecem feixes de fibras e vasos que, juntos, possuem a função de conduzir os nutrientes e produtos produzidos pela planta [44].

2.1.1. Importância Econômica

O Brasil é o maior produtor de cana-de-açúcar do mundo,sendo responsável por cerca de 416 milhões de toneladas, seguido pela Índia e China. Em média, 55% da produção brasileira destina-se à produção de etanol e 45% à produção de açúcar. A cana-de-açúcar é cultivada nas regiões Centro-Sul e Nordeste, o que permite dois períodos de safra. Na região Centro-Sul, a safra ocorre de abril a novembro e na região Nordeste ela ocorre de novembro a abril [39]. Para o Brasil, e particularmente para o Estado de São Paulo, a cana-de-açúcar é uma cultura muito importante. A cana-de-açúcar foi introduzida no Brasil em 1532 e já teve grande importância na economia do país no passado. Ao longo dos últimos anos, esta cultura vem se destacando novamente, sendo a principal cultura explorada no Estado de São Paulo (terceiro maior produtor mundial de cana-de-açúcar), que permanece como o maior produtor do Brasil, com 51,7% (4.552,0 mil hectares) da área plantada [39].

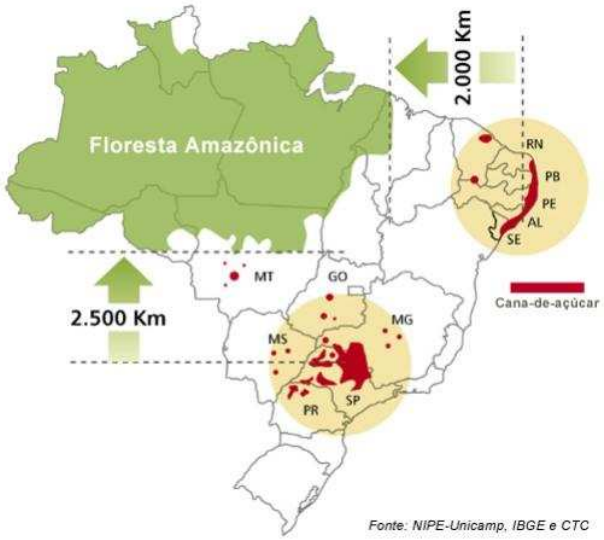


Figura 1. Mapa da produção de cana-de-açúcar no Brasil

De acordo com informações da Companhia Nacional de Abastecimento [39], a produção total de cana-de-açúcar moída na safra 2013/14 foi de 658,8 milhões de toneladas, com aumento de 11,9% em relação à safra 2012/13, que foi de 588,9 milhões de toneladas, significando um aumento de 69,9 milhões de toneladas maior que na safra anterior. A produção de cana-de-açúcar da Região Centro-Sul foi de 602,1 milhões de toneladas, 13,0%maior que a produção da safra anterior. Deste total de cana-de-açúcar acrescido nesta safra, 60,2% proveio de São Paulo, 13,7% de Minas Gerais, 13,3% de Goiás, 6,5% de Mato Grosso do Sul e 3,6% do Paraná, totalizando 97,3% deste crescimento. A Região Norte/Nordeste teve aumento de 1,4%, passando de 55,9 milhões de toneladas da safra 2012/13, para 56,7 milhões na safra 2013/14. A agroindústria do açúcar e do álcool gera para o Brasil, um produto final de dez bilhões de dólares por ano, um milhão de empregos diretos e indiretos e o seqüestro de 20% das emissões de carbono que o setor de combustíveis fósseis emite no país.

2.2. Lignina

A palavra lignina é proveniente do latim lignum, que significa madeira, e pode ser definida como um polímero tridimensional, amorfo, heterogêneo e altamente ramificado, que preenche espaços entre os polissacarídeos. Nesse sentido, a lignina é um dos componentes principais dos tecidos vasculares de gimnospermas e angiospermas. Sabe-se que a lignina é responsável pela rigidez, proteção contra patógenos e pela baixa reatividade dos materiais lignocelulósicos [45].

A lignina é uma das substâncias naturais mais abundantes da face da terra, ocupando cerca de 30% dos carbonos da biosfera e são exclusivamente formadas dentro da parede celular [46]. Quimicamente, a lignina é um composto de estrutura irregular, de alto peso molecular, altamente insolúvel e recalcitrante. Estruturalmente, este biopolímero apresenta inúmeros grupos aromáticos e alifáticos, formados pela polimerização oxidativa de três precursores fenilpropanóides monoméricos: álcool sinapílico (propanol siringil), álcool coniferílico (propanol guaiacil) e álcool p-cumerílico (propanol p-hidroxifenil), unidos entre si e com os polissacarídeos da parede celular, por meio de diferentes tipos de ligações, como do tipo éter (hidroxilas primárias e secundárias, carbonilas, carboxilas, ésteres e ligações etilênicas) ou carbono-carbono [47,48].

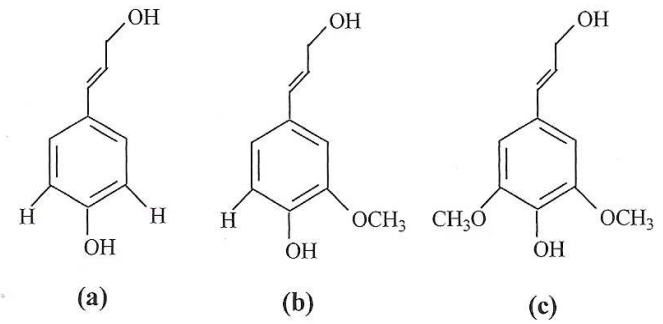


Figura 2. Precusores primários das ligninas: álcool p-cumarílico (I), álcool coniferílico (II) e álcool sinapílico (III).

Na madeira de eucalipto, a lignina é geralmente formada, principalmente, pelas unidades siringil (S) e guaiacil (G) (lignina S-G); em coníferas, é formada por unidades guaiacil e p-hidroxifenil (H) (lignina G-H). Em gramíneas, como a cana-de-açúcar, a formação da lignina envolve a polimerização dos três tipos de unidades monoméricas (lignina H-G-S) [49].

Apesar de todos os estudos e esforços realizados até hoje, a estrutura da lignina é bastante complexa e ainda não conhecida completamente. Isto acontece, pois a proporção dos precursores da lignina varia entre as diferentes espécies de plantas e da grande diversidade da estrutura das ligninas quando se passa de uma espécie vegetal para outra ou, até mesmo, dentro da mesma espécie, em partes diferentes do vegetal. Na literatura há uma série de modelos de ligninas, todos construídos a partir de análises de grupos funcionais e espectroscópicas. Na Figura 4é apresentado um modelo estrutural da lignina proposto por FENGEL,et al. [46].

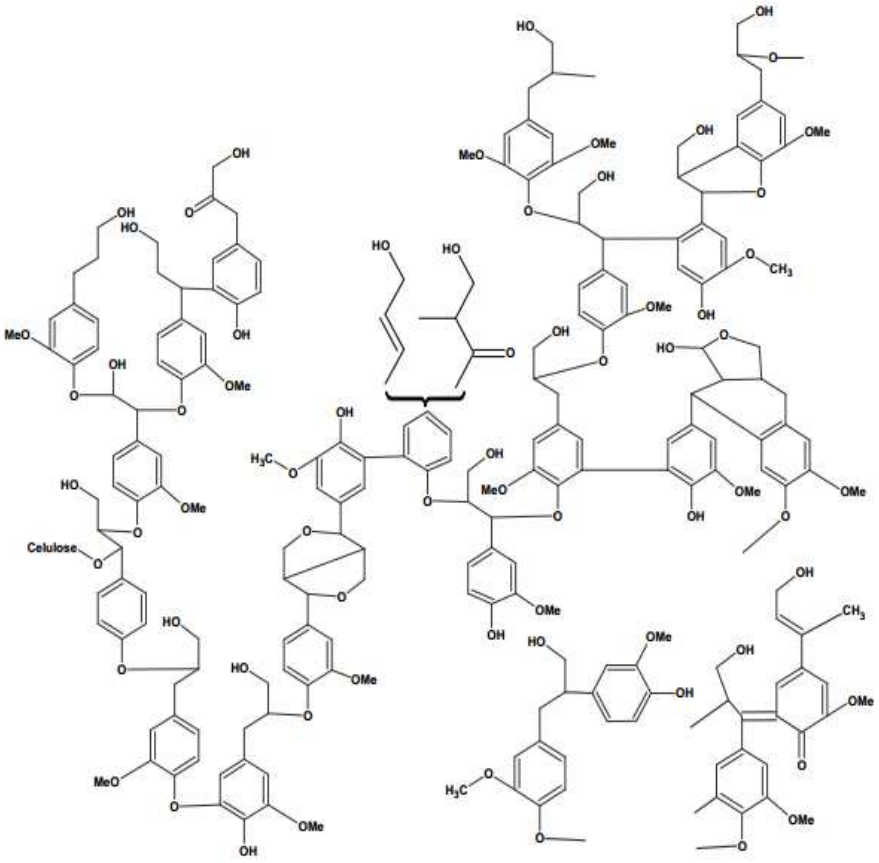


Figura 3. Modelo esquemático da lignina [46].

A lignina representa um dos maiores estoques de carbono/energia da natureza e é o maior depósito de estruturas químicas aromáticas, constituindo-se em uma fonte potencial de valiosos insumos para a indústria química**.** Apesar de ser possível produzir diversos produtos com base na lignina, atualmente o foco dos estudos tem se voltado para o uso desse material como fonte de energia para os processos, garantindo a auto-suficiência energética brasileira. Nesse sentido, maiores quantidades de celulose, hemicelulose e lignina, além dos açúcares, são desejadas na cana para produção de energia [50]. Logo, as quantificações destes biopolímeros e dos açúcares na cana são necessárias para tomadas de decisão em relação à produção de clones e ao melhoramento genético [51].

2.3. Espectroscopia no Infravermelho Próximo (NIR)

A espectroscopia consiste, basicamente, em um método analítico baseado nas interações de radiação eletromagnética com a matéria. Pela análise do espectro obtido é possível obter informações importantes sobre a estrutura molecular e o modo de

interação entre as moléculas. A energia eletromagnética pode ser ordenada de maneira contínua em função de seu comprimento de onda ou de sua frequência, sendo esta disposição denominada de "espectro eletromagnético", que apresenta subdivisões de acordo com as características de cada região. Dessa forma, o espectro eletromagnético se estende desde comprimentos de onda muito curtos (raios cósmicos), até as ondas de rádio de baixa frequência e grandes comprimentos de onda, de acordo com a Figura 5 [52]. De acordo com o valor de energia da radiação eletromagnética, as transições entre os estados podem ser de vários tipos, dos quais as principais são as transições eletrônicas, vibracionais e rotacionais [53].

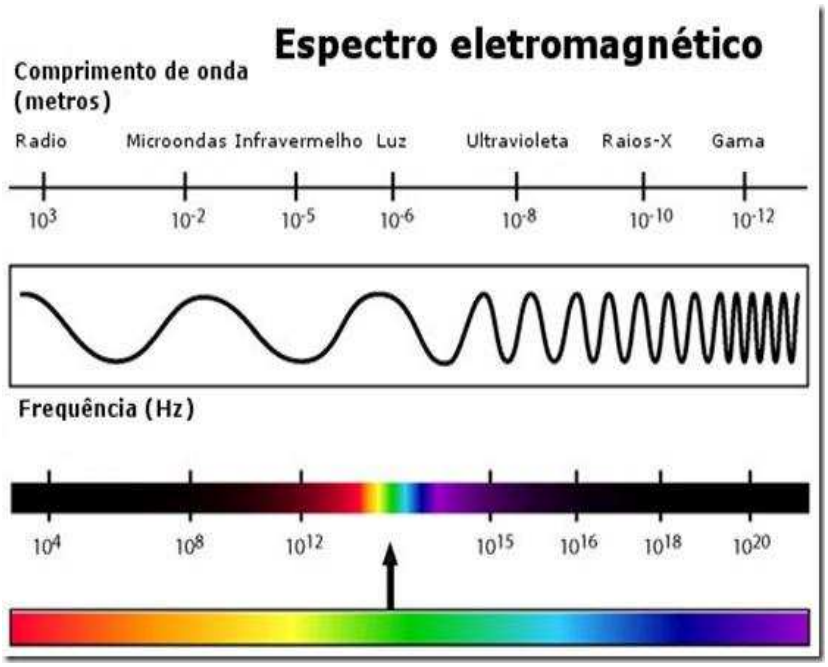


Figura 4. Espectro Eletromagnético.

A região espectral no infravermelho compreende a faixa de radiação com números de onda no intervalo de aproximadamente 12800 a 4000cm^{-1} [54]. O espectro na região do infravermelho é usualmente dividido em infravermelho próximo (NIR), infravermelho médio (MID) e infravermelho distante (FAR). Os limites aproximados para cada região espectral são mostrados na Tabela 1.

Tabela 1. Regiões espectrais do infravermelho

Região	Comprimento de Onda (nm)	Número de Onda (cm^{-1})	Frequência (Hz)
Próximo (NIR)	780 – 2500	12800 - 4000	$3,8x10^{14}$ - $1,2x10^{14}$
Médio (MID)	2500 -50000	4000 - 200	$1,2x10^{14}$ - $6,0x10^{12}$
Distante (FAR)	50000 - 1mm	200 - 10	$6,0x10^{12}$ - $3,0x10^{11}$

A espectroscopia NIR, técnica utilizada nesse trabalho, surgiu há aproximadamente 40 anos e tem sido largamente aplicada para determinação de moléculas orgânicas e parâmetros qualitativos de produtos agrícolas e alimentares, bem como no campo industrial, tais como petróleo, têxtil, carvão, cosméticos, polímeros, química, tintas e farmacêuticas [55]. O NIR é a denominação dada à região do espectro eletromagnético logo após a região visível (780 a 2500 nm). De forma geral, as ocorrências espectrais na região NIR provêm de ligações das moléculas em que participam o hidrogênio, o que torna a técnica útil para a determinação de compostos orgânicos contendo ligações C-H, N-H, e O-H [56].

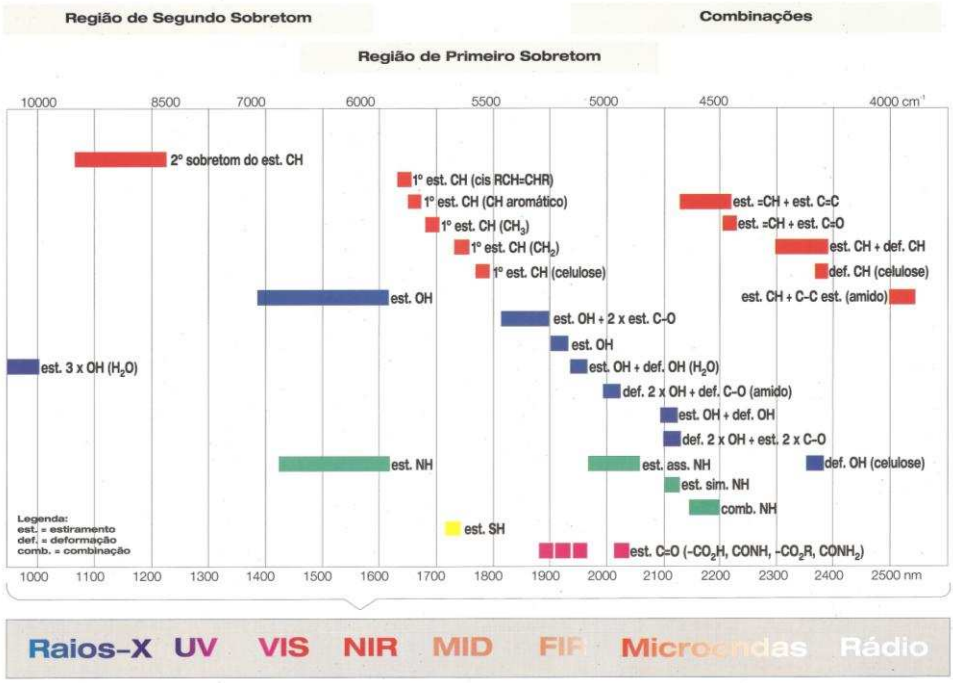


Figura 5. Bandas e posições relativas de picos de absorção em infravermelho próximo.

As vibrações observadas na espectroscopia NIR são referentes a sobretons e combinações de ligações fundamentais sendo, portanto, de pequena intensidade, restringindo a ligações C-H, N-H, O-H e S-H. Como essas ligações estão presentes na

maior parte das moléculas, a utilização do NIR em análise qualitativa para identificação de compostos é bastante restrita, ao contrário do infravermelho médio. Além disso, também é rara a observação de um comprimento de onda seletivo, que permita o desenvolvimento de um método de quantificação univariado. Por isso, a espectroscopia NIR ficou estagnada [57] por um longo período de tempo, até que o desenvolvimento da quimiometria permitisse a aplicação quantitativa desta técnica. Dentre as vantagens da espectroscopia NIR, podemos citar que é uma técnica rápida (um minuto ou menos na leitura por amostra), não invasiva, adequada para uso em linha de produção e exige preparo mínimo da amostra [58].

2.4. Calibração Multivariada

Calibração pode ser definida como uma série de operações que estabelecem uma relação entre medidas de um conjunto de padrões de referência e valores de uma propriedade de interesse. Essa relação, em calibração, é denominada modelo e o principal objetivo é realizar previsões a partir de respostas obtidas de amostras com valores desconhecidos da propriedade modelada. Para a construção de um modelo é necessário, primeiramente, calibrar um conjunto definido de padrões de referência para a propriedade de interesse.As amostras desconhecidas, de onde serão obtidas as respostas previstas das concentrações ou propriedades de interesse, apresentam, necessariamente, as mesmas características dos padrões de referência usados na construção do modelo [59].

Diversos são os modelos que podem ser empregados para relacionaras respostas com as concentrações/propriedades. Atualmente, muitos modelos exigem certo tipo de arranjo das respostas para serem construídos. Estes arranjos dependem do tipo de resposta coletada e podem ser definidos da seguinte maneira: (1) quando uma resposta é obtida para cada amostra, a resposta é um escalar e pode-se definir como arranjo de ordem zero; (2) quando muitos escalares são obtidos para cada amostra i.e., um vetor, define-se como arranjo de primeira ordem; (3) quando muitos vetores de mesmo comprimento são obtidos para cada amostra i.e., uma matriz, define-se como arranjo de segunda ordem e assim, sucessivamente [60].

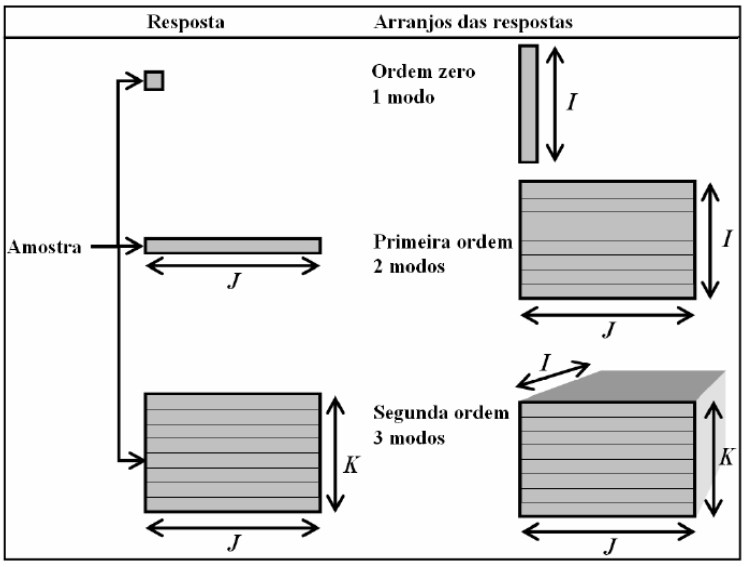


Figura 6. Representação de arranjo de dados para calibração[59].

Dentre os métodos de calibração existentes, os mais conhecidos são os métodos de calibração univariada, ou de ordem zero. Neste caso, tem-se apenas uma medida instrumental (resposta) para cada uma das amostras de calibração, isto é, para cada amostra tem-se apenas um escalar. Esses métodos são descritos na literatura em vários trabalhos [61-64] e sua aplicação e validação são relativamente fáceis. No entanto, a aplicação da calibração univariada é restrita, pois não é permitido a modelagem na presença de interferentes.

Em calibração multivariada, por sua vez, mais de uma resposta instrumental é relacionada com a propriedade de interesse. Uma das principais vantagens deste método é que possibilita a análise mesmo na presença de interferentes, desde que esses interferentes estejam presentes nas amostras utilizadas para a construção do modelo de calibração. Isso faz com que os modelos de calibração multivariada sejam uma alternativa quando os métodos univariados não encontram aplicação [65]. Neste tipo de calibração a resposta instrumental é representada na forma de matriz, onde as colunas representam as variáveis e as linhas as amostras, enquanto a propriedade de interesse, determinada por uma metodologia padrão, é representada por um vetor, de dimensão igual ao número de amostras. A Figura 7 ilustra como uma matriz de dados pode ser construída.

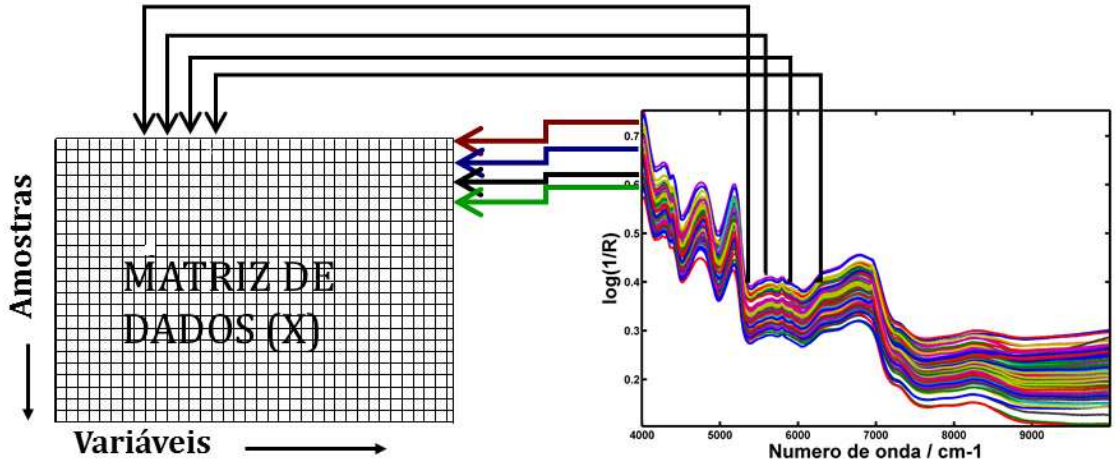


Figura 7. Construção da matriz **X** para calibração multivariada.

Diversos métodos de regressão vem sendo utilizados em química analítica para a construção de modelos de calibração multivariada, dentre esses os mais empregados tem sido a regressão linear múltipla (MLR), regressão por componentes principais (PCR) e regressão por quadrados mínimos parciais (PLS), que são métodos para ajuste linear entre as variáveis.

2.5. Regressão por Quadrados Mínimos Parciais (PLS)

A regressão por quadrados mínimos parciais (PLS) é considerada o método de regressão mais utilizado para a construção de modelos de calibração multivariada a partir de dados de primeira ordem. Herman Wold [66] foi o grande responsável pelo desenvolvimento do PLS. Isto ocorreu por volta do ano de 1975 quando ele trabalhava com dados na área de econometria [67]. Atualmente, a regressão por quadrados mínimos parciais tornou-se uma ferramenta padrão para modelagem de dados multivariados.

Para construção de um modelo utilizando PLS é necessário, primeiramente, realizar a compressão dos dados, o que gerará as variáveis latentes (equivalentes às componentes principais). Essas variáveis latentes descrevem o espalhamento máximo das amostras, contendo informações diferentes e complementares em ordem decrescente de variância. Vale ressaltar que, durante a construção das variáveis latentes, leva-se em consideração a correlação entre a matriz **X** e o vetor **y** [68].

Diversos algoritmos PLS estão disponíveis na literatura em softwares, tais como NIPALS, bidiagonal, SIMPLS e Kernel. Dentre estes o considerado mais rápido e

correto matematicamente é o bidiagonal, proposto por Manne [69]. Dessa forma, a regressão através do algoritmo bidiagonal foi escolhida para a realização desse trabalho, que consiste na decomposição da matriz **X** em outras matrizes, conforme equação 1 [68]:

$$\mathbf{X} = \mathbf{URV}^t \quad (1)$$

em que **UR** é a matriz de escores e **V** a matriz de loadings.

O algoritmo pode ser descrito na seguinte formulação resumida [70]:

1. Inicialize o algoritmo para a primeira componente,

$$\upsilon_1 = \mathbf{X}^t\mathbf{y} / \left\|\mathbf{X}^t\mathbf{y}\right\|; \alpha_1\mu_1 = \mathbf{X}\upsilon_1$$

2. Para i = 2, ..., h componentes:

2.1 $y_{i-1}\upsilon_i = \mathbf{X}^t\mu_{i-1} - \alpha_{i-1}\nu_{i-1}$

2.2 $\alpha_i\mu_i = \mathbf{X}\nu_i - y_{i-1}\mu_{i-1}$

Com $\mathbf{V}_h = (\upsilon_i,...,\upsilon_h)$, $\mathbf{U}_h = (\mu_1,...,\mu_i)$ e $\mathbf{R}_h = \begin{pmatrix} \alpha_1 & y_1 & & \\ & \ddots & \ddots & \\ & & \alpha_{k-1} & y_{k-1} \\ & & & \alpha_k \end{pmatrix}$ prova-se ainda que $\mathbf{XV}_h = \mathbf{U}_h\mathbf{R}_h$ e, portanto, $\mathbf{R}_h = \mathbf{U}^t_h\mathbf{XV}_h$..

Com as matrizes **U**, **V** e **R** calculadas para h variáveis latentes, pode-se estimar a pseudoinversa de Moore-Penrose de **X** e resolver o problema dos quadrados mínimos, como mostrado a seguir

$$\mathbf{y = Xb} \rightarrow \mathbf{y = U_hR_hV_h^t} \rightarrow \hat{\mathbf{b}} = \mathbf{U_hR_h^{-1}V_h^ty} \quad (2)$$

em que h é o número de variáveis latentes selecionado para a construção do modelo.

A escolha do número de variáveis latentes h é de extrema importância para evitar sub-ajuste ou sobre ajuste do modelo, pois as matrizes reconstruídas dependem do valor de h. O primeiro caso implica a modelagem de dados insuficientes para explicar toda informação e o segundo, a inclusão de excesso de informação no modelo, que pode ser aleatória ou estar sistematicamente relacionada à presença de erros. De um modo geral,

o risco de sobre-ajuste em calibração multivariada é muito maior do que o de sub-ajustes [71]. Uma das técnicas mais empregadas para a escolha do número de variáveis latentes (embora nem sempre é a mais exata) é a validação cruzada. Esta técnica baseia-se no procedimento de reamostragem. A partir de um gráfico que relaciona os erros na reamostragem versus h, seleciona-se o ponto de menor erro. Normalmente usa-se como cálculo do erro a raiz quadrada do erro quadrático médio de validação cruzada (RMSECV), conforme Equação 3:

$$RMSECV = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{I_{cv}}} \quad (3)$$

em que $\hat{y}_i$ o valor estimado para a amostra i, e I_{cv} é o número de amostras da validação interna.

Os principais métodos de validação cruzada são: i) leave-one-out, que remove uma amostra de cada vez, sendo indicado apenas para modelos com pequeno número de amostras (aproximadamente 20); ii) blocos contíguos, que separa amostras em blocos de amostras sequenciais; iii) venezianas (venetian blinds), que separa amostras sistematicamente espaçadas; e iv) subconjuntos aleatórios, que separa, aleatoriamente, conjuntos de amostras [71].

Além disso, a qualidade dos modelos foi avaliada pelo RPD [72] (desvio do resíduo da predição) e viés. O RPD é calculado de acordo com a Equação 4:

$$RPD = \frac{D_p}{RMSECV} \quad (4)$$

em que D_p é o desvio padrão dos dados de referência.

De acordo com Dunn et al. [73] e Chang et al. [74], valores de RPD acima de 2,0 já pode ser considerado como modelos excelentes (A), de 1,4 à 2,0 modelos aceitáveis (B) e menor que 1,4 (C) modelos não confiáveis.

Já o viés, ou bias, definido como a diferença entre o valor de uma medida e o valor de referência, é a medida de erros sistemáticos no modelo, sendo calculado apenas com as amostras de validação. O cálculo de viés para calibração multivariada é proposto pela norma ASTM conforme equação 5:

$$\text{viés} = \frac{\sum_{i=1}^{nv} (y_i^{ref} - \hat{y}_i)}{I_m} \quad (5)$$

em que y_i^{ref} é o valor de referência e $\hat{y}_i$ é o valor previsto de cada amostra, e I_m é o número de amostras do conjunto de validação.

O desvio padrão de erros de validação (SDV – standard deviation of validation errors) é usado para verificar se o viés é significativamente diferente de zero, por meio de um teste t de Student, com nv graus de liberdade:

$$SDV = \sqrt{\frac{\sum \left[\left(y_i^{ref} - \hat{y}_i \right) - \text{bias} \right]^2}{nv - 1}} \quad (6)$$

$$t_{calc} = \frac{|\text{viés}| \sqrt{nv}}{SDV} \quad (7)$$

2.6. Seleção de Variáveis

A escolha adequada de variáveis na matriz de dados **X** pode melhorar significativamente a capacidade de previsão do modelo de calibração multivariada. Desta forma, a seleção de variáveis envolve a escolha de determinadas regiões do espectro (um conjunto de comprimentos de onda), que minimizam o erro na previsão. Como consequência, têm-se a construção de modelos mais robustos, simples de interpretar e com melhor exatidão [75]. Em espectroscopia,comprimentos de onda que se referem a ruídos, informações irrelevantes ou não-linearidades, podem ser eliminados [76]. Diversos métodos para seleção de variáveis foram desenvolvidos para calibração multivariadas, sendo que eles variam em relação à estratégia de escolha das melhores variáveis. Em espectroscopia dois métodos tem sido muito utilizados, que são o algoritmo genético (GA) [77] e o método de quadrados mínimos parciais por intervalo (iPLS) e suas variações [76,78]. Mais recentemente, o método de seleção dos preditores ordenados (OPS) foi apresentado e poucas aplicações foram realizadas em espectroscopia [79]. Estes algoritmos têm se mostrado muito eficiente para seleção de descritores em estudos da relação estrutura-atividade quantitativa (QSAR) [80]. Nesta dissertação os algoritmos GA, iPLS, biPLS e OPS serão executados e comparados

quanto à capacidade de previsão, número de variáveis selecionadas, facilidade de execução e tempo de cálculo.

2.6.1. Quadrados Mínimos Parciais por Intervalos (iPLS)

O iPLS tem como fundamento encontrar uma ou mais regiões do espectro (faixa) que produza melhores resultados para previsão que o espectro completo. No método iPLS é feita uma regressão por quadrados mínimos parciais em cada sub-intervalo equidistante ao longo de toda a extensão das variáveis na matriz de dados [78]. Deste modo é possível avaliar e identificar os subconjuntos de variáveis que apresentam informações mais relevantes. As regiões espectrais cujas variáveis se apresentam como de menor importância e responsáveis por informações ruidosas são removidas. O modelo é construído, portanto, apenas com o subconjunto ou subconjuntos selecionados.

A otimização dos subintervalos iPLS consiste de duas etapas: selecionar o número de intervalos empregados e escolher a amplitude de cada intervalo. Os diferentes modelos obtidos com os intervalos espectrais são avaliados através do RMSECV, RMSEP e Rcv. Para cada intervalo, é construído um modelo PLS apresentado na forma gráfica para facilitar a comparação com toda a faixa espectral utilizada. O método é planejado para dar uma visão geral dos dados e pode ser útil para selecionar as variáveis mais representativas na construção de um modelo de calibração adequado [81].

Uma versão backward PLS por intervalos (biPLS) foi proposta por Leardi e Norgaard [81]. A ideia básica é similar ao procedimento executado no iPLS, contudo intervalos são usados no lugar de variáveis individuais. O espectro está dividido em um determinado número de intervalos e os modelos PLS são calculados a cada intervalo deixado de fora. A definição do número de intervalos no biPLS é uma tarefa de extrema importância pois, se o número de intervalos for demasiadamente pequeno, as regiões espectrais serão largas e, consequentemente, há perda de informação dos picos menores. Por outro lado, se o número de intervalos for muito grande, os resultados ficarão numa escala local e será preciso um tempo computacional maior.

2.6.2.Algoritmo Genético (GA)

O GA se enquadra nos chamados métodos de inteligência artificial [82] e é muito utilizado na seleção de comprimentos de ondas em calibração multivariada. O algoritmo funciona simulando matematicamente a teoria de Darwin: as condições experimentais (cromossomos), que levam as melhores respostas (indivíduos mais adaptados ao ambiente) tem maior chance de serem selecionadas (sobreviver), sendo transmitidas às novas gerações através da reprodução. Dessa forma, a otimização das respostas (evolução da espécie) é alcançada por meio da recombinação das variáveis (cruzamento de cromossomos) e de algumas modificações aleatórias (mutações de genes).

A etapa inicial do GA consiste em executar um grande número de seleções aleatórias das variáveis independentes e calcular o RMSECV para cada subconjunto. O RMSECV usado para cada modelo PLS é o menor obtido entre todos os calculados, limitado pelo número h previamente definido [83]. Cada subconjunto de variáveis selecionado é chamado de indivíduos (ou cromossomos) e o código sim (1) ou não (0) é o gene para aquele indivíduo, indicando quais variáveis serão usadas (1) e quais não serão usadas (0). A população é formada pela combinação de todos os indivíduos testados. Os valores de RMSECV indicam a capacidade de previsão das variáveis selecionadas naquele indivíduo. Devido à presença de variáveis ruidosas e interferentes, a aptidão dos diferentes indivíduos na população se estenderá por uma faixa de valores.

Na etapa de seleção os indivíduos com valores maiores de RMSECV que o valor mediano são descartados. Neste ponto, há uma redução considerável do tamanho da população. Dessa forma, o cruzamento dos indivíduos restantes é realizado, através de dois métodos: cross-over simples ou duplo [59]. No simples, os genes de dois indivíduos aleatoriamente selecionados são divididos em algum ponto aleatório do gene. A primeira parte do gene do individuo A é trocada com a primeira parte do individuo B, e os dois genes híbridos formam dois novos indivíduos (C e D) para a população. No cross-over duplo, dois pontos aleatórios no gene são selecionados e a porção do meio dos dois genes é trocada. A grande diferença, na prática, é que o duplo gera tipicamente novos subconjuntos de variáveis com maior número de variáveis iguais aos dos cromossomos de origem .

Após a adição de novos indivíduos à população, todos os genes individuais podem sofrer mutação aleatória. Isto permite, dentro de uma chance finita de adição ou remoção, considerar variáveis que possam estar super ou sub-representadas na população. As mutações são necessárias para superar alguns problemas que podem

ocorrer durante a seleção. O problema mais essencial a ser resolvido é que se uma variável não for selecionada em qualquer cromossomo original, ela nunca será selecionada nas gerações futuras se a mutação não existir [59].

Finalmente, após o emparelhamento e cruzamento de todos os indivíduos, a população retorna para o tamanho original e o processo pode novamente continuar na etapa de avaliação da aptidão (RMSECV), i.e., na segunda etapa [59].A finalização do GA pode ocorrer depois de um número de iterações finito ou depois de alguma percentagem de indivíduos na população estarem usando subconjunto de variáveis idênticas.

Como principais vantagens do GA, podemos citar: realiza busca simultâneas em várias regiões do espaço amostral, aplicável a uma ampla gama de otimizações, não requer informações sobre a superfície de resposta, fornece uma lista de variáveis ótimas, entre outros [84].

2.6.3. Seleção dos Preditores Ordenados (OPS)

O método OPS foi proposto por Teófilo e colaboradores no ano de 2008 [85] e tem como principal objetivo automatizar a seleção de variáveis usando vetores que trazem informações sobre os preditores mais importantes na matriz original.

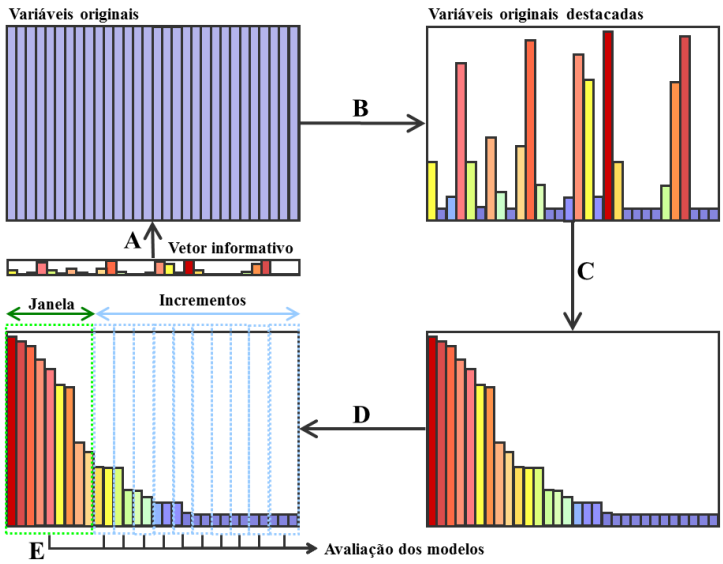


Figura 8. Método de seleção de variáveis usando a seleção dos preditores ordenados.

De forma geral, o objetivo do método é obter um vetor (vetor informativo), que contém informações com a localização das variáveis com melhor resposta para a predição (Figura 8A) [59]. Esses vetores são obtidos diretamente com os cálculos

efetuados com as variáveis dependentes e independentes, apenas com as variáveis independentes ou a partir de combinações de diferentes vetores obtidos com a mesma finalidade.

Posteriormente, na segunda etapa (Figura 8B), as variáveis originais são destacadas de acordo com os correspondentes valores absolutos dos elementos do vetor informativo obtidos anteriormente na etapa A. Quanto maior o valor absoluto, mais importante é a variável, o que permite a sua classificação em ordem decrescente de magnitude na terceira etapa (Figura 8C) [59].

Na quarta etapa (Figura 8D), são construídos diferentes modelos de regressão multivariada e os mesmos são avaliados através da estratégia de validação cruzada. Um primeiro subconjunto de variáveis (janela) é selecionado para construir e avaliar o primeiro modelo. Em seguida, essa matriz é expandida pela adição de um número fixo de variáveis (incremento) e um novo modelo é construído e avaliado. Incrementos novos são adicionados até que todos ou algum percentual das variáveis sejam analisados. Parâmetros estatísticos de qualidade dos modelos são calculados para cada avaliação e armazenados para futuras comparações. Finalmente (Figura 8E), os conjuntos de variáveis avaliadas (janela inicial e os incrementos) são comparados com a qualidade dos parâmetros calculados durante as validações. O modelo com os melhores parâmetros de qualidade contém variáveis com a melhor capacidade de previsão e, por esta razão, são as variáveis selecionadas [59].

Ao se usar o vetor informativo para a seleção de variáveis, deve-se dar atenção especial ao número de variáveis latentes, h, a ser utilizado na obtenção desse vetor. Primeiramente, deve-se determinar o número de variáveis, h = hMod, para construir e validar o modelo. Mas, geralmente, hMod não gera um vetor de regressão suficientemente informativo para seleção de variáveis. Para encontrar o melhor vetor de regressão para seleção de variáveis, um estudo deve ser realizado no conjunto completo de dados, aumentando o número de h do modelo, a partir de hMod e realização da seleção de variáveis usando o algoritmo OPS até um número de variáveis latentes ótimo para construir um vetor de regressão exclusivo para seleção de variáveis. Este número de variáveis latentes é definido como hOPS, número de variáveis latentes usado na construção do vetor de regressão usado para seleção de variáveis no algoritmo OPS [87].

Deste modo, dois números ótimos de h são empregados, um representando o número de h para construção do modelo (hMod) e outro representando o número de h empregados para gerar o melhor vetor informativo no método OPS (hOPS) [59].

2.7. Validação e Figuras de Mérito

O desenvolvimento de um método analítico, sua adaptação e implementação, envolve diferentes processos de avaliações que estimem sua eficiência na rotina do laboratório. A validação é um processo de averiguação da performance de um método, com o intuito de avaliar se este é adequado para as condições nas quais será aplicado. Dessa forma, o processo de validação deve ser realizado sempre que um procedimento analítico é proposto ou desenvolvido [88]. A validação de métodos baseados em espectroscopia no infravermelho próximo é fiscalizada pela American Society for Testing and Materials (ASTM) [89].

A validação pode ser obtida através do cálculo de parâmetros estatísticos conhecidos como figuras de mérito que são, nesse caso, os indicadores quantitativos do bom desempenho de um modelo [90]. Sendo assim, sensibilidade, limite de detecção (LOD), limite de quantificação (LQ), seletividade, entre outros, são parâmetros que constituem as figuras de mérito essenciais à validação de métodos analíticos.

2.7.1. Sensibilidade

É definida como a fração do sinal responsável pelo acréscimo de uma unidade de concentração à propriedade de interesse. Para modelos de calibração multivariada, como PLS, pode ser determinada como [91,92]:

$$\mathrm{SEN} = \frac{1}{\|\mathrm{b}\|} \quad (8)$$

em que $\|\mathrm{b}\|$ é a norma do vetor dos coeficientes de regressão estimados pelo PLS.

2.7.2. Sensibilidade Analítica

A sensibilidade analítica, γ, não é abordada em normas ou guias de validação. No entanto, esse parâmetro apresenta a sensibilidade do método em termos da unidade de concentração que é utilizada, sendo definida como a razão entre a sensibilidade e o desvio padrão do sinal de referência (∂_x) [93,94]:

$$\gamma = \frac{\mathrm{SEN}}{\|\partial_x\|} \quad (9)$$

em que ∂_x é o desvio padrão do sinal de referência estimado através do desvio padrão do valor de NAS para espectros do sinal de referência.

O inverso desse parâmetro, ou seja, γ^{-1}, permite estabelecer a menor diferença de concentração entre amostras, que pode ser distinguida pelo método.

2.7.3. Seletividade

É a medida do grau de sobreposição entre o sinal da espécie de interesse e os interferentes presentes na amostra, indicando também, a parte do sinal que é perdida por essa sobreposição [95]. Para modelos de calibração multivariada, a seletividade, $SEL_{k,i}$, é definida como [96,97]:

$$SEL = \frac{nas_{k,i}}{\|\mathbf{x}_{k,i}\|} \quad (10)$$

em que, $nas_{k,i}$ é o valor escalar do sinal analítico líquido para a amostra i e $x_{k,i}$ representa o vetor de respostas instrumental para a amostra i.

2.7.4. Limite de Detecção (LOD) e Limite de Quantificação (LOQ)

O limite de detecção (LOD) e o limite de quantificação (LOQ) de um procedimento analítico expressam as menores quantidades da espécie de interesse que podem ser detectadas e determinadas quantitativamente, respectivamente. Para um conjunto de dados que apresenta comportamento homoscedástico (variância constante ao longo da faixa de trabalho, erros na previsão que seguem uma distribuição normal e não são correlacionados), os LD e LQ na calibração multivariada podem ser calculados pelas seguintes equações [91]:

$$LOD = 3\|\partial_r\|\|\mathbf{b}\| \quad (11)$$

$$LOQ = 10\|\partial_r\|\|\mathbf{b}\| \quad (12)$$

em que ∂_r é o vetor de desvios padrões das colunas da matriz de ruídos e $\mathbf{b}$ é o vetor de regressão.

2.7.5. LOD e LOQ proposto por Teófilo, RF (2007)

Neste trabalho, será utilizado um método modificado [59] baseado em um dos métodos proposto pela IUPAC para o cálculo de LOD que utiliza os valores medidos e preditos do modelo de regressão, conforme equação 13:

$$\text{LOD} = \frac{\hat{a} + (3\sigma_a)}{\hat{b}} \quad (13)$$

em que â é o intercepto na equação de regressão, $\hat{b}$ é o coeficiente de regressão e σ_a é o erro relacionado ao intercepto, calculado a partir de repetições. A proposta apresentada consiste em realizar a estimativa do erro no intercepto usando o cálculo da matriz de variância e covariância considerando os valores estimados como variável independente do modelo [59]. O intercepto é estimado com uma coluna de uns, adicionada ao lado esquerdo dos valores estimados. Dessa forma:

$$\text{LOD} = \frac{\hat{a} + e_a}{\hat{b}} \quad (14)$$

3. MATERIAIS E MÉTODOS

3.1. Banco de Germoplasma

As amostras utilizadas nesse trabalho são pertencentes ao banco de germoplasma do Programa de Melhoramento Genético de Cana-de-Açúcar (PMGCA/UFV) Ridesa, constituído por aproximadamente 300 genótipos de cana-de-açúcar e localizado na área experimental da UFV - Universidade Federal de Viçosa, Viçosa – MG, Brasil.

A colheita dos colmos ocorreu aos 18 meses após o plantio e em cada parcela foram aleatoriamente colhidos 10 colmos. Os colmos foram cortados manualmente sem despalha prévia a fogo para obtenção das amostras visando análise química e espectrofotométrica. Por ocasião da colheita, foram obtidas amostras da folha +3 de cada genótipo. A Tabela 2 abaixo indica os genótipos utilizados no experimento.

Tabela 2. Identificação dos genótipos

Etiqueta Quadra 2	Genótipo	Gen. Feminino	Gen. Masculino	Origem
1842	SP80-1842	SP71-1088	H57-5028	UFV/RIDESA
7515	RB867515	RB72454	?	UFV/RIDESA
194	UFV09194	CO62-175	?	UFV/RIDESA
195	UFV09195	CO62-175	?	UFV/RIDESA
200	UFV09200	CO62-175	?	UFV/RIDESA
207	UFV09207	F150	?	UFV/RIDESA
210	UFV09210	F150	?	UFV/RIDESA
217	UFV09217	F150	?	UFV/RIDESA
227	UFV09227	NG57-6	?	UFV/RIDESA
230	UFV09230	CO223	?	UFV/RIDESA
233	UFV09233	B70710	?	UFV/RIDESA
253	UFV09253	B70710	?	UFV/RIDESA
259	UFV09259	B70710	?	UFV/RIDESA
260	UFV09260	B70710	?	UFV/RIDESA
263	UFV09263	B70710	?	UFV/RIDESA
264	UFV09264	B70710	?	UFV/RIDESA
269	UFV09269	B70710	?	UFV/RIDESA
273	UFV09273	B70710	?	UFV/RIDESA
280	UFV09280	IN84-105	?	UFV/RIDESA
285	UFV09285	IN84-105	?	UFV/RIDESA
288	UFV09288	IN84-105	?	UFV/RIDESA
289	UFV09289	IN84-105	?	UFV/RIDESA
290	UFV09290	IN84-105	?	UFV/RIDESA
297	UFV09297	CANA BLANCA	?	UFV/RIDESA
298	UFV09298	CANA BLANCA	?	UFV/RIDESA
900	UFV09900	CANA BLANCA	?	UFV/RIDESA

902	UFV09902	CANA BLANCA	?	UFV/RIDESA
904	UFV09904	CANA BLANCA	?	UFV/RIDESA
906	UFV09906	CANA BLANCA	?	UFV/RIDESA
909	UFV09909	NG21-21	?	UFV/RIDESA
910	UFV09910	RB93509	IN84-88	UFV/RIDESA
917	UFV09917	PT47-1010E	?	UFV/RIDESA
Banco de Germoplasma - Fundão/UFV				
2	SP80-1816	SP71-1088	H57-5028	Copersucar
4	RB001915	RB83594	?	UFAL/RIDESA
5	RB001906	?	?	
7	RB00509	RB931530	RB83594	UFAL/RIDESA
9	RB00416	CB47-355	?	UFAL/RIDESA
10	RB946903	RB765418	RB72454	UFPR/RIDESA
16	RB835089	RB72454	NA56-79	UFSCar/RIDESA
18	RB00507	RB931530	RB83594	UFAL/RIDESA
24	RB00412	SP80-1770	RB75126	UFAL/RIDESA
26	RB767647	?	?	
27	RB001913	RB931602	?	UFAL/RIDESA
28	RB867515	RB72454	?	UFV/RIDESA
29	SP79-1011	NA56-79	Co775	Copersucar
36	SP80-3280	SP71-1088	H57-5028	Copersucar
37	RB947501	SP71-1406	RB72454	UFV/RIDESA
40	RB835054	RB72454	NA56-79	UFSCar/RIDESA
42	SP83-2847	HJ5741	SP70-1143	Copersucar
47	RB93509	RB72454	?	UFAL/RIDESA
51	RB855113	SP70-1143	RB72454	UFSCar/RIDESA
53	RB867515	RB72454	?	UFV/RIDESA
54	RB855046	SP70-1143	TUC71-7	UFSCar/RIDESA
56	RB867515	RB72454	?	UFV/RIDESA
58	RB001914	RB72454	?	UFAL/RIDESA
65	MARIA PELADINHA	?	?	
66	RB867515	RB72454	?	UFV/RIDESA
67	RB931555	SP71-6113	?	UFAL/RIDESA
68	RB945961	RB855206	?	UFSCar/RIDESA
69	Q124	NCo310	54N7096	BSES
77	RB867515	RB72454	?	UFV/RIDESA
78	IN84-82	SAC.SPONTANEUM	?	Estrangeira
80	28NG289	SAC.ROBUST	?	Estrangeira
84	NG57-6	SAC.OFFIC. HYBRID	?	Estrangeira
87	RB72454	CP53-76	?	UFAL/RIDESA
110	RB00507	RB931530	RB83594	UFAL/RIDESA
111	RB00412	SP80-1770	RB75126	UFAL/RIDESA
113	UVA			
115	Co285	STR.MAURITIUS	SAC.SPONTANEUM	ICAR, SBI
116	28NG289	SAC.ROBUST	?	Estrangeira
118	RB739359	IANE55-34	?	UFRRJ/RIDESA

128	RB855156	RB72454	TUC71-7	UFSCar/RIDESA
129	RB855453	TUC71-7	?	UFSCar/RIDESA
133	RB037210	?	?	
134	RB985523	SP80-1520	SP70-1143	UFSCAR
135	RB977512	SP80-1842	?	UFV/RIDESA
136	RB975947	RB855563	RB735200	UFSCar/RIDESA
139	RB965911	RB855546	?	UFSCar/RIDESA
151	RB928064	SP70-1143	?	UFV/RIDESA
152	RB987649	RB72454	RB739359	UFV/RIDESA
153	SP70-1143	IAC48/65	?	Copersucar
154	RB008336	SP84-2029	SP82-6108	UFV/RIDESA
156	NA56-76	?	?	
158	RB037133	SP77-5181	RB835486	UFV/RIDESA
159	RB037049	RB855546	RB957689	UFV/RIDESA
161	RB047128	SP85-3877	?	UFV/RIDESA
162	RB037142	RB855511	SP80-1816	UFV/RIDESA
164	RB047157	SP85-3877	?	UFV/RIDESA
165	RB037064	RB925345	RB855156	UFV/RIDESA
166	RB037231	?	?	
167	RB037228	?	?	
169	RB975019	RB855563	RB735200	UFSCar/RIDESA
171	RB008340	SP80-3280	?	UFV/RIDESA
172	SP71-1406	NA56-79	?	Copersucar
176	RB975138	RB855598	RB845257	UFSCar/RIDESA
177	IM76-227	ERIANTHUS	?	Estrangeira
178	RB037042	RB947501	SP80-3280	UFSCar/RIDESA
180	RB037109	?	?	
182	RB037041	SP80-3280	RB855156	UFV/RIDESA
183	RB037193	RB951015	?	UFV/RIDESA
186	RB037189	SP80-185	RB912512	UFV/RIDESA
187	CB49-260	CB44-36	?	Campos Brasil
188	RB037044	RB835486	RB855511	UFV/RIDESA
189	SP77-5181	HJ57-41	?	Copersucar
191	RB987935	RB72454	RB83102	UFV/RIDESA
192	SP79-1011	NA56-79	Co775	Copersucar
194	CB49-15	Co213	?	CB
196	RB037163	SP80-3280	RB855156	UFV/RIDESA
197	RB047156	SP83-2847	?	UFV/RIDESA
199	RB037213	SP80-1816	?	UFV/RIDESA
200	RB047143	RB955993	?	UFV/RIDESA
201	RB037136	RB845197	SP77-5197	UFV/RIDESA
204	IAC86-2210	CP52-58	Co798	IAC
205	RB925211	RB855206	?	UFSCar/RIDESA
206	RB863129	RB763411	?	UFRPE/RIDESA
208	RB988078	RB83102	RB7245	UFV/RIDESA
209	RB988079	RB83102	RB7245	UFV/RIDESA
216	RB037229	SP83-5073	RB855511	UFV/RIDESA
218	RB037127	RB8495	SP91-1049	UFV/RIDESA

219	RB037055	SP83-2847	RB855035	UFV/RIDESA
220	RB037060	SP77-5181	RB835486	UFV/RIDESA
222	RB037174	RB928064	RB945962	UFV/RIDESA
227	RB988105	RB72454	RB83102	UFV/RIDESA
228	RB988037	?	?	
230	RB985571	RB835486	?	UFSCar/RIDESA
231	RB937568	SP70-1143	RB72454	UFV/RIDESA
234	RB008041	SP84-2025	SP80-3280	UFV/RIDESA
235	RB047031	RB955993	?	UFV/RIDESA
237	RB037032	RB928064	RB945962	UFV/RIDESA
244	RB008029	RB845197	?	UFV/RIDESA
245	RB977514	SP80-1842	?	UFV/RIDESA
248	RB997671	SP80-185	SP80-3280	UFV/RIDESA
249	RB998211	SP85-162	SP82-3530	UFV/RIDESA
250	SP81-3250	CP70-1547	SP71-1279	Copersucar
252	RB941537	RB83160	RB855035	UFAL/RIDESA
253	RB027053	SP85-3877	RB855536	UFV/RIDESA
254	RB037000	RB855156	RB957689	UFAL/RIDESA
255	RB037002	RB855511	RB925345	UFV/RIDESA
256	RB037158	RB912850	SP80-1816	UFAL/RIDESA
258	RB027057	SP85-3877	RB855536	UFV/RIDESA
259	IN84-73	E. ARUNDINACEUS	?	Estrangeira
260	RB047122	RB957510	?	UFV/RIDESA
261	RB9438	RB83160	RB855202	UFAL/RIDESA
262	RB987667	?	?	
263	SP83-2847	HJ5741	SP70-1143	Copersucar
264	RB008026	RB845197	?	UFV/RIDESA
265	RB008041	SP84-2025	SP80-3280	UFV/RIDESA
267	RB997504	SP82-3620	?	UFV/RIDESA
268	RB987932	RB72454	RB83102	UFV/RIDESA
271	RB987955	RB72454	RB739359	UFV/RIDESA
276	SP87-365	SP77-3067	?	Copersucar
277	RB027061	SP85-3877	RB855536	UFV/RIDESA
279	RB988082	RB83102	RB72454	UFV/RIDESA
281	RB988090	IAC86-2210	?	UFV/RIDESA
284	SP91-1049	SP80-3328	SP81-3250	Copersucar
285	RB008296	SP80-1816	RB855589	UFV/RIDESA
286	RB008300	SP80-1816	RB855589	UFV/RIDESA
287	SP88-819	SP71-6106	?	Copersucar
288	RB918625	?	?	
289	RB971754	F150	RB739359	UFAL/RIDESA

O objetivo, portanto,foi obter nesta população de genótipos, uma amostragem representativa, contemplando variabilidade para teor fibra e características morfo-anatômicas das plantas.

3.2. Preparo das amostras

3.2.1. Análise via úmida

Após a colheita o material foi levado ao Campo experimental de cana-de-açúcar (CECA), localizado em Oratórios – MG, pertencente à Universidade Federal de Viçosa.Os colmos foram submetidos à desintegração e homogeneização. Uma alíquota de 500 g foi submetida à prensa hidráulica, obtendo-se o caldo extraído.

Após a extração do caldo, o bagaço foi seco em estufa por 24 horas a 105 ^{o}C. Em seguida foi moído e separado através de uma malha de 0,4 mm. Este material foi utilizado para a determinação gravimétrica do teor de lignina.

Os extrativos da amostra (2 g de matéria seca) foram sucessivamente extraídos com solvente orgânico (etanol comercial) em uma unidade de extração de Soxhlet, por um período de 5 horas. Ao final, as amostras foram levadas à estufa (65 ^{o}C)para secagem. Tal procedimento é necessário para remover compostos que não fazem parte da biomassa e podem interferir na análise.

3.2.2. Análises espectrométricas no NIR

Para as análises espectrofotométricas, foram utilizados os mesmos genótipos selecionados para a determinação de lignina via úmida. Os espectros das amostras foram obtidos em diferentes locais e condições da matriz: bagaço seco com caldo, bagaço seco, colmo e folha. O objetivo, nesta etapa, foi encontrar o melhor material para realizar a previsão de lignina de forma rápida e com alta exatidão. Para o bagaço com caldo, os colmos dos genótipos foram submetidos à desintegração e homogeneização, sendo triturados juntamente com o caldo e armazenados. Já os espectros do bagaço seco foram obtidos com as mesmas amostras utilizadas para a determinação de lignina. Além disso, foram tomados os espectros no limbo foliar da Folha +3 de cada genótipo, de acordo com sistema Kuijper citado na revisão feita por Cheavegatti-Gianotto et al [98]. Para obtenção dos espetros utilizou-se o terço médio da folha excluindo-se a nervura central da mesma.Para o colmo, diferentes condições experimentais foram executadas, todas sem a necessidade de pré-tratamento: um seção do colmo, na parte superior (Terço Médio do Colmo) foi obtida e armazenada em freezer -80°C. Da mesma forma, foram obtidos amostras de colmo da região do meio (Terço Médio do Colmo) e pé (Terço Inferior do Colmo). Vale lembrar que o vetor y (variável dependente) foi construído com as mesmas amostras, para todas as condições

experimentais. Já a matriz X (variáveis independentes) foi diferente para cada condição (folha, colmo, bagaço seco e bagaço com caldo), de acordo com a figura abaixo.

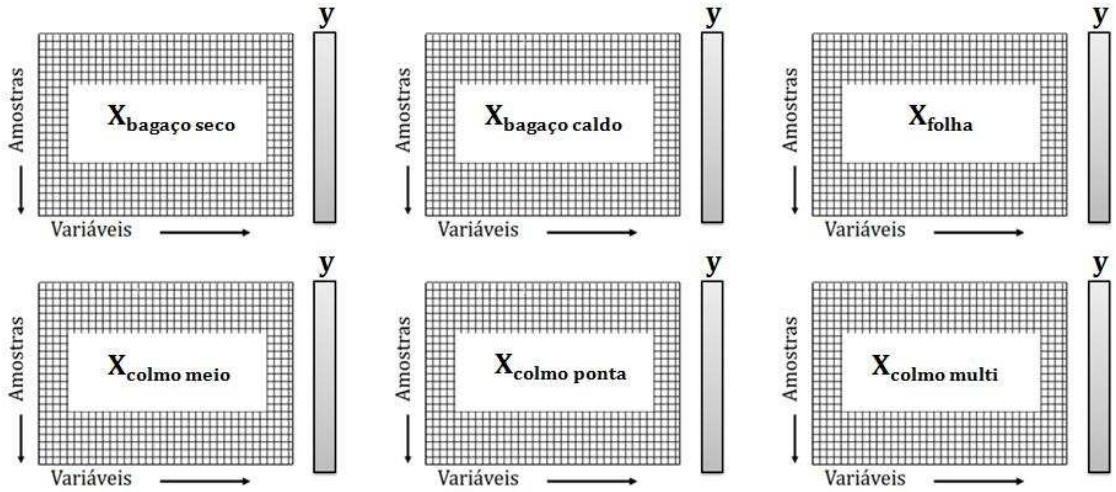


Figura 9. Esquema obtenção dos dados (**X** colmo meio = terço médio do colmo, **X** colmo ponta = terço superior do colmo e X multi = terço superior, médio e inferior do colmo).

3.3. Determinação das variáveis dependentes

Neste projeto foi utilizado o método descrito nos testes de NREL [99]para determinar a lignina da biomassa, conhecida como lignina Klason [100]. Neste método, uma massa triturada da biomassa (aproximadamente 0,3 g) é tratada com 3mL de solução de H_2SO_4 72% durante 2 h, à temperatura ambiente, para hidrolisar e solubilizar as frações de celulose e hemicelulose. A amostra é então diluída com 84 mL de água, para reduzir a concentração de ácido sulfúrico e autoclavada por um período de 1h. Em seguida, a lignina é deixada em repouso antes de ser filtrada. O resíduo é lavado com água quente até meio neutro. O resíduo é filtrado em cadinho filtrante e colocado na estufa por um período de 5h. A massa do resíduo seco insolúvel, representa o teor de lignina. As análises foram realizadas em duplicata.

3.4. Determinação das variáveis independentes

Os espectros NIR foram obtidos em um espectrômetro com transformada de Fourier (FT) Agilent 660 com auxílio do acessório de refletância usando esfera de integração adquirido da PIKE Technologies. Este acessório coleta a energia refletida a

partir de uma perspectiva esférica. A faixa investigada foi de 10000 a 4000 cm^{-1}com um incremento de 4 cm^{-1}. Os espectros foram obtidos através do software Resolutions Pro Versão 5.1, armazenando a informação como log(1/R), em que R é a refletância coletada. Para cada amostra um total de 64 varreduras foram realizadas e a média foi armazenada.

3.5. Construção dos modelos de calibração

Os espectros foram exportados para a extensão xls. e importados pelo Sofware Matlab7.8 (Math Works, Natick, USA). Uma matriz de dados com todos os espectros foi montada e denominada matriz **X**, que são as variáveis independentes. As linhas da matriz **X** correspondem às amostras e as colunas correspondem às variáveis (número de onda). Um vetor contendo os valores determinados de lignina foi construído e denominado de **y**, que é a variável dependente. O vetor **y** possui um número de linhas igual ao número de amostras na matriz **X**.

A matriz **X** e o vetor **y** foram importados pelo software MatLab 7.8. Para construção dos modelos foi utilizado a regressão PLS, aliada aos métodos GA, iPLS ou OPS. Os algoritmos para a construção e validação dos modelos foram escritos no Laboratório de Instrumentação e Quimiometria em função .m para Matlab. Todos os cálculos foram realizados no Software Matlab 7.8. O pacote computacional iToolbox foi usado para os cálculos iPLS. O pacote PLS-Toolbox 4.0 [101] para Matlab foi usado para o cálculo de seleção de variáveis usando algoritmo genético.

As variáveis **X** e **y** foram pré-processadas usando a centragem na média em todos os cálculos. Transformações foram realizadas nas linhas da matriz **X** com o objetivo de encontrar o melhor modelo para previsão. As transformações executadas foram: (1) alisamento usando o algoritmo Savitz Golay; (2) primeira derivada; (3) segunda derivada e (4) correção do espalhamento multiplicativo (MSC).

A qualidade dos modelos foi avaliada pela raiz quadrática média do erro (RMSE), o qual foi calculado de acordo com a Equação 15. R,o coeficiente de correlação, foi calculado pela Equação 16, em que $\hat{y}$ e $\hat{\mathbf{y}}$ são os escalares e vetor de valores estimados,respectivamente, $\bar{y}$ é um escalar dos valores médios de y, e I_m é o número das amostras (59):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{Im}\left(y_i - \hat{y}_i\right)^2}{I_m}} \quad (15)$$

$$R = \frac{\sum_{i=1}^{Im}\left(\hat{y}_i - \bar{\hat{y}}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{Im}\left(y_i - \hat{y}_i\right)^2\left(y_i - \bar{y}\right)^2}} \quad (16)$$

Quando a validação interna (cross validação-CV) é aplicada, I_m é o número de amostras no conjunto de calibração. Neste caso, o erro e o coeficiente de correlação são chamados RMSECV e Rcv, respectivamente. Para a validação externa (um novo conjunto de amostras), I_m é o número de amostras de previsão (P) e, neste caso, os coeficientes de correlação e o erro são nomeados Rp e RMSEP, respectivamente (59).

Para selecionar o número de variáveis latentes (h) do modelo, foi utilizado o método de validação cruzada com remoção aleatória de dez amostras. O gráfico número de componentes versus RMSECV foi utilizado para escolher o número de h.

A matriz de dados foi separada em dois conjuntos, um de calibração e outro de previsão. A construção propriamente dita dos modelos foi realizada usando o conjunto de calibração. Os valores de RMSECV e o coeficiente de correlação dos valores medidos e preditos na validação cruzada (Rcv) foram usados como parâmetros de verificação do ajuste do modelo. Uma vez realizada a regressão PLS usou-se o conjunto de previsão para verificar a capacidade preditiva do modelo e assim validar o modelo. Os valores de RMSEP e o coeficiente de correlação dos valores medidos e preditos na predição (Rp), bem como os valores dos erros relativos das amostras foram usados como parâmetro para verificar a capacidade de predição do modelo construído.A escolha do número de amostras do conjunto de previsão foi segundo a norma ASTM E1665 [86] que recomenda um total de 4 vezes o número de variáveis latentes (4h). Já a escolha das amostras para o conjunto de previsão e calibração foi realizada utilizando o algoritmo Kennard-Stone [102], que seleciona as amostras com base em suas distâncias. A primeira amostra selecionada pelo algoritmo é a que apresenta a maior distância em relação à amostra média. A segunda amostra a ser selecionada será a que apresentar maior distância em relação à primeira amostra selecionada. A próxima amostra a ser selecionada apresentará maior distância em relação à ultima amostra selecionada, e assim sucessivamente até atingir o número de amostras desejadas [48].

O GA foi empregado usando os seguintes parâmetros previamente otimizados: população = 54, gerações = 300, taxa de mutação = 0,008, largura da janela = 1, convergência = 80, inicialização = 50 e cross over = 2.

4. RESULTADOS E DISCUSSÕES

4.1. Bagaço seco com Caldo

Os espectros NIR das amostras (N = 232) de bagaço com caldo (variáveis independentes), na faixa de 4000 a 10000 cm^{-1}, com incremento de 4nm, são apresentados na figura abaixo:

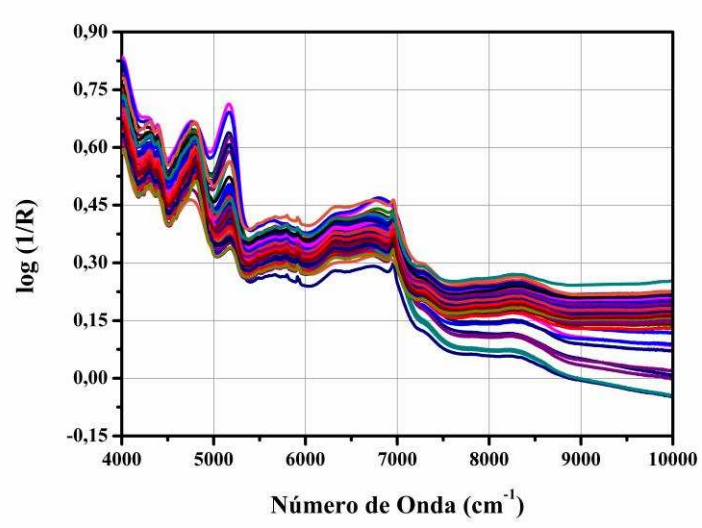


Figura 10. Espectro NIR bagaço com caldo

Os valores de lignina variaram na faixa de 18,35 a 28,37 % (m/m). Diversos trabalhos são encontrados na literatura [103-105] aplicando de forma eficaz a espectroscopia NIR aliada a métodos quimiométricos para determinação de lignina em biomassa. Porém, em todos os exemplos citados, a determinação é realizada no material seco e moído. A determinação direta de propriedades em bagaço de cana-de-açúcar com caldo é relativamente nova.

4.1.1. Construção do modelo

A Tabela 3 mostra os resultados para as diferentes transformações aplicadas, i.e., MSC, primeira derivada, segunda derivada. As colunas foram centradas na média (CM). O objetivo é obter o menor valor de RMSECV.

Tabela 3. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h= 10

	Modelo	
	RMSECV	**RMSEC**
MSC + CM	1,52	0,70
1ª Derivada + CM	1,51	0,90
2ª Derivada + CM	1,37	0,62

Analisando os valores de RMSECV e RMSEC apresentados na Tabela 3, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, representando pela Figura 11.

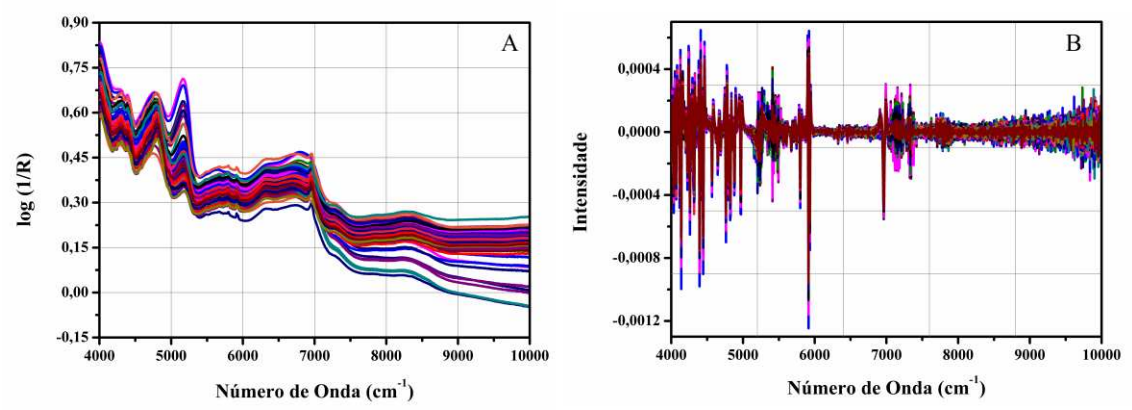


Figura 11. (A) Espectro original; (B) Espectro Derivada Segunda.

4.1.2. Seleção de Variáveis

Na construção dos modelos para determinação de lignina em bagaço com caldo, utilizaram-se previamente os algoritmos iPLS, biPLS, GA e OPS. Dessa forma, foi possível a seleção de regiões do espectro que apresentam informações relevantes e que melhor estão correlacionadas com a concentração de lignina em amostras de cana de açúcar.

Os parâmetros estatísticos calculados para todos os modelos estão representados na Tabela 4.

Tabela 4. Resultados estatísticos para os modelos.

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
h	10	10 (hOPS=19)	10	10	10
nVars	1038	265	346	520	352
RPD	1,42 (B)	2,77(A)	1,68(C)	1,51(C)	2,57(A)
RMSECV	1,31	0,71	1,20	1,22	0,74
RMSEC	0,30	0,32	0,33	0,38	0,32
Rc	0,99	0,99	0,99	0,98	0,99
Rcv	0,79	0,94	0,83	0,81	0,93
RMSEP	1,16	0,65	1,48	1,37	0,67
Rp	0,86	0,94	0,80	0,81	0,95
Viés	0,03	0,003	-0,093	-0,044	0,03
%ER	3,17	1,94	3,31	3,74	2,35

A Figura 12 compara os valores de RMSECV, RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

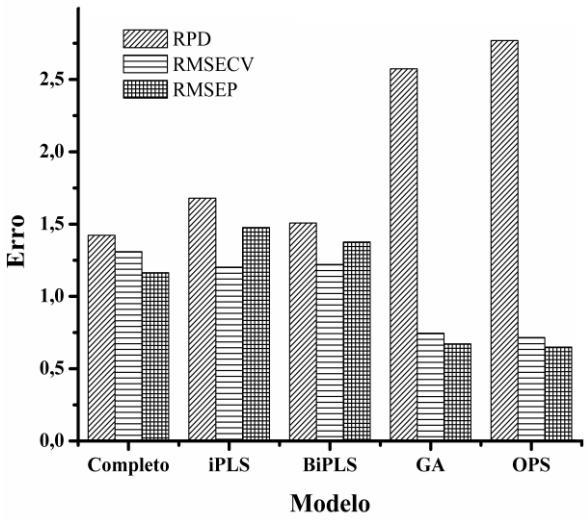


Figura 12. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos.

Analisando a Figura 12 e a Tabela 4, pode-se observar que para todos os modelos o viés não foi significativo, indicando que não há tendência nos resíduos. Além disso, a média dos erros relativos (% ER) foi pequena, em todos os casos, porém, erros altos de RMSEP e RMSECV para o modelo com todas as variáveis confirma a necessidade de se realizar seleção de variáveis em calibração multivariada. Os piores erros na previsão (RMSEP) foram obtidos quando se realizou seleção usando iPLS e biPLS. A razão deste mau desempenho pode estar ligada com o fato de que a lignina é uma molécula complexa e, conseqüentemente, é ativa em diferentes regiões do NIR.

Como o iPLS e o biPLS realizam seleção por intervalos, eles não funcionam bem quando a informação está espalhada por todo o espectro.

De modo diferente, uma comparação entre os algoritmos GA e OPS mostra que não houve diferença significativa nos valores de RMSEP e RMSECV, porém um maior valor de RPD foi observado para o OPS. Além disso,o algoritmo OPS selecionou um número menor de variáveis, tornando o modelo mais simples. Nesse contexto, é importante ressaltar o tempo computacional de ambos os métodos.Enquanto o OPS realizou os cálculos para estes conjuntos em minutos,o GA demorou horas para cada ensaio. Adicionalmente um tempo significativo foi consumido para otimizar os parâmetros usados no GA.

Dessa forma, o modelo utilizando o algoritmo OPS pode ser considerado mais eficiente e robusto. As variáveis selecionadas para o modelo OPS estão representadas na figura 13.

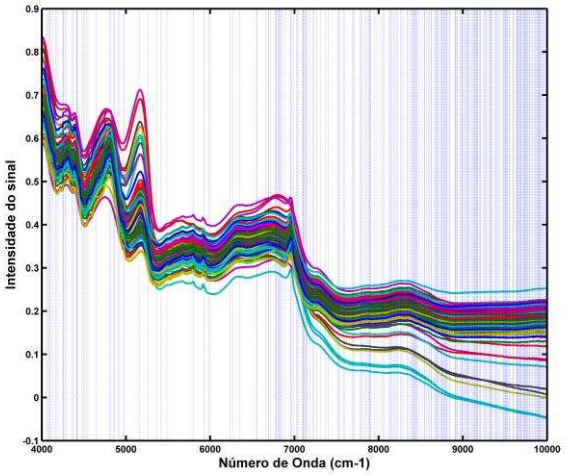


Figura 13. Variáveis selecionadas pelo algoritmo OPS

Os cálculos de validação/figuras de mérito foram realizados para o modelo OPS. As figuras de mérito calculadas para o modelo OPS são apresentadas na Tabela 5.

Tabela 5. Figuras de Mérito para os modelos Completo e OPS

	Bagaço Caldo- OPS	
	OPS	Completo
RPD	2,77	1,42
RMSEP	0,65	1,16
Rp	0,94	0,86
SEL	0,057	0,038
SEN.	$1{,}80\times10^{-5}$	$5{,}87\times10^{-5}$
γ	6,31	6,91
γ^{-1}	0,16	0,14
$\mathbf{LOD_{ruído}}$	7,49	8,55
$\mathbf{LOQ_{ruído}}$	22,72	25,91
$\mathbf{LOD_{prop}}$	3,69	3,65
$\mathbf{LOQ_{prop}}$	11,18	11,07

Os resultados apresentados na Tabela 5 indicam que o modelo apresenta alta capacidade para prever lignina em bagaço de cana-de-açúcar com caldo com alta exatidão em relação ao método de referência. O inverso da sensibilidade analítica γ-1mostrou que o modelo é sensível e apresentou maior seletividade (SEL). Dois métodos foram usados para o cálculo do limite de detecção e quantificação (LOD e LOQ). Um deles baseado no ruído e outro proposto por Teófilo, R. F (2007), que se baseia na extrapolação da regressão. Observa-se que o método proposto por Teófilo (2007) foi mais coerente, pois o valor de LOQ_{prop} ficou mais próximo do valor mínimo encontrado no conjunto de calibração.

A Figura 14 apresenta os valores de erro relativo para o conjunto calibração e previsão.

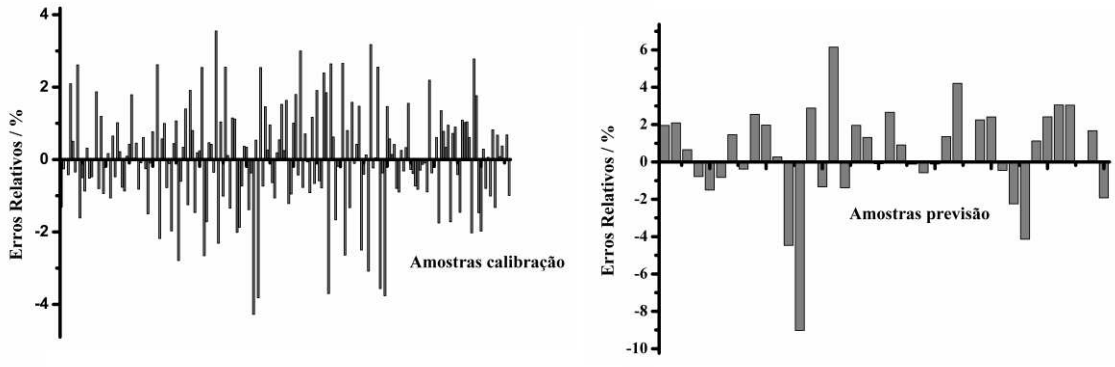


Figura 14. Erros relativos para o conjunto calibração e previsão

Pela análise do gráfico acima percebe-se o maior erro encontrado (em módulo) foi de 4,28% e o menor 0,01 % (conjunto calibração). Já para o conjunto previsão, os

valores variavam na faixa de (em módulo) 9,02% e 0,002%. Como os valores de erros relativos são pequenos, pode-se concluir que o modelo está apto para realizar previsão de lignina em cana-de-açúcar utilizando bagaço com caldo, com alta exatidão.

Para finalizar, os valores medidos versus preditos se encontram na Figura 15.

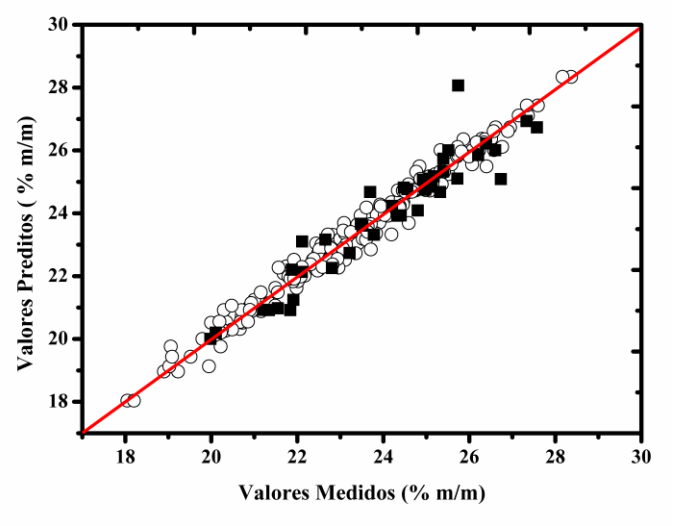


Figura 15. Valores medidos versus preditos. Círculos representam o conjunto calibração e quadrados representam o conjunto previsão.

4.2. Bagaço Seco

Os espectros NIR das amostras (N = 378) de bagaço seco (variáveis independentes), na faixa de 4000 a 10000 cm^{-1}, com incremento de 4nm, são apresentados na Figura 16.

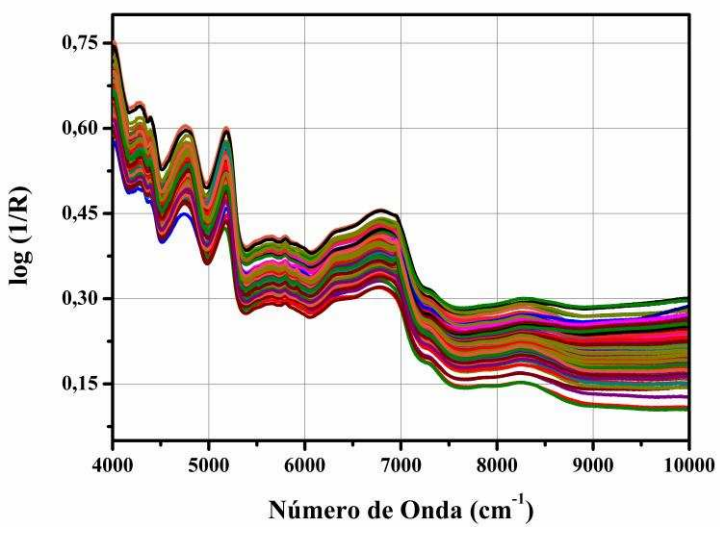


Figura 16. Espectro NIR bagaço seco.

Os valores da propriedade medida (lignina) variaram na faixa de 16,23 a 33,85%.

A faixa de variação do bagaço seco foi consideravelmente maior que a do bagaço com caldo pois, no primeiro, as análises foram realizadas em diferentes partes do colmo: pé, meio, ponta e colmo inteiro. Diferente do modelo do bagaço com caldo, que fornece o teor de lignina na cana inteira, o modelo do bagaço seco fornecerá uma estimativa do teor de lignina na cana inteira, como o bagaço com caldo, e de diferentes partes do colmo (pé, meio e ponta). O modelo pode então ser considerado multiproduto.

4.2.1. Construção do modelo

Conforme realizado para o bagaço com caldo, foram testados diferentes transformações, i.e., MSC, primeira derivada, segunda derivada. O pré-processamento realizado em todos os casos foi a centragem na média (CM).

Tabela 6. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h= 10

	Modelo	
	RMSECV	RMSEC
MSC + CM	1,88	1,73
1ª Derivada + CM	1,76	1,30
2ª Derivada + CM	1,39	0,67

Analisando os valores apresentados na Tabela 6, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, de acordo com a figura 17.

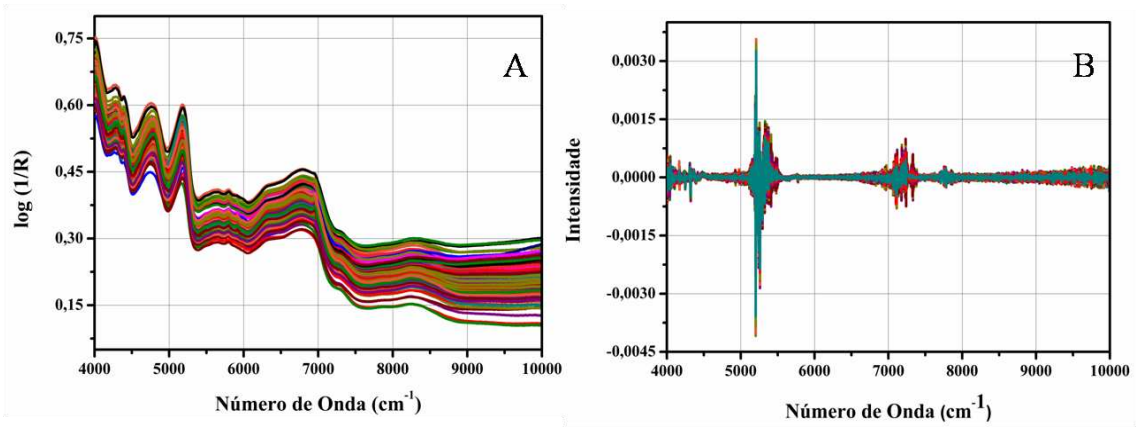


Figura 17. (A) Espectro original; (B) Espectro Derivada Segunda.

4.2.2. Seleção de Variáveis

Da mesma forma que foi realizado para o bagaço com caldo, os algoritmos iPLS, biPLS, OPS e GA foram usados para seleção de regiões do espectro que apresentam informações relevantes e que melhor estão correlacionadas com a concentração de lignina em amostras de cana de açúcar.

Os parâmetros estatísticos calculados para todos os modelos estão representados na tabela 7.

Tabela 7. Resultados estatísticos para os modelos

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
h	10	10 (hOPS=24)	10	10	10
nVars	1038	445	346	519	387
RPD	1,84 (B)	2,87 (A)	1,84 (B)	1,64 (B)	2,70(A)
RMSECV	1,42	0,89	1,43	1,58	0,95
RMSEC	0,53	0,44	0,76	0,63	0,50
Rc	0,98	0,99	0,96	0,97	0,98
Rcv	0,87	0,94	0,86	0,83	0,94
RMSEP	0,98	0,85	1,39	1,08	0,88
Rp	0,93	0,97	0,87	0,93	0,96
Viés	-0,052	-0,0066	0,018	-0,0004	-0,0033
%ER	3,22	2,82	4,9	3,48	2,7

A Figura 18 compara os valores de RMSECV, RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

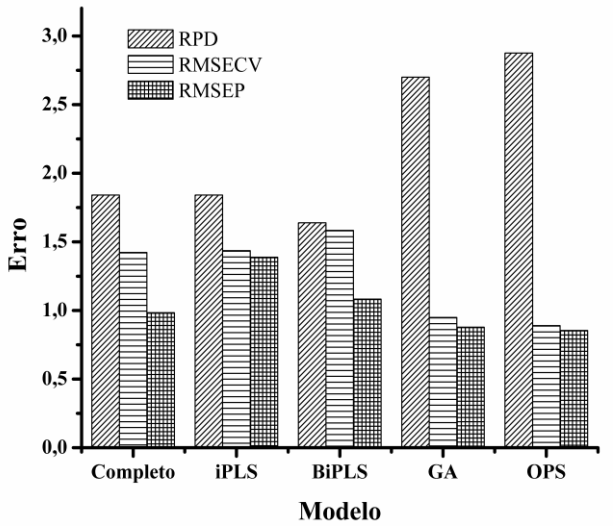


Figura 18. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos.

Analisando a Figura 18 e a Tabela 7, pode-se observar que,assim como o modelo bagaço com caldo, em todos os modelos o viés não foi significativo, indicando que não há tendência nos resíduos. Além disso, a média dos erros relativos (% ER) foi pequena, em todos os casos, porém, erros altos de RMSEP e RMSECV para o modelo com todas as variáveis. Os piores erros na previsão (RMSEP) foram obtidos quando se realizou seleção usando iPLS e biPLS.

Uma comparação entre os algoritmos GA e OPS mostra que não houve diferença significativa nos valores de RMSEP e RMSECV, porém um valor ligeiramente maior de RPD foi observado para o OPS. Além disso,o algoritmo OPS selecionou um número menor de variáveis, tornando o modelo mais simples. A Figura 19 é uma representação das variáveis selecionadas pelo algoritmo OPS.

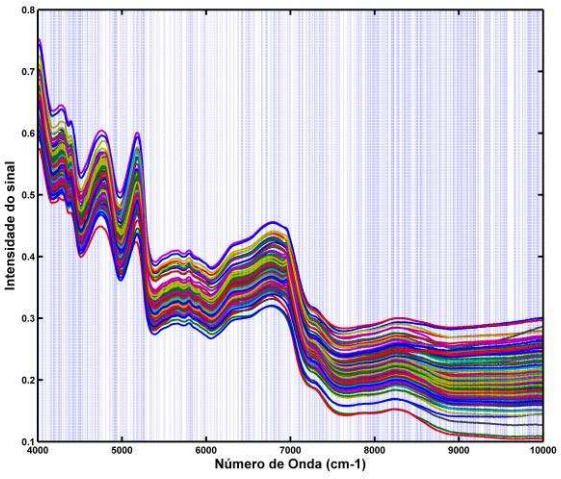


Figura 19. Variáveis selecionadas pelo algoritmo OPS

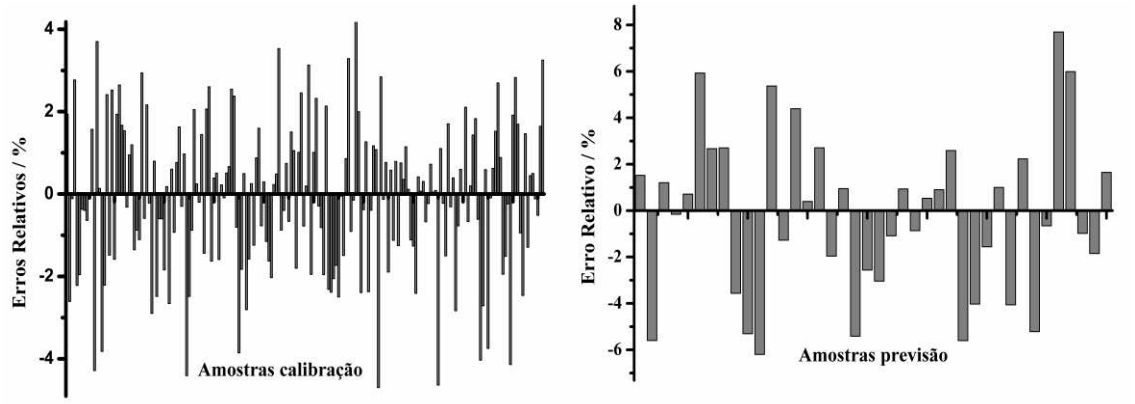
Dessa forma, o modelo OPS será escolhido para os cálculos das figuras de mérito, indicados na tabela abaixo.

Tabela 8. Figuras de mérito para os modelos Completo e OPS

	Bagaço Seco - OPS	
	OPS	Completo
RPD	2,87	1,84
RMSEP	0,85	0,98
Rp	0,97	0,93
SEL	0,0503	0,0373
SEN.	$1.22x10^{-5}$	$1,52x10^{-5}$
γ	1,72	1,77
γ^{-1}	0,58	0,56
$\mathbf{LOD_{ruido}}$	15,65	10,84
$\mathbf{LOQ_{ruido}}$	47,42	32,85
$\mathbf{LOD_{prop}}$	2,53	2,71
$\mathbf{LOQ_{prop}}$	7,67	8,22

Os resultados apresentados na Tabela 8 indicam que o modelo apresenta alta capacidade para prever lignina em bagaço de cana-de-açúcar seco, em diferentes partes (pé, meio e ponta) com alta exatidão em relação ao método de referência. O inverso da sensibilidade analítica γ^{-1} mostrou que o modelo é sensível e apresentou maior seletividade (SEL). Pelo análise dos valores de LOD e LOQ, conclui-se, mais uma vez, que o método proposto por Teófilo foi mais coerente, pois o valor de $LOQ_{ruído}$ é maior que o menor valor de teor encontrado através das analises via úmida.

A Figura 20 apresenta os valores de erro relativo para o conjunto calibração e previsão.



F**igura 20**. Erros relativos para o conjunto calibração e previsão.

Pela análise do gráfico acima percebe-se o maior erro encontrado (em módulo) foi de 5,57% e o menor 0,03 % (conjunto calibração). Já para o conjunto previsão, os valores variavam na faixa de (em módulo) 7,70% e 0,16%. Como os valores de erros relativos são pequenos, pode-se concluir que o modelo está apto para realizar previsão de lignina em cana-de-açúcar utilizando bagaço seco, com alta exatidão.

Para finalizar, o gráfico dos valores medidos versus preditos para o conjunto calibração e previsão.

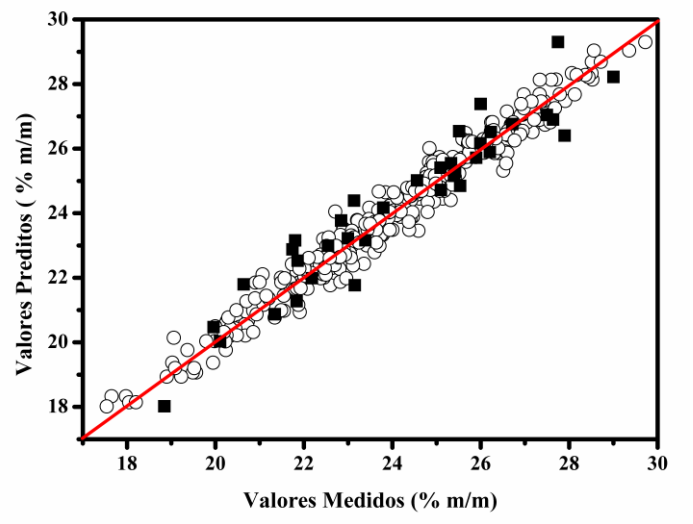


Figura 21. Valores medidos x preditos. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

4.3. Folha

Os espectros NIR das amostras (N = 256) de folha (variáveis independentes), na faixa de 4000 a 10000 cm^{-1}, com incremento de 4nm, são apresentados na Figura 22.

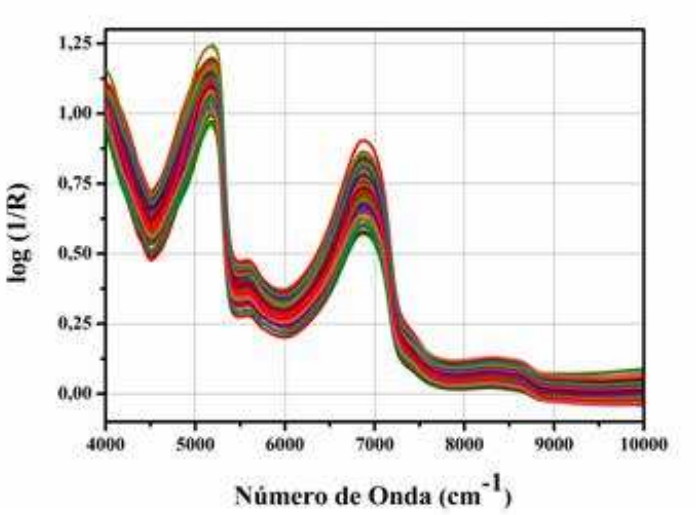


Figura 22. Espectro NIR folha.

Os valores da propriedade medida (lignina) variaram na faixa de 18,05 a 28,37%. Para o modelo da folha, as estimativas serão do teor de lignina na cana inteira. Diferentes trabalhos são encontrados na literatura relacionando o espectro NIR [104, 106-107] com alguma propriedade da folha. Nesses casos, a folha é seca e triturada, diferente do trabalho proposto em que o espectro é obtido diretamente sobre a folha, sem a necessidade de preparo da amostra. Menesatti e colaboradores [106] avaliaram propriedades nutricionais de laranja obtendo o espectro diretamente sobre a folha seca. Neste caso, há uma correlação entre uma propriedade da folha e o espectro da mesma. O trabalho proposto se dedica a correlacionar o espectro da folha (obtido sem nenhum preparo da amostra) com o conteúdo de lignina do bagaço, considerando o mesmo genótipo.

4.3.1 Construção do modelo

Conforme realizado para os outros modelos, foram testados diferentes transformações, com o objetivo de obter o menor valor de RMSECV. O pré-processamento realizado em todos os casos foi a centragem na média (CM).

Tabela 9. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h = 10

	Modelo	
	RMSECV	RMSEC
MSC + CM	1,87	1,70
1ª Derivada + CM	1,67	1,17
2ª Derivada + CM	1,22	0,57

Analisando os valores de RMSECV e RMSEC apresentados na Tabela 9, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, representando pela Figura 21.

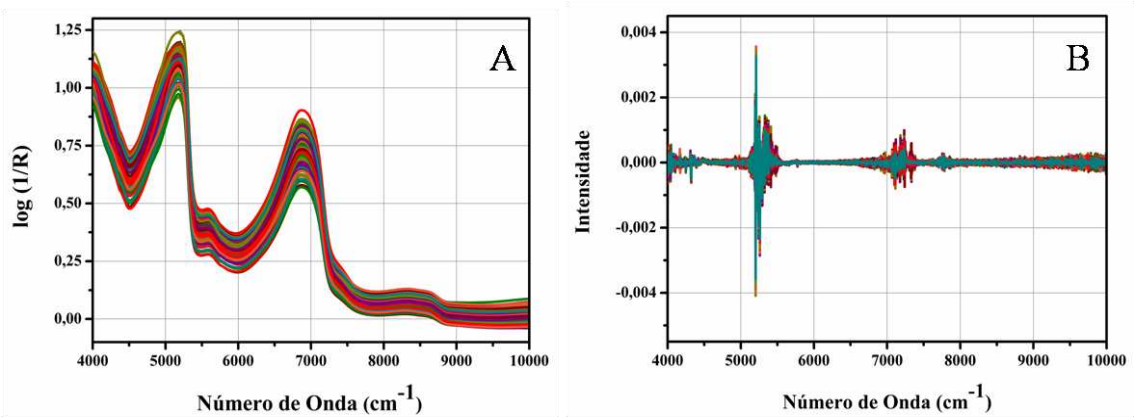


Figura 23. (A) Espectro original; (B) Espectro Derivada Segunda.

4.3.2 Seleção de Variáveis

Da mesma forma que foi realizada para os demais conjuntos de dados, diferentes algoritmos de seleção de variáveis foram testados (iPLS, biPLS, OPS e GA). Os parâmetros estatísticos calculados para todos os modelos estão representados na tabela 10.

Tabela 10. Resultados estatísticos para os modelos

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
h	10	10 (hOPS=25)	10	10	10
nVars	1038	305	346	519	324
RPD	1,41 (B)	2,56 (A)	1,15 (C)	1,17 (C)	2,72 (A)
RMSECV	1,32	0,76	1,84	1,55	0,72
RMSEC	0,43	0,35	0,54	0,69	0,35
Rc	0,98	0,98	0,97	0,94	0,99
Rcv	0,78	0,93	0,62	0,69	0,94
RMSEP	0,77	0,67	0,99	1,01	0,64
Rp	0,93	0,96	0,90	0,91	0,95
Viés	-0,015	0,017	0,003	-0,045	0,0043
%ER	2,62	2,47	3.47	3,27	2,30

A Figura 24 compara os valores de RMSECV, RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

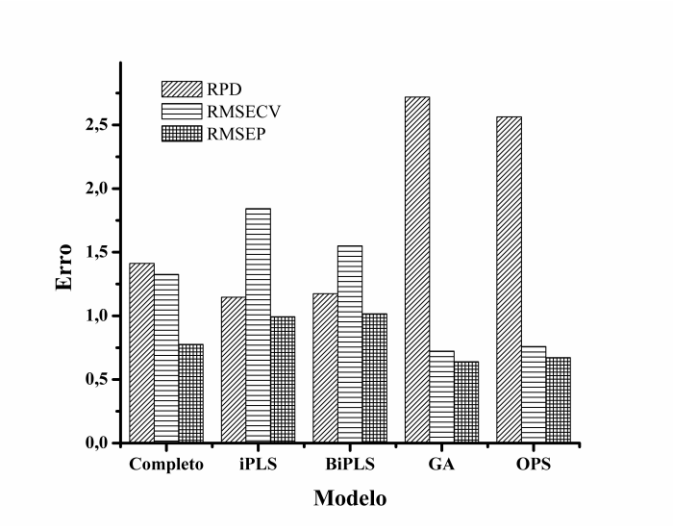


Figura 24. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos

Pela análise da tabela e do gráfico, percebe-se que os algoritmos OPS e GA foram mais eficientes, se comparados com os demais e com o modelo completo. Nesse exemplo, os parâmetros estatísticos para o modelo GA (principalmente RMSECV, RPD e RMSEP) foram ligeiramente melhores. Se analisar de maneira mais detalhada, percebe-se que não há diferença significativa entre os resultados obtidos para os dois algoritmos. Como o algoritmo OPS selecionou um conjunto menor de variáveis em um intervalo consideravelmente menor de tempo, o modelo OPS é o escolhido para futuras

previsões. As variáveis selecionadas pelo algoritmo OPS estão representadas na Figura 25.

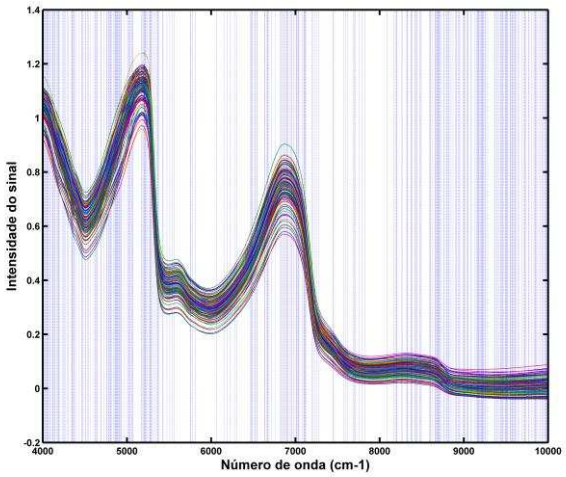


Figura 25. Variáveis selecionadas pelo algoritmo OPS.

Os parâmetros estatísticos foram calculados para o modelo OPS e estão representados na Tabela 11.

Tabela 11. Figuras de mérito para os modelos Completo e OPS

	Folha – OPS	
	OPS	Completo
RPD	2,56	1,41
RMSEP	0,67	0,77
Rp	0,96	0,93
SEL	0,0571	0,0422
SEN.	$3{,}78 \times 10^{-5}$	$4{,}37 \times 10^{-5}$
γ	8,69	12,31
γ^{-1}	0,11	0,081
LODruido	12,72	26,84
LOQruido	38,55	81,33
LODprop	3,74	4,10
LOQprop	11,33	12,43

A análise dos parâmetros estatísticos confirma, mais um vez, a eficiência no cálculo do LOD e LOQ proposto por Teófilo. O $LOQ_{ruído}$ encontrado foi consideravelmente superior ao mínimo encontrado via análise Klason. O inverso da sensibilidade analítica γ^{-1} mostrou que o modelo é sensível e apresenta alta seletividade

(SEL). Dessa forma, através da obtenção do espectro da folha de um genótipo (sem qualquer pré-tratamento), é possível obter, com alta exatidão, o teor de lignina de uma amostra de cana-de-açúcar.

A Figura 26 apresenta os valores de erro relativo para o conjunto calibração e previsão.

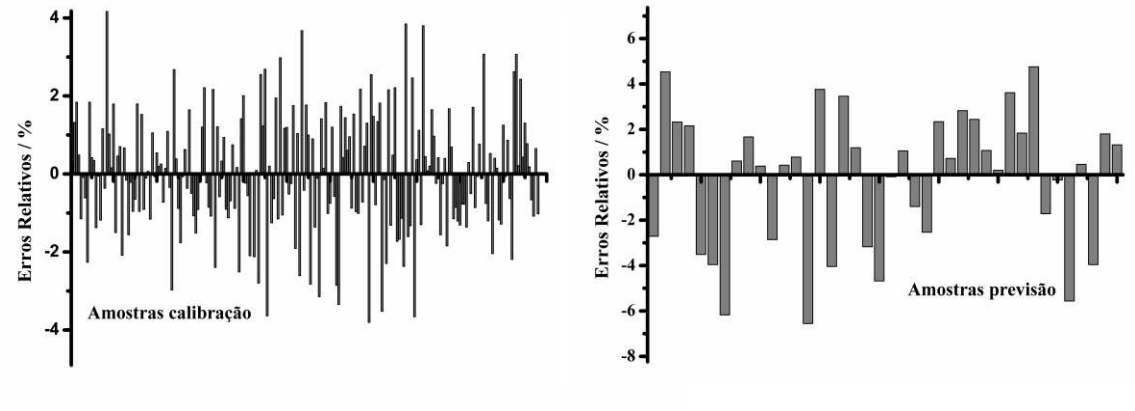


Figura 26. Erros Relativo (%) para o conjunto calibração e conjunto previsão.

Pela Figura 26, percebe-se que a faixa de erro relativo obtida é (em módulo) 8,94 - 0,04% (conjunto previsão) e 9,06 – 0,07% (conjunto previsão). Dessa forma, comprova-se que é viável realizar a previsão do teor de lignina, em cana-de-açúcar, através do espectro da folha, reduzindo, significativamente, o tempo gasto com as análises e o consumo de reagentes químicos.

Finalmente, a Figura 27 contém os valores medidos versus preditos para o conjunto calibração e previsão.

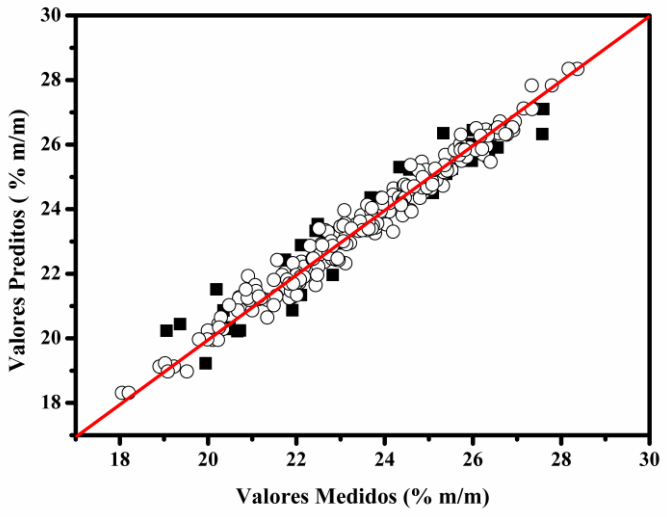


Figura 27. Valores medidos x preditos. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

4.4 Terço Médio do Colmo

Os espectros NIR das amostras (N = 221) de colmo (meio), na faixa de 4000 a 10000 cm^{-1}, com incremento de 4nm, são apresentados na Figura 28.

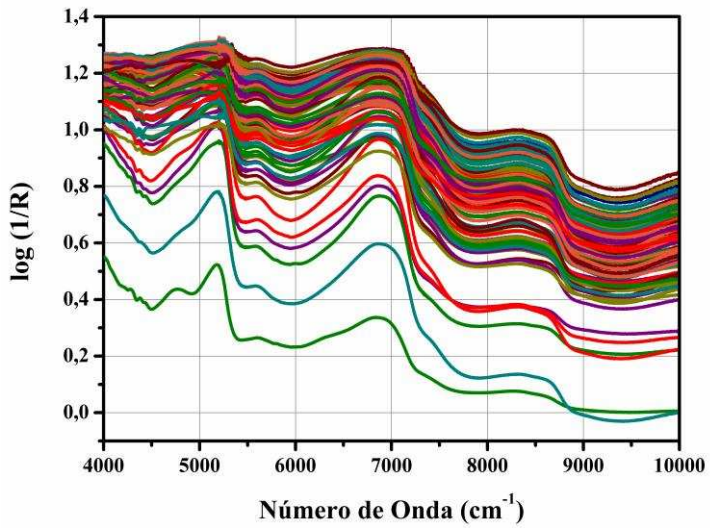


Figura 28. Espectro NIR Colmo Meio.

Os valores da propriedade medida (lignina) variaram na faixa de 18,05 a 28,37%. Diferentes trabalhos na literatura [108-109] já exploraram a construção de modelos multivariados utilizando o espectro obtido diretamente sobre o colmo. Porém, a determinação rápida de lignina obtida a partir de espectros NIR coletados diretamente sobre o colmo é, pelo nosso conhecimento, apresentado pela primeira vez neste trabalho.

O objetivo de se construir um modelo utilizando o espectro do colmo (meio) está na facilidade em se obter o espectro. Para realizar a análise, a preparação da amostra é mínima. Diferente do bagaço seco que necessita de moer, peneirar e secar a amostra, o colmo, na região do meio, é apenas cortado da cana e levado ao laboratório para análise. Dessa forma, pode-se obter centenas de resultados em um único dia.

4.4.1Construção do modelo

Os valores de RMSECV e RMSEC obtidos para as diferentes transformações, no espectro, estão indicados na Tabela 12.

Tabela 12. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h = 8

	Modelo	
	RMSECV	RMSEC
MSC + CM	2,00	1,82
1ª Derivada + CM	1,99	1,72
2ª Derivada + CM	0,87	1,55

Analisando os valores de RMSECV e RMSEC apresentados na Tabela 12, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, representando pela Figura 29.

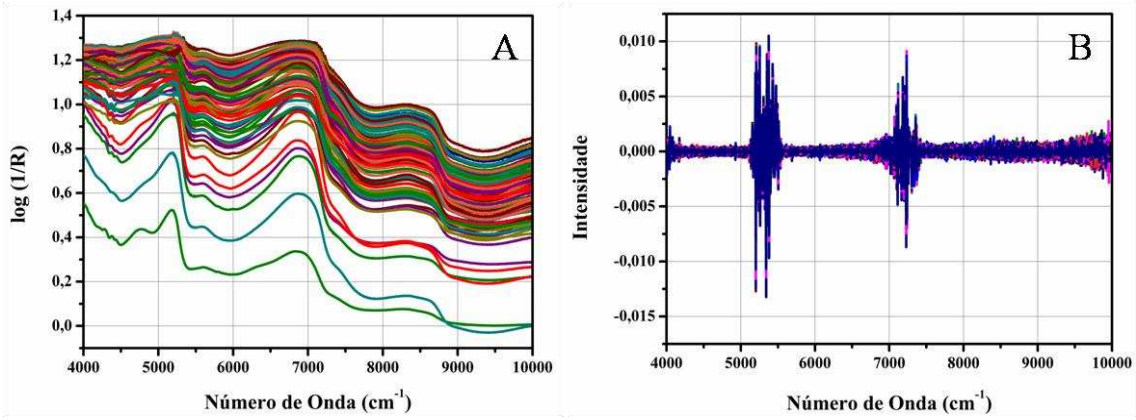


Figura 29. (A) Espectro original; (B) Espectro Derivada Segunda

4.4.2 Seleção de Variáveis

Os parâmetros estatísticos obtidos para os diferentes algoritmos (iPLS, biPLS, OPS e GA) são apresentados na Tabela 13.

Tabela 13. Resultados estatísticos para os modelos

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
H	8	8 (hOPS=16)	8	8	8
Nvars	1038	205	346	519	338
RPD	1,77 (B)	3,24 (A)	1,38 (C)	1,31 (C)	2,16 (A)
RMSECV	1,21	0,63	1,76	1,57	0,87
RMSEC	0,30	0,31	0,33	0,38	0,35
Rc	0,99	0,99	0,99	0,98	0,99
Rcv	0,84	0,96	0,71	0,72	0,91
RMSEP	0,73	0,61	0,93	0,90	0,62
Rp	0,90	0,95	0,87	0,89	0,94
Viés	0,017	0,018	-0,083	-0,0099	-0,039
%ER	2,52	1,97	2,82	2,83	2,00

A Figura 30 compara os valores de RMSECV,RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

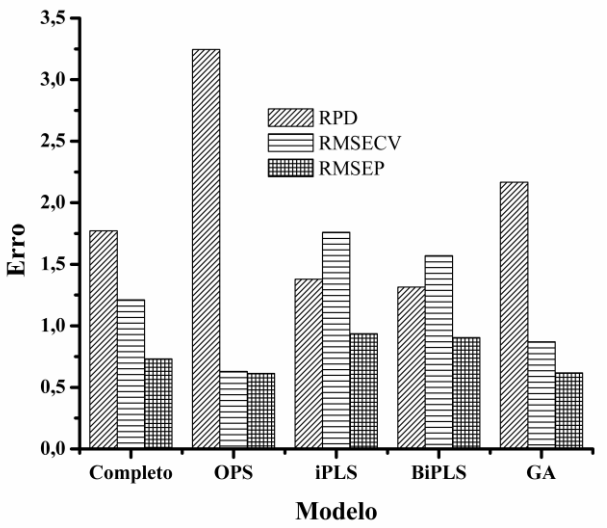


Figura 30. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos.

Conforme todos os exemplos anteriores, o viés não foi significativo para nenhum dos modelos. Analisando o gráfico acima, é fácil notar que o algoritmo OPS se destaca, selecionando um número menor de variáveis com maior capacidade de previsão. Percebe-se que houve melhoria significativa em todos os parâmetros estatísticos utilizando o algoritmo OPS. As variáveis selecionadas pelo algoritmo OPS estão representadas na Figura 31.

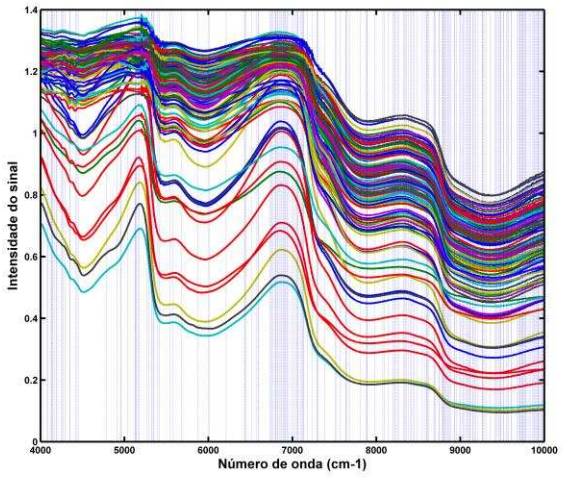


Figura 31. Variáveis selecionadas pelo algoritmo OPS

Dessa forma, os cálculos das figuras de mérito realizados no modelo OPS estão na Tabela 14.

Tabela 14. Figuras de Mérito para os modelos Completo e OPS

	Colmo M - OPS	
	OPS	Completo
RPD	3,24	1,77
RMSEP	0,61	0,73
Rp	0,95	0,90
SEL	0,062	0,039
SEN.	$2{,}00 \times 10^{-4}$	$8{,}90 \times 10^{-4}$
γ	36,34	10,49
γ^{-1}	0,0275	0,0953
LODruido	12,57	17,38
LOQruido	38,1	52,69
LODprop	3,75	3,66
LOQprop	11,38	11,10

Pela análise das figuras de mérito, pode-se compreender que a obtenção do espectro do colmo, na parte média, sem qualquer pré-tratamento, fornece o teor de lignina da cana, com alta exatidão. O inverso da sensibilidade analítica γ^{-1}mostrou que o modelo é sensível e apresentou alta seletividade (SEL). Conforme observado para os demais conjuntos de dados, o cálculo de LOQ e LOD foi mais eficiente utilizando o método proposto.

A Figura 32 apresenta os valores de erro relativo para o conjunto calibração e previsão.

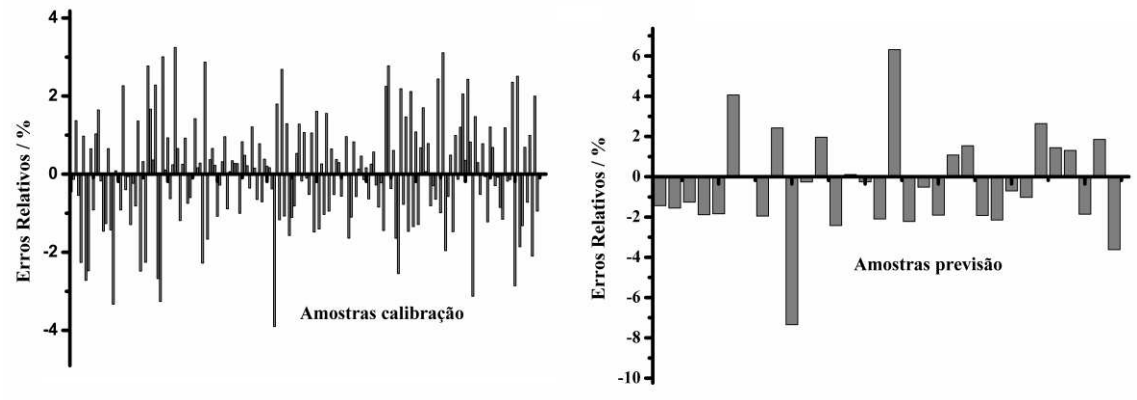


Figura 32. Erros Relativo (%) conjunto calibração e conjunto previsão.

Para o conjunto calibração, a faixa de erro relativo obtida (em módulo) foi 4,66 - 0,002%. Já para o conjunto previsão, a faixa é de 12,05-0,08%. Como os erros relativos encontrados foram pequenos, pode-se concluir que o modelo OPS-Colmo M é eficaz para prever lignina em cana-de-açúcar.

A Figura 33 apresenta os valores medidos versus preditos para o conjunto de dados (calibração e previsão).

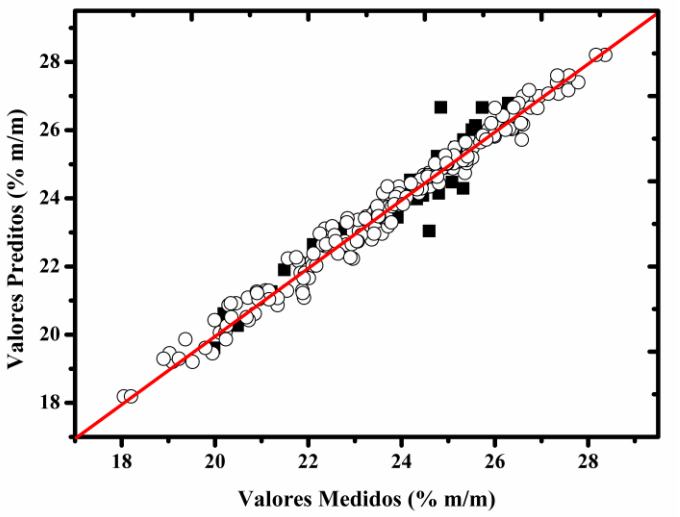


Figura 33. Valores medidos x preditos. Bolas representam o conjunto calibração e quadrados o conjunto previsão.

4.5 Terço Superior do Colmo

Os espectros NIR das amostras (N = 223) de colmo (ponta), na faixa de 4000 a 10000 cm-1, com incremento de 4nm, são apresentados na Figura 34.

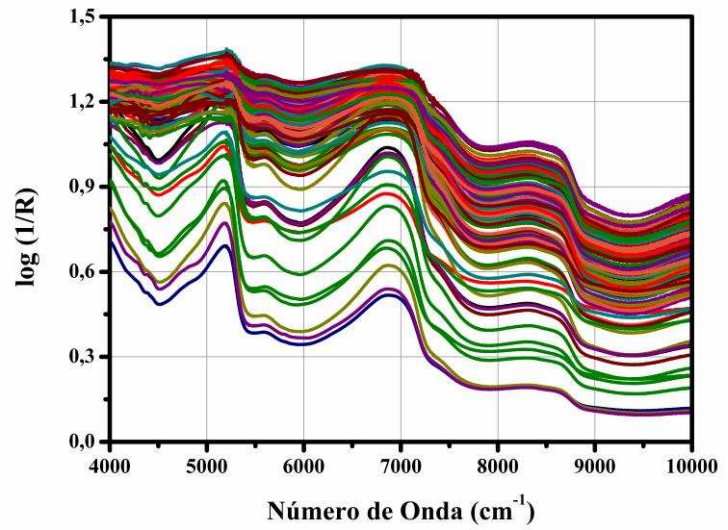


Figura 34. Espectro NIR Colmo Ponta.

Os valores da propriedade medida (lignina) variaram na faixa de 18,05 a 28,37%.

De forma semelhante ao colmo meio, o objetivo de se construir um modelo utilizando o espectro do colmo (ponta) está na facilidade na obtenção do espectro. Para realizar a análise, a preparação da amostra é mínima. Dessa forma, pode-se obter centenas de resultados em um único dia, sem gasto excessivo de reagentes.

4.5.1 Construção do modelo

A Tabela 15 indica os valores de RMSECV e RMSEC obtidos quando utilizado diferentes transformações.

Tabela 15. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h= 10

	Modelo	
	RMSECV	RMSEC
MSC + CM	1,95	1,42
1ª Derivada + CM	1,90	1,41
2ª Derivada + CM	1,27	0,63

Analisando a Tabela acima, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, representando pela Figura 35.

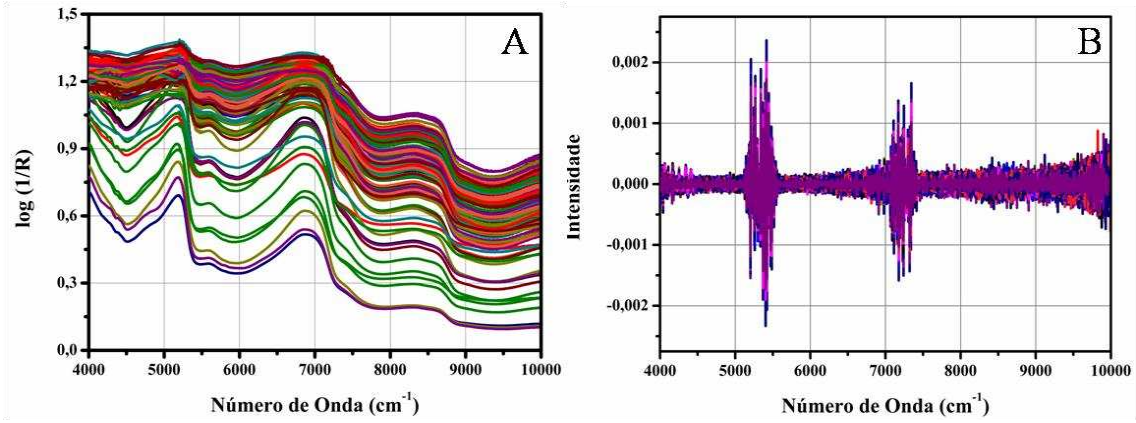


Figura 35. (A) Espectro original; (B) Espectro Derivada Segunda.

4.5.2 Seleção de Variáveis

Os parâmetros estatísticos calculados para todos os modelos (Completo, iPLS, biPLS, OPS e GA) estão representados na tabela 16.

Tabela 16. Resultados estatísticos para os modelos

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
h	10	10 (hOPS=15)	10	10	10
nVars	1038	300	346	519	357
RPD	1,47 (B)	2,33 (A)	1,21 (C)	1,43 (B)	2,55 (A)
RMSECV	1,43	0,89	2,10	1,58	0,83
RMSEC	0,29	0,30	0,37	0,38	0,33
Rc	0,99	0,99	0,98	0,98	0,99
Rcv	0,78	0,91	0,61	0,74	0,93
RMSEP	0,58	0,58	0,65	0,65	0,57
Rp	0,96	0,96	0,95	0,95	0,96
Viés	0,019	0,045	-0,088	-0,065	0,051
%ER	2,01	1,94	2,3	1,98	1,91

A Figura 36 compara os valores de RMSECV, RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

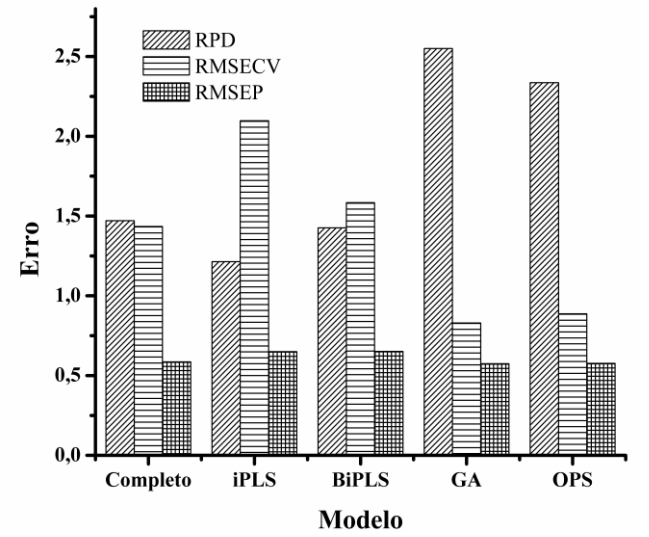


Figura 36. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos.

Pela análise da tabela e do gráfico acima, percebe-se que, da mesma forma que observado para o conjunto colmo meio, os algoritmos OPS e GA foram mais eficientes, se comparados com os demais e com o modelo completo. Nesse exemplo, os parâmetros estatísticos para o modelo GA (principalmente RMSECV, RPD e RMSEP) foram ligeiramente melhores. Se analisar de maneira mais detalhada, percebe-se que não há

diferença significativa entre os resultados obtidos para os dois algoritmos. Como o algoritmo OPS selecionou um conjunto menor de variáveis em um intervalo consideravelmente menor de tempo, o modelo OPS será escolhido para futuras previsões. As variáveis selecionadas pelo algoritmo OPS estão representadas na Figura 37.

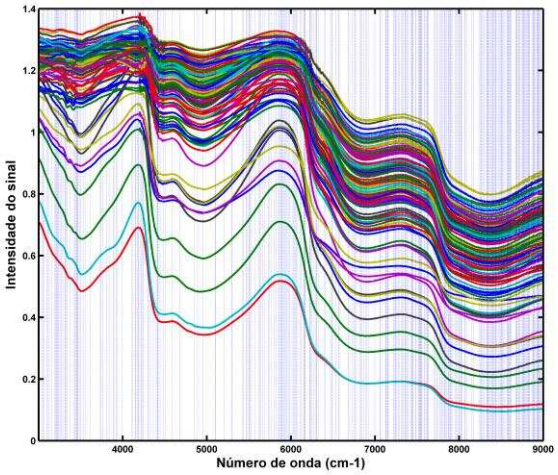


Figura 37. Variáveis selecionadas pelo algoritmo OPS.

Os parâmetros estatísticos foram calculados para o modelo OPS e estão representados na tabela 17.

Tabela 17. Figuras de Mérito para os modelos Completo e OPS

	Colmo PT-OPS	
	OPS	Completo
RPD	2,33	1,47
RMSEP	0,57	0,59
Rp	0,96	0,96
SEL	0,057	0,042
SEN.	$7{,}6\text{x}10^{-5}$	$6{,}9\text{x}10^{-5}$
γ	8,54	10,49
γ^{-1}	0,12	0,09
LODruido	10,35	17,38
LOQruido	31,36	52,69
LODprop	3,58	3,66
LOQprop	10,84	11,10

Os resultados apresentados na Tabela 17 indicam que o modelo apresenta alta capacidade para prever lignina em bagaço de cana-de-açúcar com alta exatidão em

relação ao método de referência. O inverso da sensibilidade analítica γ-1mostrou que o modelo é sensível e apresentou maior seletividade (SEL). Vale ressaltar, mais uma vez, que os resultados dos parâmetros LOQ e LOD foram mais satisfatórios, quando utilizado o método proposto.

A Figura 38 apresenta os valores de erro relativo para o conjunto calibração e previsão.

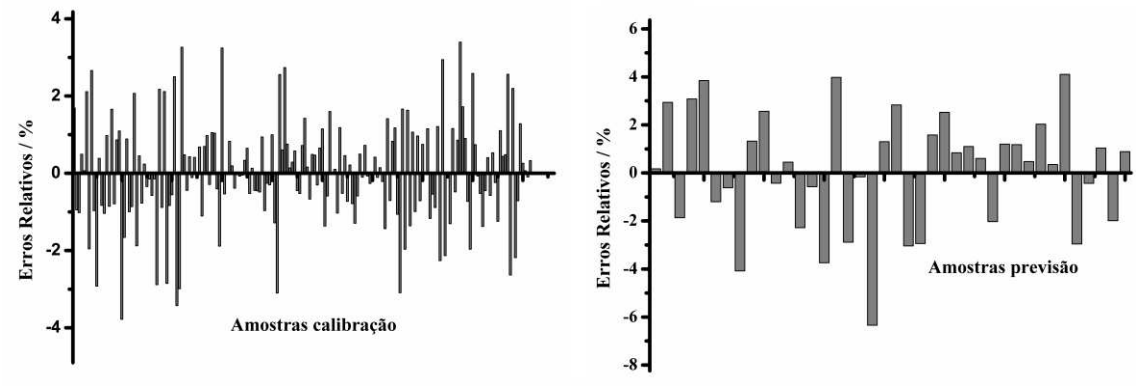


Figura 38. Erros Relativo (%) conjunto calibração e conjunto previsão.

Os valores de erro relativo encontrados para o conjunto calibração e previsão foram, respectivamente, 6,18– 0,009% e 6,53 – 0,02%. Tais valores confirmam a habilidade do modelo em prever lignina em cana-de-açúcar com alta exatidão.

A Figura 39 apresenta os valores medidos versus preditos para o conjunto calibração e previsão.

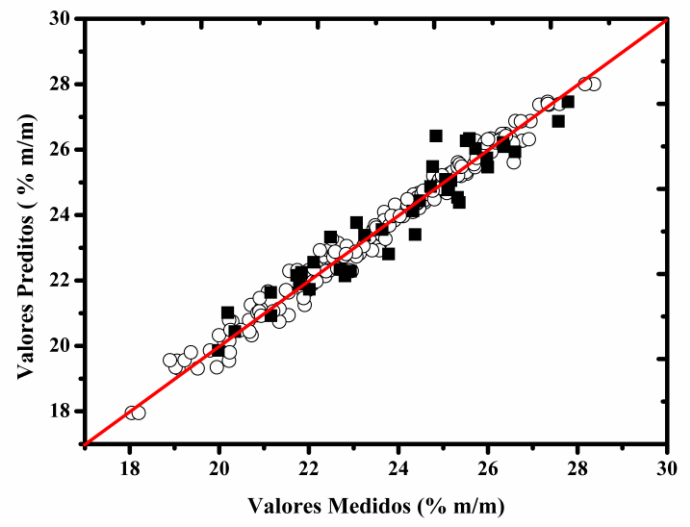


Figura 39. Valores medidos x preditos. Círculos representam o conjunto calibração e quadrados o conjunto previsão.

4.7. Terço Inferior, Médio e Superior do Colmo

Os espectros NIR das amostras (N = 150) de colmo (pé, meio e ponta), na faixa de 4000 a 10000 cm-1, com incremento de 4nm, são apresentados na Figura 40.

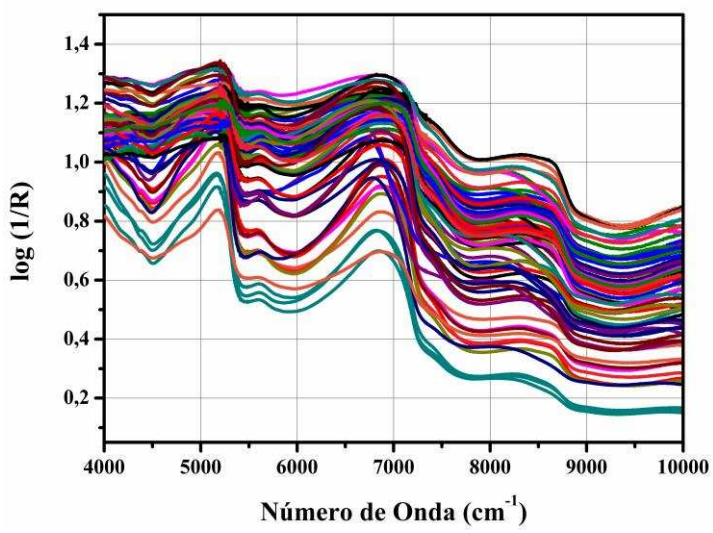


Figura 40. Espectro NIR Terço Inferior, Médio e Superior do Colmo.

Os valores da propriedade medida (lignina) variaram na faixa de 14,86 a 33,85%. A faixa de valores para o conjunto de dados das diferentes partes do terço do colmo é maior pois foi realizado análise química nestas três diferentes partes do colmo. Dessa

forma, neste modelo, teremos informações do teor de lignina no pé, meio e ponta sendo classificado, portanto, como um modelo multiproduto.

4.7.1 Construção do modelo

Os valores de RMSECV e RMSEC para as diferentes transformações i.e., MSC, primeira derivada, segunda derivada estão representados na Tabela 18.

Tabela 18. Valores de RMSECV e RMSEC para diferentes pré-processamentos usando h = 7

	Modelo	
	RMSECV	RMSEC
MSC + CM	2,94	1,49
1ª Derivada + CM	2,08	1,04
2ª Derivada + CM	1,70	0,67

Analisando os valores de RMSECV e RMSEC apresentados na Tabela 18, verifica-se que o melhor pré-processamento foi a segunda derivada. Logo, o modelo será construído com base neste pré-processamento, representando pela Figura 41.

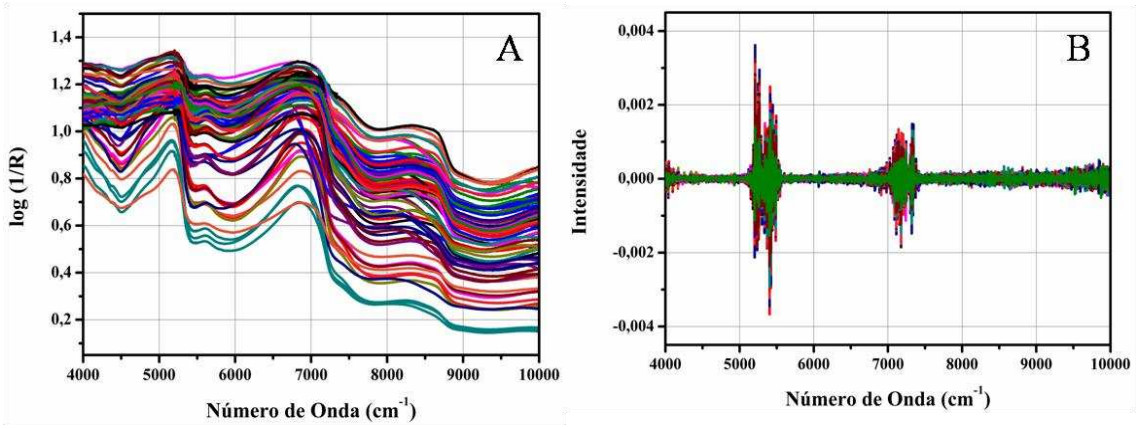


Figura 41. (A) Espectro original; (B) Espectro Derivada Segunda.

4.7.2 Seleção de Variáveis

Os parâmetros estatísticos calculados para os diferentes modelos (iPLS, biPLS, OPS e GA) estão representados na tabela 19.

Tabela 19.Resultados estatísticos para os modelos

	Modelos				
	Completo	**OPS**	**iPLS**	**biPLS**	**GA**
h	7	7 (hOPS=20)	7	7	7
nVars	1038	250	346	518	336
RPD	1,66 (A)	2,79 (A)	1,30 (C)	1,14 (C)	1,20 (A)
RMSECV	2,25	1,31	2,66	3,14	1,19
RMSEC	0,58	0,51	1,04	0,46	0,42
Rc	0,99	0,99	0,97	0,99	0,99
Rcv	0,84	0,95	0,77	0,68	0,96
RMSEP	0,93	0,80	1,58	0,87	0,91
Rp	0,98	0,99	0,94	0,94	0,99
vies	-0,06	-0,01	-0,08	-0,23	-0,02
%ER	2,90	2,90	5,29	2,79	2,94

A Figura 42 compara os valores de RMSECV, RPD e RMPSEP para os modelos obtidos com os diferentes algoritmos de seleção de variáveis.

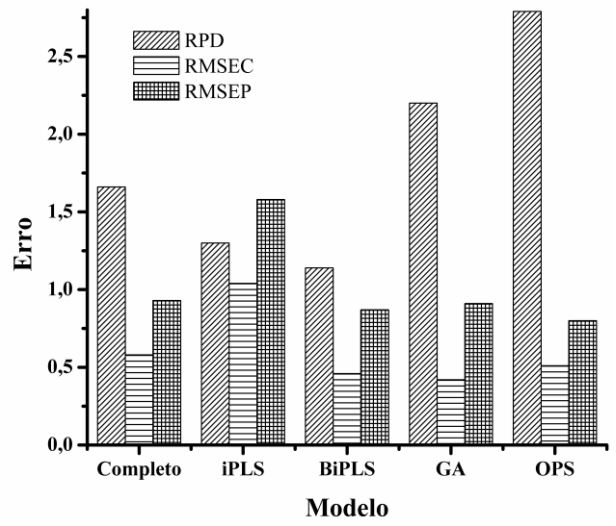


Figura 42. Comparação de RMSECV, RPD e RMSEP obtidos para os diferentes modelos.

Pela análise da tabela e do gráfico acima, percebe-se que, a utilização do algoritmo OPS melhorou, consideravelmente, todos os parâmetros estatísticos. Como o algoritmo OPS selecionou um conjunto menor de variáveis em um intervalo consideravelmente menor de tempo, as figuras de mérito foram calculados para o

modelo OPS. As variáveis selecionadas pelo algoritmo OPS estão representadas na Figura 43.

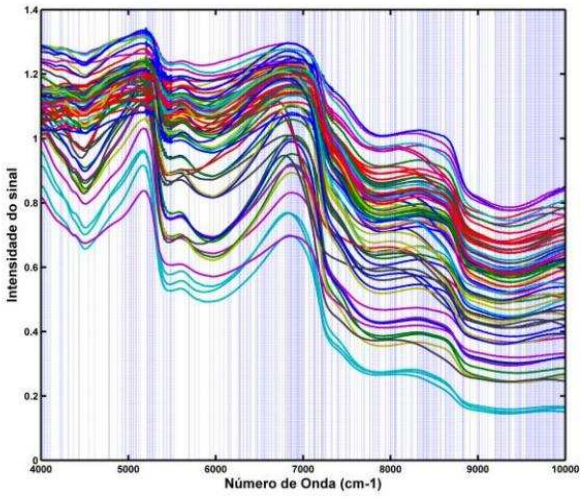


Figura 43. Variáveis selecionadas pelo algoritmo OPS

As figuras de mérito calculadas para o modelo OPS estão representadas na Tabela 20.

Tabela 20. Figuras de Mérito para os modelos Completo e OPS

	Colmo Médio, Inferior e Superior	
	OPS	Completo
RPD	2,79	1,66
RMSEP	0,80	0,93
Rp	0,99	0,98
SEL	0,076	0,056
SEN.	$4{,}6x10^{-5}$	$1{,}5x10^{-5}$
γ	5,57	17,91
γ^{-1}	0,18	0,056
LODruido	12,37	16,82
LOQruido	37,5	50,98
LODprop	2,2	2,52
LOQprop	6,68	7,64

Os resultados apresentados na Tabela 20 indicam que o modelo apresenta alta capacidade para prever lignina em bagaço de cana-de-açúcar com alta exatidão em relação ao método de referência. Com esse modelo, é possível prever o teor de lignina nas diferentes partes do colmo: pé, meio e ponta. O inverso da sensibilidade analítica γ^{-1} mostrou que o modelo é sensível e apresentou alta seletividade (SEL). Da mesma forma

que os demais conjuntos de dados, observa-se que o método proposto por Teófilo foi mais coerente, pois o valor de LOQ_{prop} ficou mais próximo dos do valor mínimo encontrado.

A Figura 44 apresenta os valores de erros relativos para o conjunto calibração e previsão.

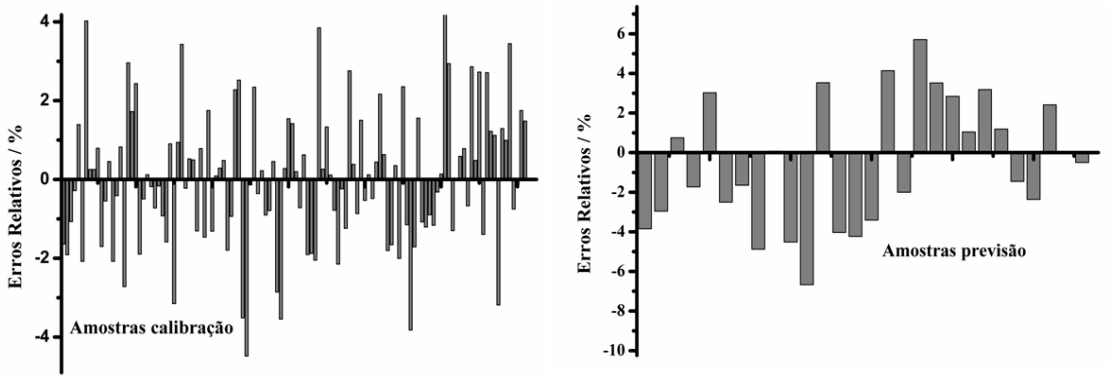


Figura 44. Erros Relativo (%) conjunto calibração e conjunto previsão.

A Figura 45 apresenta os valores medidos versus preditos para o conjunto calibração e previsão.

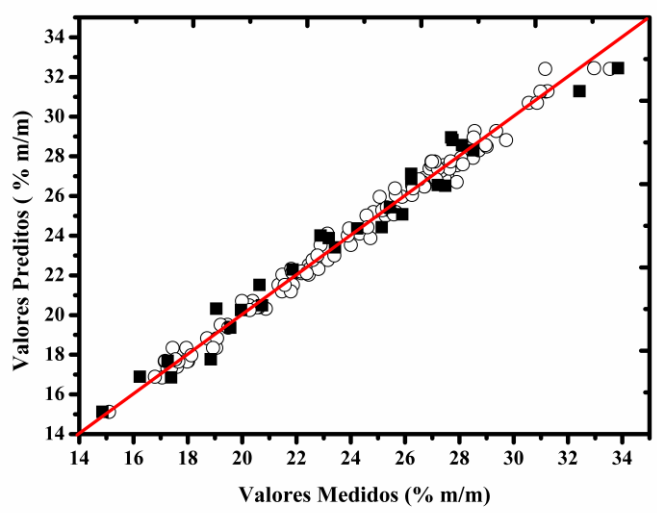


Figura 45. Valores medidos x preditos. Bolas representam o conjunto calibração e quadrados o conjunto previsão.

5. CONCLUSÕES

A análise dos parâmetros estatísticos indica que, tanto o algoritmo OPS como o GA, diminuíram significativa os erros de validação cruzada e de previsão. Vale lembrar que o tempo computacional do algoritmo OPS é consideravelmente menor que o do GA. Por essa razão, os modelos finais escolhidos foram, todos, utilizando o algoritmo OPS.

O método OPS aliado à regressão PLS proporciona a construção de modelos mais simples, interpretáveis e preditivos.

Os modelos apresentaram boa capacidade de previsão quando aplicados para prever o teor de lignina e cana-de-açúcar, podendo ser perfeitamente aplicados, substituindo análises morosas, dispendiosas e não corretas ambientalmente, por uma análise extremamente rápida, de baixo custo e que não exige gasto de reagentes em análises via úmida.

6. REFERÊNCIAS BIBLIOGRÁFICAS

1. Mota, C. J. A.; Silva, C. X. A.; Gonçalves, V. L. C. Gliceroquímica: novos produtos e processos a partir da glicerina de produção de biodiesel. Química Nova, **2009**, (32), 639-648.

2. Leite, R. C. C.; Leal, M. R. L. V.; Cortez, L. A. B.; Griffin, W. M.; Scandiffio, M. I. G. Can Brazil replace 5 of the 2025 gasoline world demand with ethanol? Energy, **2009,** (34) ,655-661.

3.Schmer, M. R.; Vogel, K. P.; Mitchell, R. B.; Perrin, R. K. Net energy of cellulosic ethanol from switchgrass., Proceedings of the National Academy of Sciences of The United States of America, **2008,** (105), 464-469.

4.De Oliveira, M.; Vaughan, B. E.; Rykiel, E. J. Ethanol as fuels: Energy, carbon dioxide balances, and ecological footprint. Bioscience, **2005, (**55),593-602.

5. De Gorter, H.; Just, D. R. The Social Costs and Benefits of Biofuels: The Intersection of Environmental, Energy and Agricultural Policy. Applied Economic Perspectives and Policy, **2010,** (32), 4-32.

6.Demirbas, A.,Progress and recent trends in biofuels. Progress in Energy and Combustion Science, **2007,** (33), 1-18.

7. Balat, M.; Balat, H. Recent trends in global production and utilization of bio-ethanol fuel. Applied Energy, **2009,** (86), 2273-2282.

8. Santos, A.F.; Queiróz, J.H.; Colodette, J.L.; Fernandes, S.A.; Guimarães, V.M.; Rezende, S.T. Potencial da palha de cana-de-açúcar para produção de etanol. Química Nova, **2012**, (35), 1004-1010.

9. Unica: União da Indústria de Cana-de-açúcar Disponível em: <http://www.unica.com.br> acessado em 07 de novembro, 2013.

10. Farrell, A. E.; Plevin, R. J.; Turner, B. T.; Jones, A. D.; O'Hare, M.; Kammen, D. M. Ethanol can contribute to energy and environmental goals. Science, **2006, (**311), 506-508.

11.Santchurn, D.; Ramdoyal, K.; Badaloo, M. G. H.; Labuschagne, M. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica, **2012,** (185), 543-558.

12.Santos, M. S. M.; Madalena, J. A.; Soares, L.; Ferreira, P. V.; Barbosa, G. V. S. Repetibilidade de características agroindustriais em cana-de-açúcar. Pesquisa agropecuária brasileira, **2004**, (39), 301-306.

13. Barbosa, G. V. S.; Souza, A, J, R.; Rocha, A. M. C.; Ribeiro, G. A. G.; Ferreira, J. L. C.; Soares, L.; Cruz, M. M.; Silva, W. C. M. Novas variedades RB de cana-de-açúcar para Alagoas, Maceió: UFAL, Programa de Melhoramento Genético de Cana-de-Açúcar, Boletim Técnico - Programa de Melhoramento Genético de Cana-de-Açúcar, 2000, 1, 16.

14. Kim, M.; Day, D. Composition of sugar cane, energy cane, and sweet sorghum suitable for ethanol production at Louisiana sugar mills. Journal of Industrial Microbiology & Biotechnology, **2011**, **(**38), 803-807.

15. Berding, N.; Pendrigh, R. S. Breeding implications of diversifying end uses of sugarcane. International Sugar Journal, **2009,** 111, (1331), 676.

16. Byrt, C. S.; Grof, C. P. L.; Furbank, R. T. C-4 Plants as biofuel feedstocks: optimising biomass production and feedstock quality from a lignocellulosic perspective free access. Journal Integraty Plant Biology, **2011**, **(**53), 120-135.

17. Balat, M.; Balat, H. Recent trends in global production and utilization of bio-ethanol fuel. Applied Energy, **2009**, (86), 2273-2282.

18. Sun, Y.; Cheng, J. Y. Hydrolysis of lignocellulosic materials for ethanol production: a review. Bioresource Technology, **2002**, (83), 1-11.

19. Carvalheiro, F.; Duarte, L. C.; Girio, F. M. Hemicellulose biorefineries: a review on biomass pretreatments. Journal of Scientific Industrial Research, **2008**, (67), 849-864.

20. Girio, F. M.; Fonseca, C.; Carvalheiro, F.; Duarte, L. C.; Marques, S.; Bogel-Lukasik, R. Hemicelluloses for fuel ethanol: A review. Bioresource Technology, **2010,** 101, (13, SI), 4775-4800.

21. Kim, M.; Day, D. Composition of sugar cane, energy cane, and sweet sorghum suitable for ethanol production at Louisiana sugar mills. Journal of Industrial Microbiology & Biotechnology, **2011, (**38), 803-807.

22. Santchurn, D.; Ramdoyal, K.; Badaloo, M. G. H.; Labuschagne, M. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica, **2012,** (3), 543-558.

23. Barbosa, M. H. P., Study of genetic divergence in sugarcane varieties grown in Brazil using the parentage coefficient. International Sugar Journal, **2001**, (103), 294-295.

24. da Silva, P. P.; Soares, L.; da Costa, J. G.; Viana, L. d. S.; Farias de Andrade, Julio Cesar; Goncalves, E. R.; dos Santos, J. M.; de Souza Barbosa, Geraldo Verissimo; Nascimento,V. Pathanalysis for selection of drought tolerant sugar cane genotypes through physiological components. Industrial Crops And Products, **2012**, (37), 11-19.

25. Masarin, F.; Gurpilhares, D. B.; Baffa, D. C. F.; Barbosa, M. H. P.; Carvalho, W.; Ferraz, A.; Milagres, A. M. F. Chemical composition and enzymatic digestibility of sugarcane clones selected for varied lignin content. Biotechnology for Biofuels, **2011**, **(**4),55.

26. Marim, B. G.; Silva, D. J. H.; Carneiro, P. C. S.; Miranda, G. V.; Mattedo, A. P.; Caliman, F. R. B. Variabilidade genética e importância relativa de caracteres em acessos de germoplasma de tomateiro.Pesquisa agropecuária brasileira, Brasília, **2009**, (44), 1283-1290.

27. Santchurn, D.; Ramdoyal, K.; Badaloo, M. G. H.; Labuschagne, M. From sugar industry to cane industry: investigations on multivariate data analysis techniques in the identification of different high biomass sugarcane varieties. Euphytica, **2012**, (185), 543-558.

28. Carrier, M.; Loppinet-Serani, A.; Denux, D.; Lasnier, J.-M.; Ham-Pichavant, F.; Cansell, F.; Aymonier, C. Thermogravimetric analysis as a new method to determine the lignocellulosic composition of biomass. Biomass Bioenergy, **2011**, (35), 298-307.

29. Sills, D. L.; Gossett, J. M. Using FTIR to predict saccharification from enzymatic hydrolysis of alkali-pretreated biomasses. Biotechnology and Bioengineering, **2012**, (109), 353-362.

30. Dang, V. Q.; Bhardwaj, N. K.; Hoang, V.; Nguyen, K. L. Determination of lignin content in high-yield kraft pulps using photo acoustic rapid scan Fourier transform infrared spectroscopy. Carbohydrate Polymers, **2007**, (68),489-494.

31. Sena, M. M.; Trevisan, M. G.; Poppi, R. J. PARAFAC: uma ferramenta quimiométrica para tratamento de dados multidimensionais. Aplicações na determinação direta de fármacos em plasma humano por espectrofluorimetria. Química Nova, **2005**, (28), 910-920.

32. Sun, B.; Liu, J.; Liu, S.; Yang, Q. Application of FT-NIR-DR and FT-IR-ATR spectroscopy to estimate the chemical composition of bamboo (Neosinocalamusaffinis Keng). Holzforschung, **2011**, (65), 689-696.

33. Uner, B.; Karaman, I.; Tanriverdi, H.; Ozdemir, D. Determination of lignin and extractive content of Turkish Pine (Pinusbrutia Ten.) trees using near infrared spectroscopy and multivariate calibration. Wood Science and Technology, **2011**, (45), 121-134.

34. Trafela, T.; Strlic, M.; Kolar, J.; Lichtblau, D. A.; Anders, M.; Mencigar, D. P.; Pihlar, B. Nondestructive analysis and dating of historical paper based on IR Spectroscopy and chemometric data evaluation. Analytical Chemistry, **2007**, (79), 6319-6323.

35. Nuopponen, M. H.; Birch, G. M.; Sykes, R. J.; Lee, S. J.; Stewart, D. Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared Fourier transform (DRIFT-MIR) spectroscopy. Journal of Agricultural and Food Chemistry, **2006**, (54), 34-40.

36. Kankanala, K. S.; Schenzel, K. C. Determination of cellulose I crystallinity of plant fibres using NIR FT Raman spectral data and multivariate calibrations. Abstracts of Papers of the American Chemical Society, **2010**, (239).

37. Jones, R. W.; Meglen, R. R.; Hames, B. R.; McClelland, J. F. Chemical analysis of wood chips in motion using thermal-emission mid-infrared spectroscopy with projection to latent structures regression. Analytical Chemistry, **2002**, (74), 453-457.

38. Sheng, Y.; Guo-feng, W.; Yi-fei, J.; Xiao-dong, F.; Hong-kun, L.; Mei, S.; Jun-wen, P. Extending Hemicelluloses Content Calibration of Acacia Spp Using NIR to New Sites. Spectroscopy and Spectral Analysis, **2010**, (30), 1206-1209.

39. Conab: Companhia Nacional de Abastecimento. Central de informações agropecuárias: safras – cana. **2012**. Disponível em: <http://www.conab.gov.br> acessado em 25 de outubro, 2013.

40. Cesnik, K, R.; Miocque, J. Melhoramento da cana-de-açúcar. Brasília: Embrapa - Informações Tecnológicas, **2004**, 307.

41. Van Dillewijn, C. B. Botanyofsugarcane. Waltham: The ChronicaBotanica, **1952**, 371.

42. Wheals, A.E.; Basso, L.C.; Alves, D.; Amorim, H.V. Fuel ethanol after 25 years. Trends in Biotechnology, **1999**, 17, (12), 482-487.

43. Lange, J.P. Lignocellulose conversion: an introduction to chemistry, process and economics. Biofuels Bioproducts and Biorefinery, **2007**, (1), 39-48.

44. INSTITUTO CUBANO DE INVESTIGACIONES DE LOS DERIVADOS DE LA CANÃ DE AZUCAR (ICIDCA). Manual de los derivados de La cana de azucar. México: CEPLACEA, 1990, 447.

45. FONSECA, B. G. Destoxicação biológica de hidrolisado hemicelulósico de bagaço de cana-de-açúcar empregando as leveduras Issatchenkiaoccidentalis e Issatchenkiaorientalis. Dissertação (Mestrado em Biotecnologia Industrial). Universidade de São Paulo, **2009**.

46. Fengel, D.; Wegener, G. Wood chemistry ultra structure, reactions. New York: Walter de Gruyter, **1984**, 167-181.

47. Hofrichter M. Review: ligin conversion by manganese peroxidase (MnP). Enzime and Microbial Technology. New York, **2002**, (30), 454-466.

48. Onnerud, H., Zhang, L., Gellerstedt, G.; Henriksson, G. Polymerization of monolignols by redox shttle – mediated enzymatic oxidation. The Plant Cell. Rockville, **2002**, 14, 1953 – 1968.

49. Rio, J.D.C. Determining the influence of eucalypt lignin composition in paper pulp yield using Py-GC/MS. Journal of Analytical and Applied Pyrolysis, **2005**, (74), 110-115.

50. De Gorter, H.; Just, D. R. The Social Costs and Benefits of Biofuels: The Intersection of Environmental, Energy and Agricultural Policy. Applied Economic Perspectives and Policy, **2010**, (32), 4-32.

51. Masarin, F.; Gurpilhares, D. B.; Baffa, D. C. F.; Barbosa, M. H. P.; Carvalho, W.; Ferraz, A.; Milagres, A. M. F. Chemical composition and enzymatic digestibility of sugarcane clones selected for varied lignin content. Biotechnology for Biofuels, **2011,** 4.

52. Brown, T. L.; Lemay Jr, H. E.; Bursten, B.E.; Burdge, J.R.; Química: a ciência central. São Paulo: Pearson Prentice Hall, **2005**, 5, 972.

53. Oliveira, L. F. C. Espectroscopia Molecular. Cadernos Temáticos de Química Nova na Escola, **2001**, (53).

54. Skoog, D.A.; West, D. M.; Holler, F.J.; Stanley, R.C. Fundamentos de Química Analítica, Tradução da 8ª Edição norte-americana, Editora Thomson, São Paulo-SP, **2006**.

55. Drennen. J.K.; Kraemer, E.G.; Lodder, R.A. Advances and perspectives in near-infrared spectrophotometry.Critical Reviews in Analytical Chemistry, **1991**, (22), 443-475.

56. Scafi, S. H. F. Sistema de Monitoramento em Tempo Real de Destilações de Petróleo e Derivados Empregando a Espectroscopia no Infravermelho próximo. Tese (Doutorado em Química) – Instituto de Química, Universidade Estadual de Campinas, Campinas, **2005.**

57. Wetzel, D. L. Near-Infrared Reflectance Analysis - Sleeper Among Spectroscopic Techniques. Analytica lChemistry, **1983**, (55), 1165-1176.

58. Pasquini, C. Espectroscopia no Infravermelho Próximo: fundamentos, aspectos práticos e aplicações analíticas. Jornal da Sociedade Brasileira de Química, **2003**, (14), 198-219.

59. Teófilo, R. F. Métodos Quimiométricos em Estudos Eletroquímicos de Fenóis sobre Filmes de Diamante Dopado com Boro. Tese (Doutorado em Química) - Universidade Estadual de Campinas, **2007**.

60. Escandar, G. M.; Damiani, P. C.; Goicoechea, H. C.; Olivieri, A. C. A review of multivariate calibration methods applied to biomedical analysis. Microchemistry Journal, **2006**, (82), 29-42.

61. Charne, R.; De Luna Freire, C. A.; Charnet, E. M. R.; Bovino, H. Análise de Modelos de Regressão Linear com Aplicações. Campinas: Unicamp, **1999**.

62. Miller, J. N.; Miller, J. C. Statistics and Chemometrics for Analytical Chemistry. London: Prentice Hall, **2000**.

63. Chui, Q. S. H.; Zucchini, R. R.; Lichtig, J. Qualidade de medições em química analítica. Estudo de caso: determinação de cádmio por espectrofotometria de absorção atômica com chama, Química Nova. **2001**, (24), 374-380.

64. Barros Neto, B.; Scarminio, I. S.; Bruns, R. E. Como fazer experimentos: Pesquisa e desenvolvimento na ciência e na industria. 2. ed. Campinas: Unicamp, 2001.

65. Martens, H.; Naes, T. Multivariate calibration. New York: Wiley, **1996.**

66. Wold, H. The basic design and some extensions. In Systems under indirect observation, Amsterdam, 1982, (2), 1-53.

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

68. Barlow, J.L.; Bosner, N.; Drmac, Z. A new stable bidiagonal reduction algorithm. Linear Algebra and Its Applications, **2005**, (397), 35-84.

69. Manne, R. Analysis of 2 partial-least-squares algorithms for multivariate calibration. Chemometrics and Intelligent Laboratory Systems, **2001**, (58),187-197.

70. Martins, J.P. A.; Teófilo, R. F.; Ferreira, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. Journal of Chemometrics, **2010**, (24), 320-332.

71. Brereton, R. G. Applied Chemometrics for Scientists. Chichester, Inglaterra: Wiley, **2007.**

72. WILLIAMS, P. C. Implementation of near-infrared technology. Near Infrared Technology in the Agricultural and Food Industries, **2001**, 145-169.

73. Beecher, B. W.; Ciavarella, H. G. S. The potential of near-infrared reflectance spectroscopy for soil analysis – a case study from the Riverine Plain of south-eastern Australia. Australian Journal of Experimental Agriculture, **2002**, (42), 607-614.

74. Chang, C.; Laird, D.A.; Mausbach, M.J.; Hurburg, C.R. Near infrared reflectance spectroscopy – principal components regression analyses of soil properties. Soil Science Societyof American Journal, **2001**, (25), 480-490.

75. Oliveira, F. C.; Souza, A. T. P. C.; Dias, J. A.; Dias, S. C. L.; Rubim, J. C. A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos, Química Nova, **2004**, (27), 218-225

76. Osborne, S. D.; Jordan, R. B.; Künnemeyer, R., Method of wavelength selection for partial least square. Analyst, **1997**, (122), 1531-1537.

77. Costa Fillho, P. A.; Poppi, R. J. Algoritmo genético em química. Química Nova, **1999**, (22), 405-411.

78. Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L; Engelsen, S. B. Interval partial least-square regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, Applied Spectroscopy, **2000**, (54), 413-419.

79. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. Journal of Chemometrics, **2009**, (23), 32-48.

80. Ferreira, M. M. C. Multivariate QSAR. Jounal of the Brazilian Chemistry Society, **2002**, (13), 742-753.

81. Leardi, R.; Norgaad, L. Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, Journal of Chemometrics, **2004**, (18), 486-497.

82. Cartwright, H. M. Applications of Artificial Intelligence in Chemistry, Oxford Science Publications, New York, **1995**.

83. Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M. PLS_Toolbox 3.01, Eigenvector Research, Inc.: Manson, **2003**.

84. Goicoechea, H. C.; Olivieri, A. C. A new family of genetic algorithms for wavelength interval selection in multivariate analytical spectroscopy. Journal of Chemometrics, **2003**, (17), 338-345

85. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. OPS Toolbox 1.0, Registro de Software no INPI - 0000270703255138: Brasil, **2007**.

86. Teófilo, R. F.; Martins, J. P. A.; Ferreira, M. M. C. In Ordered predictors Selection: an intuitive method to find the most relevant variable in multivariate calibration, 10thInternational Conference on Chemometrics in Analytical Chemistry, Águas de Lindóia, Brazil, **2006**, P066.

87. Martins, J. P. A.; Ferreira, M. M. C. QSAR Modeling: Um novo pacote computacional open source para gerar e validar modelos QSAR. Química Nova, **2013**, (36), 554-560.

88. Eurachem/Citac – Work Group. Guide of quality in analytical chemistry – An aid to accreditation. 2. ed. **2002.**

89. ASTM INTERNATIONAL. ASTM E1655-05 Standard Practices for Infrared Multivariate Quantitative Analysis, **2012**.

90. Martens, H.; Naes, T. Multivariate calibration. New York: Wiley, **1996**.

91. Currie, L. A. Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995), Analytical ChimicaActa, **1999**, (391), 105-126.

92. Booksh, K. S.; Kowalski, B. R. Theory of analytical chemistry. Anaytical. Chemistry, **1994**, (66), 782-791 .

93. Muñoz de la Pena, A.; Espinosa-Mansilla, A.; Acedo Valenzuela, M. I.; Goicoechea, H. C.; Olivieri, A. C. Comparative study of net analyte signal-based 128methods and partial least squares for the simultaneous determination of amoxycillin and clavulanic acid by stopped-flow kinetic analysis, Analytica Chimica Acta, **2002**, (463), 75-88.

94. Rodríguez, L. C.; Campanã, A. M. G.; Linares, C. J.; Ceba, M. R. Estimation of performance characteristics of an analytical method using the data set of the calibration experiment. Analytical Letters ,**1993**, (6), 1243-1258.

95. Vessman, J.; Stefan, R. I.; Van Staden, J. F.; Danzer, K.; Lindner, W.; Burns, D. T.; Fajgelj, A.; Muller, H. Selectivity in analytical chemistry (IUPAC Recommendations 2001). Pure an Applied chemistry, **2001**, (73), 1381-1386.

96. Lorber, A.; Faber, K.; Kowalski, B. R., Net analyte signal calculation in multivariate calibration. Analytical Chemistry, **1997**, (69), 1620-1626.

97. Ferre, J.; Brown, S. D.; Rius, F. X. Improved calculation of the net analyte signal in inverse multivariate calibration, Journal of Chemometrics, **2001**, (15), 537-553.

98. CheavegattiVianotto, A.; Abreu, H. M. C. de; Arruda, P.; Bespalhok Filho, J. C.; Burnquist, W. L.; Creste, S.; Ciero, L. di; Ferro, J. A.; Figueira, A. V. de O.; Fiulgueiras, T. de S.; Grossi de Sa, M. de F.; Guzzo, E. C.; Hoffman, H. P.; Landell, M. G. de A.; Macedo, N.; Matsuoka, S.; Reinach, F. de C.; Romano, E.; Silva, W. J. da; Silva Filho, M. de C.; Ulian, E. C. Sugarcane (Saccharum X officinarum): a

reference study for the regulation of genetically modified cultivars in Brazil. Tropical Plant Biology, **2011**, (4), 62-89.

99. National Renewable Energy Laboratory (NREL). Chemical Analysis and Testing Task: Laboratory Analytical Procedure, Golden, **1996.**

100. ACID – Insoluble lignin in wood and pulp T222 om-98. In: TAPPI test methods. Atlanta: TAPPI, **1998**.

101. MATLAB for Windows, TheMathWorks, Inc., versão 5.1.0.421, 1984-1997.

102.Bouveresse, E.; Massart, D. L. Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration. Chemometrics and Intelligent Laboratory Systems, **1996**, (32), 201-213.

103. Rambo, M. K .D.; Amorim, E. P.; Ferreira, M. M. C. Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. Analytica Chimica Acta, **2013**, (775), 41-49.

104. Liu, L.; Ye, X. P.; Womac, A. R.; Sokhansanj, S. Variability of biomass chemical composition and rapid analysis using FT-NIR techniques. Carbohydrate Polymers, **2010**, (81), 820-829.

105. Jung, H. J. G.; Lamb, J. F. S. Prediction of cell wall polysaccharide and lignin concentrations of alfalfa stems from detergent fiber analysis. Biomass and Bioenergy, **2004**, (27), 365-373.

106. Krähme, A.; Gudi, G.; Weiher, N.; Gierus, M.; Schütze, W.; Schulz, H. Characterization and quantification of secondary metabolite profiles in leaves of red and white clover species by NIR and ATR-IR spectroscopy. Vibrational Spectroscopy, **2013**, (68), 96-103.

107. Menesatti, P.; Antonucci, F.; Pallottino, F., Roccuzzo, G. ; Allegra, M.; Stagno, F.; Intrigliolo, F. Estimation of plant nutritional status by Vis–NIR spectrophotometric analysis on orange leaves [Citrus sinensis(L) Osbeck cv Tarocco]. Biosystems Engineering, **2010**, (105), 448-453.

108. Purcell, D. E.; O'Shea, M. G.; Kokot, S. Role of chemometrics for at-field application of NIR spectroscopy to predict sugarcane clonal performance. Chemometrics and Intelligent Laboratory Systems, **2007**, (87), 113-124.