ÍTALO PELIÇÃO CALIARI

PREVISÃO DAS PROPRIEDADES CRISTALINIDADE E TEORES DE CARBOIDRATOS ESTRUTURAIS EM BIOMASSA DE CANA-DE-AÇÚCAR USANDO NIR, PLS E MÉTODOS DE SELEÇÃO DE VARIÁVEIS

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
2017

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

T
C153p
2017

Caliari, Ítalo Pelição, 1991-
Previsão das propriedades cristalinidade e teores de carboidratos estruturais em biomassa de cana-de-açúcar usando NIR, PLS e métodos de seleção de variáveis / Ítalo Pelição Caliari. – Viçosa, MG, 2017.
vii, 79f. : il. (algumas color.) ; 29 cm.

Orientador: Reinaldo Francisco Teófilo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

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

CDD 22 ed. 633.61

ÍTALO PELIÇÃO CALIARI

PREVISÃO DAS PROPRIEDADES CRISTALINIDADE E TEORES DE CARBOIDRATOS ESTRUTURAIS EM BIOMASSA DE CANA-DE-AÇÚCAR USANDO NIR, PLS E MÉTODOS DE SELEÇÃO DE VARIÁVEIS

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: 13 de fevereiro de 2017.

Antônio Augusto Neves

Luiz Alexandre Peternelli

Reinaldo Francisco Teófilo
(Orientador)

SUMÁRIO

LISTA DE ABREVIAÇÕES

Abreviatura	Termo em Ingês	Termo em Português
1G	First Generation Biofuels	Biocombustíveis de Primeira Geração
2G	Second Generation Biofuels	Biocombustíveis de Segunda Geração
Baseline	Baseline Correction	Correção de Linha de Base
D1	First Derivative	Primeira Derivada
D2	Second Derivative	Segunda Derivada
DOE	Design of Experiments	Planejamento de Experimentos
ED	Electrochemical Detector	Detector Eletroquímico
ELSD	Evaporative Light Scattering Detector	Detector Evaporativo de Espalhamento de Luz
GA	Genetic Algorithm	Algoritmo Genético
GC	Gas Chromatography	Cromatografia Gasosa
GHG	Greenhouse Gases	Gases de Efeito Estufa
h	Number of latent variables	Número de variáveis latentes
HPLC	High Performance Liquid Chromatography	Cromatografia Líquida de Alta Eficiência
LOD	Limit of Detection	Limite de Detecção
MC	Mean Center	Centrar na Média
MLR	Multiple Linear Regression	Regressão Linear Multipla
MS	Mass Spectroscopy	Espectroscopia de Massas
MSC	Multiplicative Signal Correction	Correção Multiplicativa de Sinal
NAS	Net Analytical Signal	Sinal Analítico Líquido
NIR	Near infrared	Infravermelho Próximo
OPS	Ordered Predictors Selection	Seleção dos Preditores Ordenados
PCR	Principal Components Regression	Regressão por Componentes Principais
PLS	Partial Least Squares	Regressão por Mínimos Quadrados Parciais
RCV	Correlation Coefficient of Cross-Validation	Coeficiente de Correlação de Validação Cruzada
RID	Refractive Index Detector	Detector de Índice de Refração
RMSECV	Root Mean Square Error of Cross-Validation	Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada
RMSEP	Root Mean Square Error of Prediction	Raiz Quadrada do Erro Quadrático Médio de Previsão
RP	Correlation Coefficient of Prediction	Coeficiente de Correlação de Previsão
RPD	Ratio of Performance to Deviation	Índice de Desempenho do Desvio
SD	Standard Deviation	Desvio Padrão
SEL	Selectivity	Seletividade
SEN	Sensitivity	Sensibilidade
SNV	Standard Normal Variance	Variável Normal Padrão
XRD	X-Ray Diffractometry	Difratometria de Raio-X
γ	Analytical Sensitivity	Sensibilidade Analítica

RESUMO

CALIARI, Ítalo Pelição, M.Sc., Universidade Federal de Viçosa, fevereiro de 2017. **Previsão das propriedades cristalinidade e teores de carboidratos estruturais em biomassa de cana-de-açúcar usando NIR, PLS e métodos de seleção de variáveis.** Orientador: Reinaldo Francisco Teófilo. Coorientadores: Márcio Henrique Pereira Barbosa e Sukarno Olavo Ferreira.

Modelos para a previsão das propriedades cristalinidade de celulose e dos teores de glicanas e xilanas na biomassa de cana-de-açúcar (*Saccharum* spp.) foram construídos usando regressão por mínimos quadrados parciais (PLS) e espectroscopia no infravermelho próximo (NIR). Os métodos propostos surgem em substituição às tradicionais análises da biomassa por difratometria de raio-X (XRD) e cromatografia líquida de alta eficiência (HPLC). As principais vantagens do uso da espectroscopia NIR como alternativa às técnicas convencionais se dão devido à rapidez de análise (~20s), menor custo, ser não destrutiva, requerer um mínimo manuseio de amostra e facilidade de uso. Foram usados também os métodos de seleção de variáveis Seleção dos Preditores Ordenados (OPS) e Algoritmo Genético (GA) visando melhorar a capacidade preditiva e interpretativa dos modelos. Em todos os casos avaliados o algoritmo OPS mostrou-se superior ao GA. As amostras de biomassa utilizadas neste estudo apresentaram índices de cristalinidade que variaram de 50 a 81% e teores de glicanas e xilanas que variaram de 20 a 43% e 15 a 30%, respectivamente. Os modelos para cristalinidade, glicanas e xilanas foram construídos com 5, 2 e 2 variáveis latentes e 150, 431 e 451 variáveis independentes, respectivamente. As transformações espectrais que forneceram melhores modelos foram as de derivadas. Os parâmetros estatísticos raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*), raiz quadrada do erro quadrático médio de previsão (*RMSEP*), coeficiente de correlação de validação cruzada (*RCV*), coeficiente de correlação de previsão (*RP*) e índice de desempenho do desvio (*RPD*) são descritos, respectivamente, para o modelo de: *i*) cristalinidade: 3,31; 3,01; 0,86; 0,92, e 1,71; *ii*) Glicanas: 2,16; 1,81; 0,94; 0,95 e 2,00; *iii*) Xilanas: 1,27; 1,32; 0,91; 0,94 e 1,79. As correlações entre a cristalinidade e outras propriedades químicas da biomassa de cana-de-açúcar tais como lignina, α-celulose, hemicelulose e cinzas não foram significativas, e assim, a cristalinidade se mostrou uma propriedade independente na biomassa de cana-de-açúcar.

ABSTRACT

CALIARI, Ítalo Pelição, M.Sc., Universidade Federal de Viçosa, February, 2017. **Estimation of crystallinity and structural carbohydrate contents in sugarcane biomass using NIR, PLS and variable selection methods.** Advisor: Reinaldo Francisco Teófilo. Co-advisors: Márcio Henrique Pereira Barbosa and Sukarno Olavo Ferreira.

Models for the prediction of crystallinity of cellulose and glucans and xylans contents in sugarcane biomass (*Saccharum* spp.) were built using partial least squares regression (PLS) and near-infrared spectroscopy (NIR). The proposed methods appear replacing the traditional analyzes of biomass by X-ray diffractometry (XRD) and high performance liquid chromatography (HPLC). The main advantages of using NIR spectroscopy as an alternative to conventional techniques are due to rapid analysis (~20s), lower cost, being non-destructive, requiring a minimal sample handling and ease of use. Were also used the variable selection methods Ordered Predictors Selection (OPS) and Genetic Algorithm (GA) to improve the predictive and interpretive capacity of the models. In all cases, the OPS algorithm was superior to GA. The biomass samples used in this study showed crystallinity indexes ranging from 50 to 81% and glucans and xylans contents ranging from 20 to 43% and 15 to 30%, respectively. The models for crystallinity, glucans and xylans were built with 5, 2 and 2 latent variables and 150, 431 and 451 independent variables, respectively. The spectral transformations that provided the best models were the derivatives. The root mean square error of cross-validation (*RMSECV*), root mean square error of prediction (*RMSEP*), correlation coefficient of cross-validation (*RCV*), correlation coefficient of prediction (*RP*), and ratio of performance to deviation (*RPD*) are described, respectively, for the model of: *i)* crystallinity: 3.31; 3.01; 0.86; 0.92, and 1.71; *ii)* Glucans: 2.16; 1.81; 0.94; 0.95 and 2.00; *iii)* Xylans: 1.27; 1.32; 0.91; 0.94 and 1.79. The correlations between crystallinity and other chemical properties of sugarcane biomass such as lignin, α-cellulose, hemicellulose, and ash were not significant, and hence, crystallinity proved to be an independent property in sugarcane biomass.

CAPÍTULO 1

Revisão Bibliográfica

1. Introdução Geral

Uma grande preocupação econômica e ambiental relacionada ao uso de combustíveis derivados de fontes fósseis afeta os governos e o setor produtivo, uma vez que as reservas de petróleo tendem a se extinguir em alguns anos. Além disso, a utilização desse tipo de combustível está largamente associada à liberação indiscriminada de gases causadores do efeito estufa (GHG). A utilização de matérias-primas alternativas renováveis para a produção de combustíveis vem ganhando cada vez mais visibilidade e relevância nos centros de pesquisa ao redor do mundo. Nos Estados Unidos, a produção de bioalcoóis (e.g. etanol e butanol) a partir do milho é muito explorada e o mesmo ocorre no Brasil com a produção de etanol de primeira geração (1G) a partir do caldo de cana-de-açúcar (Neves, Pitarelo & Ramos, 2016). Com o uso de fontes renováveis na produção de biocombustíveis, pode-se contornar o problema da grande quantidade de gases que é gerada. Durante a queima do etanol, se comparado à gasolina, ocorre uma redução de 60% na emissão dos GHG, além do CO_2 ser reabsorvido durante o crescimento das plantações de cana-de-açúcar (Santos et al., 2012). Uma questão muito criticada relacionada aos combustíveis 1G é a relação competitiva existente entre o cultivo destinado à produção de alimentos e à produção de biocombustíveis. Uma vez que tanto o milho e a cana-de-açúcar são *commodities* muito usadas na indústria alimentícia, o uso destes vegetais para a produção de biocombustíveis acarretaria em um aumento dos preços no mercado de alimentos (Salehi Jouzani & Taherzadeh, 2015). Uma solução para esse problema seria a utilização de biomassa lignocelulósica para a produção de combustíveis, os chamados biocombustíveis de segunda geração (2G). Uma vez que esse material é o mais abundante do mundo e ainda não é aproveitado eficientemente, sua implementação de forma economicamente viável na indústria alcooleira é o próximo passo para o crescimento desse setor, trazendo vantagens tanto na esfera econômica quanto ambiental.

A produção de cana-de-açúcar no Brasil é destinada à indústria sucroalcooleira quase que em sua totalidade. Estima-se que a produção da safra de 2015/16 tenha ultrapassado a linha de 665 milhões de toneladas, um valor 4,9% superior à safra anterior (CONAB, 2016). Para cada tonelada de cana-de-açúcar são gerados 280 kg de biomassa (Huang et al., 2016) porém grande parte dessa biomassa produzida não é adequadamente

aproveitada (Lavarack, Griffin & Rodman, 2002). Ainda que atualmente grande parte da biomassa aproveitada seja constituída da fração do bagaço, espera-se que seja incorporada a utilização da palha ao longo dos anos, uma vez que existe um esforço por parte das autoridades competentes visando reduzir impactos ambientais relativos a queima dessa fração (Santos et al., 2012). Assim, a tendência é que a colheita passe a ser feita sem o corte da palha, aumentando ainda mais a quantidade de biomassa acessível às indústrias. Durante os processos industriais, a biomassa residual de cana-de-açúcar, obtida após a extração do caldo, é usada para geração de energia térmica e elétrica. Porém uma vez que as indústrias já se tornaram autossuficientes com essa prática, ainda existe um excedente de biomassa que poderia ser melhor aproveitado (Cardona, Quintero & Paz, 2010; Goldemberg, Coelho & Guardabassi, 2008; Huang et al., 2016; Macedo, Seabra & Silva, 2008; Pereira, Dias, MacLean & Bonomi, 2015). Para produções em larga escala, a matéria prima para produção de bioalcoóis constitui de 40 a 70% dos custos de produção, de forma que o aproveitamento de resíduos de biomassa é atrativo devido à sua abundância e baixo custo (Cardona et al., 2010). Assim, algumas das principais vantagens em se utilizar a biomassa para produção de biocombustíveis seriam de agregar valor a este resíduo, aumentar a produtividade sem necessariamente modificar a área plantada, e a não necessidade de acrescer significativamente custos de deslocamento.

A biomassa lignocelulósica é composta majoritariamente de celulose, hemicelulose e lignina em proporções que podem variar em função das espécies e até mesmo das diferentes partes de uma mesma planta.

Para a produção de combustíveis a partir de biomassa é necessário, primeiramente, quebrar os polímeros presentes na parede celular em monômeros, isto é, em açúcares fermentáveis na forma de hexoses e pentoses (Cardona et al., 2010; Kim, Lee & Kim, 2015; Mosier et al., 2005; Neves et al., 2016; Santos et al., 2012). Em suma, para a cana-de-açúcar, a fração celulósica é responsável pela contribuição em hexoses (i.e., glicose) enquanto que a hemicelulose é responsável pela fração de pentoses (i.e., xilose), em sua maior parte. Dessa forma, é de grande valor o conhecimento dessa proporção, assim como da recalcitrância da fração celulósica, muitas vezes associada à sua cristalinidade (Himmel et al., 2007). Através desse tipo de informação é possível estimar os custos e rendimentos existentes em processos industriais e direcionar estudos relacionados ao melhoramento genético dessa matéria prima, afim de se obter um material com características mais adequadas a uma determinada aplicação (Chen, Danao, Singh & Brown, 2014; Himmel et al., 2007; Rambo & Ferreira, 2015).

Muitas vezes os métodos de referência para obtenção das informações químicas da biomassa, apesar de fornecerem bons resultados, acabam por apresentar algumas desvantagens quando a análise se torna rotineira, e.g. custos elevados, manutenção frequente de equipamentos, consumo de tempo e reagentes, dentre outros. Porém, os métodos de referência podem ser substituídos por análises espectroscópicas, tais como a espectroscopia de infravermelho próximo (NIR). A espectroscopia NIR se apresenta como uma excelente alternativa para substituir os métodos de referência, uma vez que é uma técnica não destrutiva, pode ser realizada em poucos segundos e não necessita de preparo de amostra (Metrohm, 2013). Entretanto, as informações coletadas com a espectroscopia NIR são sobrepostas e complexas. Para extrair as informações contidas nos espectros é necessário usar métodos matemáticos e estatística multivariada. A área que aplica estes métodos a dados de origem química é conhecida como quimiometria. A quimiometria possui diversos métodos, mas em sua vertente mais promissora, busca encontrar uma relação quantitativa entre os espectros e a propriedade de interesse. A busca desta relação é realizada através dos métodos de regressão multivariada, tais como a regressão por mínimos quadrados parciais (PLS), a regressão mais usada até o momento em quimiometria.

Este trabalho teve como objetivo investigar a possibilidade de previsão das propriedades cristalinidade da celulose e dos teores de carboidratos estruturais em biomassa de cana-de-açúcar através de espectroscopia NIR e regressão PLS, uma vez que, pelo melhor de nosso conhecimento, não foram encontrados na literatura trabalhos envolvendo este tipo de aplicação para a biomassa da cana-de-açúcar.

2. Revisão Bibliográfica

2.1. Hidrólise da biomassa Lignocelulósica

A biomassa lignocelulósica é o material mais abundante no planeta (Gutiérrez-Rojas, Moreno-Sarmiento & Montoya, 2015), representando uma classe de compostos constituídos basicamente de celulose, hemicelulose, lignina, cinzas e extrativos, cada qual com sua determinada função na planta. Para a transformação desse material para uso em biorrefinarias é necessário primeiramente conhecer a estrutura de seus constituintes e seu caráter recalcitrante. Afim de transformar os polímeros presentes na parede celular vegetal em monômeros, para consumo em uma posterior etapa de fermentação alcoólica, são necessárias etapas de pré-tratamento e hidrólise, as quais são responsáveis pela maior parte dos custos industriais (Santos et al., 2012).

A etapa de pré-tratamento do material lignocelulósico é uma das etapas que mais tem efeito nos custos do processo (Santos, Colodette & Queiroz, 2013). Esta tem como objetivo deixar o material menos recalcitrante e mais acessível para a etapa de hidrólise. Uma vez que a celulose, hemicelulose e lignina estão intimamente ligadas, o acesso à celulose é facilitado se primeiramente removida a fração de lignina e hemicelulose. Esse processo pode ser realizado através de diferentes formas, podendo ser citados o uso de ácidos, álcali, enzimas, altas pressões, agentes biológicos, entre outros (Canilha et al., 2011; Huang et al., 2016; Sun, Sun, Cao & Sun, 2015; Timung et al., 2015).

Após a etapa de pré-tratamento, a fração processada deve ser despolimerizada e transformada em açúcares por um processo de hidrólise ácida ou enzimática. Uma vez que em alguns casos a hidrólise ácida leva a uma maior formação de agentes inibidores da etapa de fermentação e é necessário o uso de equipamentos resistentes à corrosão, esta rota industrial é muitas vezes evitada devido ao seu alto custo (Cardona et al., 2010; Kumar, Barrett, Delwiche & Stroeve, 2009). Em alternativa, a hidrólise enzimática vem apresentando grande potencial e melhores resultados se considerado o produto final, de forma que mesmo sendo uma rota mais lenta é considerada mais promissora (Cardona et al., 2010).

Para a celulose, a hidrólise enzimática é realizada por celulases, um complexo de enzimas constituídos por endoglucanases, exoglucanases e β-glicosidases (Santos et al., 2013). As endoglucanases agem quebrando as cadeias poliméricas em regiões de menor cristalinidade, consequentemente levando a formação de oligossacarídeos e um maior número de extremidades redutoras. As exoglucanases prosseguem com a degradação

liberando, a partir das extremidades da cadeia, unidades de celobiose. Por fim, as β-glicosidases levam à liberação dos monômeros de glicose a partir da celobiose (Sun & Cheng, 2002). Para o caso da hemicelulose, existe ainda uma serie de enzimas capazes de realizar a quebra nos monômeros, tais como a glucuronidase, acetilesterase, xilanase, β-xilosidase, galactomananase e glucomananase (Sun & Cheng, 2002).

Finalmente, através de um processo de fermentação ou catálise ocorre a conversão dos monômeros no produto de interesse. Uma vez que os monômeros gerados nas etapas anteriores se constituem em sua maior parte de glicose e xilose, é interessante a utilização de microrganismos ou catalisadores capazes de fornecer altos rendimentos durante a conversão.

2.1.1. Celulose

A celulose é composta unicamente por monômeros de D-glicose unidos através de ligações glicosídicas β-1,4 formando cadeias poliméricas lineares que interagem entre si por forças de van der Waals e por ligações de hidrogênio intermoleculares entre os grupos hidroxila presentes nas posições C-2, C-3 e C-6 (Mohanty, Misra & Hinrichsen, 2000; Peng et al., 2013; Tong et al., 2013). Uma cadeia de celulose é ilustrada na Figura 1. Devido a esse arranjo, existirão regiões de elevada ordenação, as chamadas regiões cristalinas. Há também regiões de baixa ou nenhuma ordenação, também ditas como regiões amorfas (Gutiérrez-Rojas et al., 2015; Klemm, Heublein, Fink & Bohn, 2005; Peng et al., 2013). Tipicamente, encontram-se nas diferentes fontes de material vegetal cadeias de celulose com grau de polimerização de 5000 a até mesmo 15000 unidades (Ogeda & Petri, 2010). Dentre seus diversos alomorfos, as formas cristalinas celulose I_α (triclínico) e celulose I_β (monoclínico) são as mais abundantes, sendo encontradas concomitantemente em materiais lignocelulósicos, diferindo em proporção de acordo com a fonte de biomassa (Klemm et al., 2005; Park, Baker, Himmel, Parilla & Johnson, 2010).

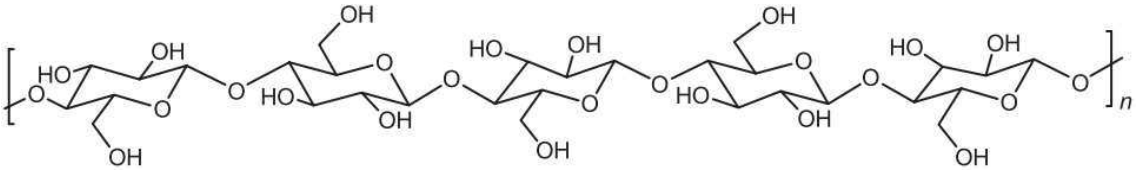

Figura 1. Representação de uma cadeia de celulose. Fonte: Santos (2012)

2.1.1.1. Cristalinidade

Definido como a fração de matéria cristalina na amostra, o chamado índice de cristalinidade está relacionado a diversas propriedades físicas dos materiais, como por exemplo módulo de Young, estabilidade dimensional, densidade, dureza, reatividade química entre outras (Jiang, Yang, So & Hse, 2007; Terinte, Ibbett & Schuster, 2011). Consequentemente, uma vez que as reações de hidrólise ocorrem preferencialmente nas regiões amorfas, o conhecimento do índice de cristalinidade é um dos fatores que pode fornecer uma estimativa do gasto energético envolvido nessas etapas e portanto da viabilidade do processo em si (Hendriks & Zeeman, 2009; Rambo & Ferreira, 2015; Sun & Cheng, 2002). Existem diferentes técnicas para se determinar a cristalinidade de materiais. Dentre elas estão a ressonância magnética nuclear (Ragauskas, 2014; Teeaar, Serimaa & Paakkarl, 1987) e a difratometria de raio-X (Jiang et al., 2007; Park et al., 2010; Teeaar et al., 1987; Timung et al., 2015), ressaltando-se que dentro de uma mesma técnica existem diferentes formas de se obter a cristalinidade (Ju, Bowden, Brown & Zhang, 2015; Park et al., 2010). Atualmente, cerca de 70% das medidas de cristalinidade são realizadas usando a difração de raio-X (Ragauskas, 2014), através do difratograma obtido pela leitura do ângulo 2θ pelo método do pó.

2.1.2. Hemicelulose

O termo hemicelulose diz respeito a um grupo de heteropolímeros que ocorrem juntamente à celulose e são compostos principalmente por glicose, galactose, manose, xilose, arabinose, ácido glucurônico, ácido 4-O-metil-glucurônico, entre outros (Gírio et al., 2010; Santos et al., 2012; Silva, Haraguchi, Muniz & Rubira, 2009). As estruturas dos monômeros presentes na fração de hemicelulose são apresentadas na Figura 2. Ao contrário do que ocorre com a celulose, esta classe de polímeros é ramificada, não apresenta caráter cristalino e tem um grau de polimerização significativamente inferior, variando de 100 a no máximo 200 (Silva et al., 2009). Estes fatores podem ser relacionados ao fato de a hemicelulose ser quebrada em monômeros mais facilmente se comparada à celulose.

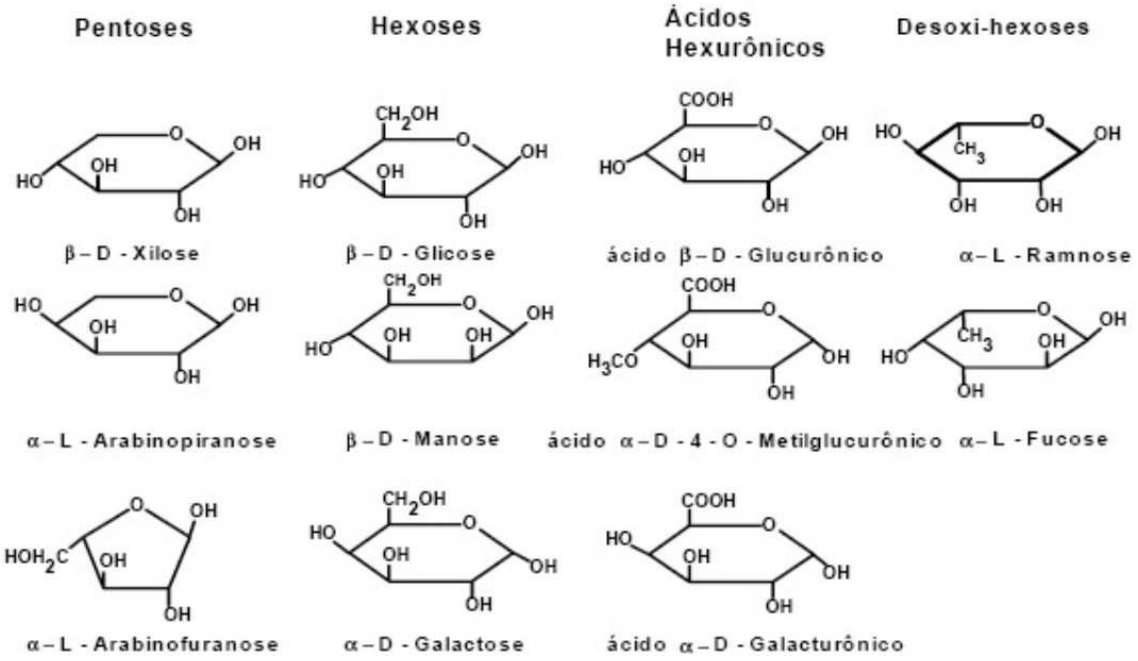

Figura 2. Estrutura de alguns monômeros constituintes da hemicelulose. Fonte: Rabelo (2010)

2.1.3. Lignina

A lignina é uma macro molécula que tem como precursores os álcoois p-coumarílico, coniferílico e sinapílico, dando origem respectivamente às ligninas tipo p-hidroxifenila (H), guaiacila (G) e siringila (S) (Boerjan, Ralph & Baucher, 2003; Colares et al., 2015; Ragauskas, 2014). A proporção dessas unidades pode variar dentre as diferentes fontes de biomassa, de forma que geralmente em angiospermas dicotiledôneas existe a maior ocorrência das formas G e S, em gimnospermas ocorre principalmente G e H, enquanto que em gramíneas apresentam-se as formas G, S e H (Barbosa, Maltha, Silva & Colodette, 2008). As estruturas dos álcoois precursores da lignina são apresentadas na Figura 3.

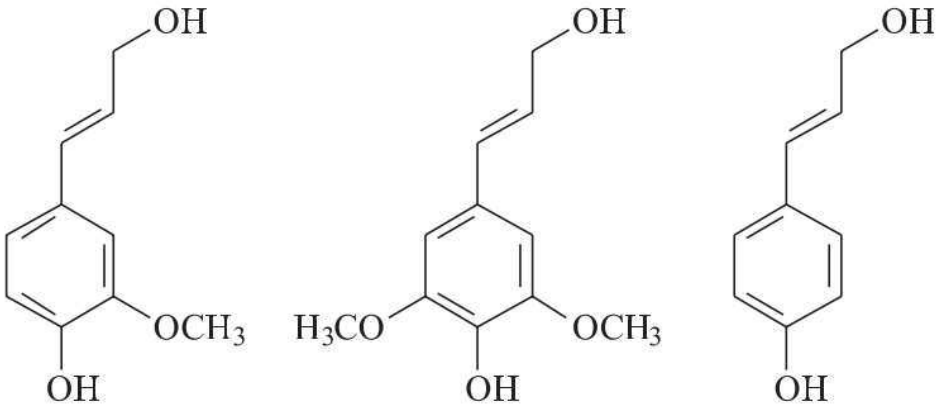

Figura 3. Precursores das unidades guaiacila (G), siringila (S) e p-hidroxifenila (H). Fonte: Barbosa (2008)

2.1.4. Compostos minoritários

Além de celulose, hemicelulose e lignina, existem ainda outras classes de compostos que se apresentam em frações muito pequenas, mas que ainda devem ser citadas devido a sua importância, sendo elas as cinzas e os extrativos.

A fração de cinzas tem grande relevância devido a esta ser composta principalmente por Si, Ca, K, Na, Mg S, P, Fe, Mn e Al existindo na forma de óxidos, silicatos, carbonatos, sulfatos, cloretos e fosfatos (Obernberger, Biedermann, Widmann & Riedl, 1997; Raveendran, Ganesh & Khilar, 1995). A presença de tais compostos inorgânicos na linha industrial está associada a diferentes problemas tais como a acumulação em superfícies e corrosão de equipamentos, acrescendo custos de manutenção (Monti, Di Virgilio & Venturi, 2008) de forma que sua presença é muitas vezes indesejada.

Os extrativos compreendem uma ampla classe de compostos orgânicos na forma de flavonóides, proteínas, ceras, terpenos, gorduras, pectinas e alguns sais inorgânicos que tem em comum seu caráter solúvel em alguns tipos de solventes (Neves et al., 2016). Nas plantas, tem ação protetora contra o ataque de insetos e fungos, funcionando também como material de reserva energética e atividade hormonal (Zanuncio, Carvalho, Trugilho & Monteiro, 2014).

Uma vez que a biomassa é composta em sua maior parte de polímeros que podem ser quebrados em hexoses (e.g. glicose, galactose e manose) e pentoses (e.g. xilose e arabinose), é interessante para as indústrias agregar valor a esse resíduo a partir da

produção de combustíveis. Na Figura 4 é apresentado um modelo de como estão dispostos os principais constituintes da parede celular de materiais lignocelulósicos. Observa-se a interação ordenada das cadeias lineares de celulose formando as microfibrilas que, por sua vez, são envoltas pela hemicelulose, que age como intermediária na interação entre celulose e lignina.

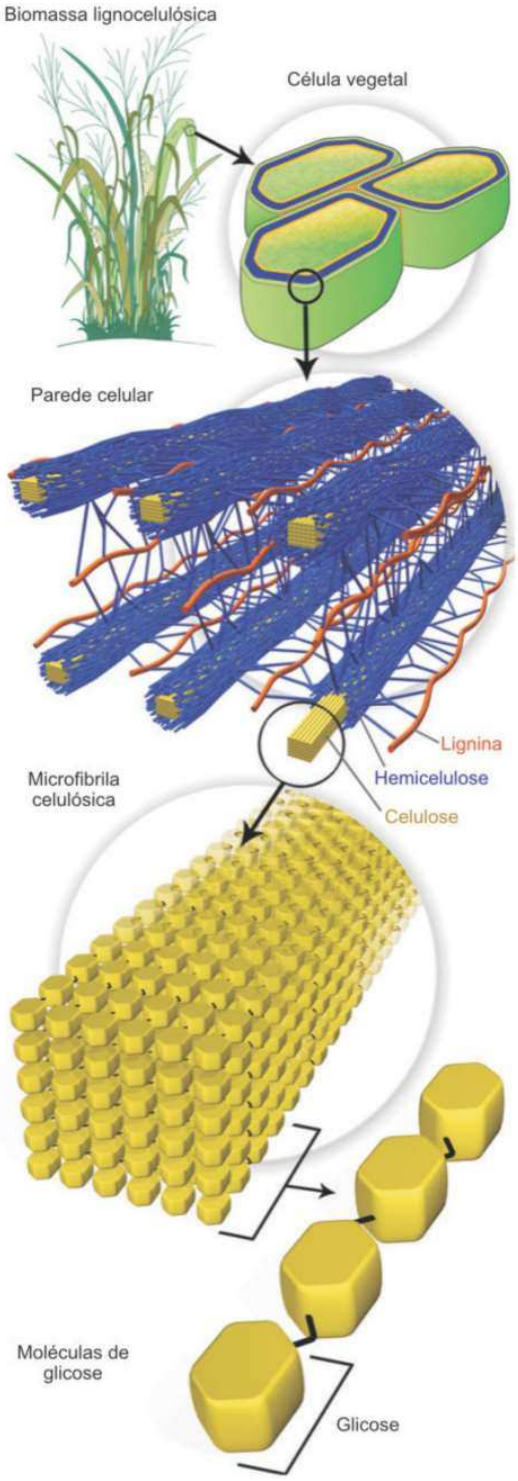

Figura 4. Modelo estrutural da biomassa lignocelulósica. Fonte: Santos (2012)

2.2. Cana-de-açúcar

A cana-de-açúcar (*Saccharum* spp.) é uma gramínea originaria da Oceania e Ásia e foi introduzida no Brasil pelos portugueses durante a colonização (Santos et al., 2013; MAPA, 2016). Pode atingir de 2 a 5 metros de altura e é composta de raízes subterrâneas, caule subdivididos em colmos, folhas e flores. Uma vez extraído o caldo, a palha (i.e., folhas) e o bagaço somam 90% da massa seca da cana-de-açúcar (Santos et al., 2013). A Figura 5 apresenta um modelo da estrutura da cana-de-açúcar.

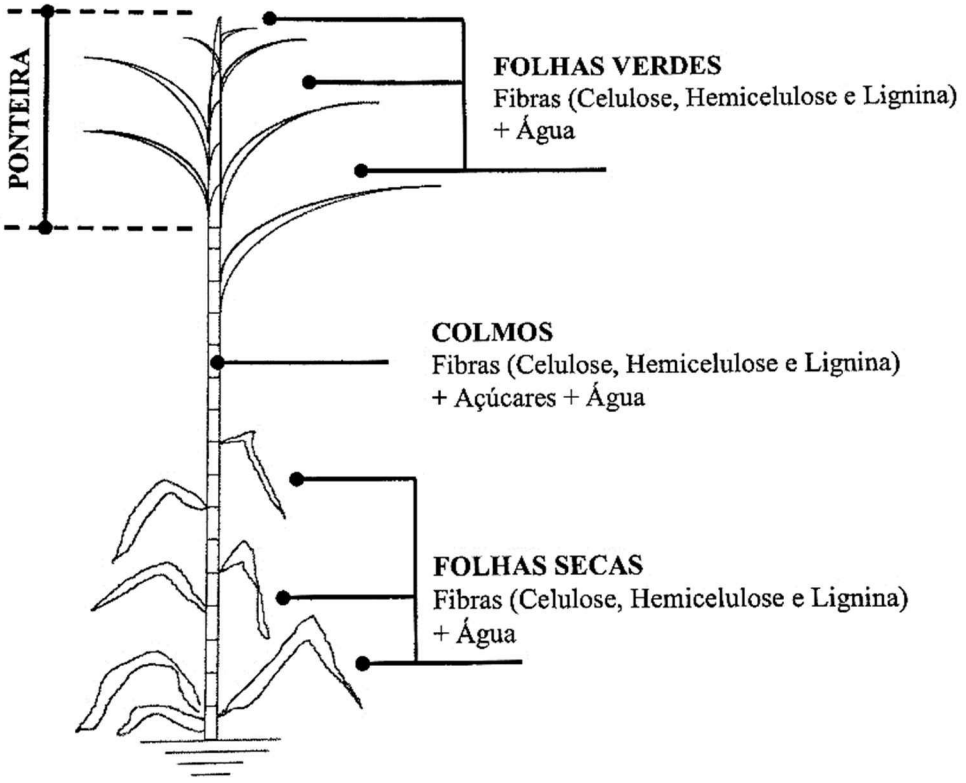

Figura 5. Estrutura da cana-de-açúcar. Fonte: Santos (2013).

Alguns valores atualmente aceitos para sua composição são apresentados na Tabela 1. Tem havido um grande esforço no país para o desenvolvimento e a implementação de novas tecnologias com base na produção de combustíveis renováveis, a partir da cana-de-açúcar. O programa Proálcool (Canilha et al., 2011; Goldemberg et al., 2008; Ogeda & Petri, 2010; Rico, Mercedes & Sauer, 2010; Silva, 2010) é considerado uma referência na produção de etanol 1G a partir dessa cultura. Outros pontos atrativos para o uso da biomassa de cana-de-açúcar se dão em relação aos seus altos teores de glicose e xilose, tornando-a um candidato promissor para produção de biocombustíveis (Batalha et al., 2015), e ao baixo teor de cinzas dessa matéria-prima, se comparado à outros materiais (Cardona et al., 2010), reduzindo assim problemas relacionados a incrustações no maquinário.

Tabela 1. Composição da parede celular de cana-de-açúcar.

	Teor / %
Celulose	40,1-48,6[a,d]
Hemicelulose	25-32[a,d]
Xilose	20,5-27,6[a,b]
Arabinose	2,3-6,3[b]
Galactose	1,6[b]
Manose	0,5-0,6[b]
Lignina	17-24[a,c,d]
Cinzas	0,1-4,0[a,c,d]
Extrativos	1,6-7,5[a,c,d]

[a] de Souza (2013); [b] Gírio (2010); [c] Batalha (2015); [d] de Vasconcelos (2013)

2.3. Difratometria de Raio-X (XRD)

A região do espectro eletromagnético atribuída ao raio-X contempla a radiação com comprimento de onda entre aproximadamente 1 e 100 angstrons. A interação do raio-X com a matéria pode levar à diferentes efeitos tais como a fotoabsorção, espalhamentos Compton e Rayleigh, dentre outros, dependendo da energia dos fótons incididos sobre o material de estudo (Eisberg & Resnick, 1994; Henke, Gullikson & Davis, 1993). O efeito explorado na difratometria de raio-X é o espalhamento Rayleigh, i.e. elástico, da radiação incidida. Uma vez que esse tipo de radiação possui aproximadamente a mesma dimensão da distância entre átomos em arranjos cristalinos, é de grande valor seu uso para investigar tais estruturas.

Como visto em estudos de física ondulatória, a condição para que haja interferência construtiva está diretamente relacionada à diferença de caminho e ao comprimento de onda utilizado. Esta condição é descrita na Equação 1 pela relação de Bragg:

$$n\lambda = 2dsen\theta \qquad \text{(Equação 1)}$$

em que n é um número natural diferente de zero, λ é o comprimento de onda incidente, d a distância inter-atômica entre os planos cristalinos e θ é o ângulo de incidência (Bragg & Bragg, 1913). A interação da radiação com materiais cristalinos pode ser visualizada na Figura 6.

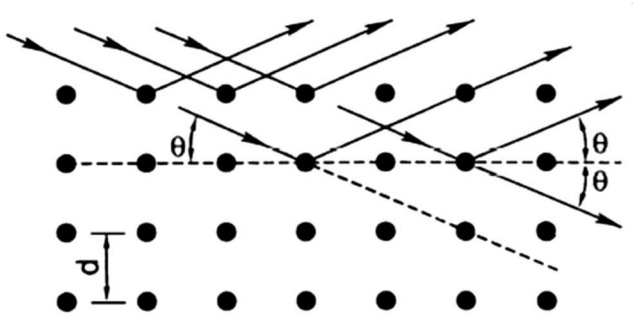

Figura 6. Interação da radiação com os planos atômicos em um cristal.

Assim, estruturas cristalinas, ou seja, de elevada ordenação, apresentarão padrões de difração referentes às interferências construtivas características para as distâncias inter-atômicas de cada plano do material cristalino (Bragg & Bragg, 1913).

Apesar de que medidas de cristalinidade absoluta para celulose em biomassa sejam possíveis (Driemeier & Calligaris, 2011), na maioria dos casos, os benefícios de se obter essa informação não superam as dificuldades experimentais. Assim, são escolhidas abordagens mais práticas onde seja possível realizar comparações diretas entre amostras. O método de Segal et. al. (1959) é um dos métodos onde é possível comparar a quantidade de material cristalino entre amostra e é muito difundido na literatura devido à sua praticidade. Este método usa a relação entre as intensidades de regiões no difratograma associadas às frações cristalina e amorfa para se obter essa relação.

2.4. Cromatografia líquida de alta eficiência (HPLC)

O termo cromatografia surgiu no ano de 1906 a partir dos estudos do botânico M.S Twett ao analisar clorofilas (Collins, 2009). A principal conclusão do trabalho de Twett foi a observação de que durante a eluição de uma amostra contendo diferentes componentes, cada qual apresentará diferentes afinidades com o que atualmente é conhecido como fases móvel e estacionária, de forma que é possível realizar a separação de uma mistura homogênea através deste princípio. A partir disso, a introdução dessa técnica em laboratórios ao redor do mundo ocorreu de diversas formas tais como a cromatografia gasosa (GC) e a cromatografia líquida de alta eficiência (HPLC).

No uso da HPLC, fatores como pressão, fluxo da fase móvel e temperatura da coluna cromatográfica são criteriosamente controlados para uma maior eficiência de separação e quantificação. Para a obtenção do sinal das diferentes frações da separação, diversos tipos de detectores podem ser usados dependendo do tipo de analito, tais como

os detectores de fluorescência, ultravioleta, eletroquímicos ou detectores evaporativos com espalhamento de luz (ELSD).

Os detectores que podem ser usados em HPLC para análise de carboidratos são os detectores de índice de refração (RID), eletroquímico (ED), de massas (MS) e o ELSD. O MS, apesar da alta sensibilidade, possui alto custo e complexidade de uso e pode ter pouca seletividade. O ED é instável e possui baixa precisão apesar da maior seletividade e alta sensibilidade. O RID, apesar da simplicidade e baixo custo, possui baixa sensibilidade e é sensível ao gradiente da fase móvel. O ELSD é intermediário, pois possui uma sensibilidade maior que o RID, alta precisão, é insensível ao gradiente da fase móvel e é mais simples que o ED e o MS. Uma exigência para o uso do ELSD é que a fase móvel possua maior volatilidade que o analito. O ideal é que o analito seja pouco volátil.

O ELSD pode ter seu funcionamento resumido em três etapas principais (Gonzalez, Bianchi, Pereira, Cassiano & Cass, 2011):

1) Nebulização: O eluente, ao sair da coluna, é nebulizado por um fluxo de gás à pressão elevada.

2) Evaporação: As gotículas formadas na etapa de nebulização são evaporadas por um aparato à temperatura controlada de forma que toda a fase móvel seja removida, restando apenas partículas secas do analito.

3) Detecção: O analito, livre da fase móvel, é carreado por um gás até uma cela óptica onde passa por um feixe de luz e a radiação espalhada é medida por um detector. A concentração de amostras pode ser então determinada usando uma curva analítica construída com padrões do analito.

2.5. Infravermelho Próximo (NIR)

A região que compreende a faixa espectral atribuída ao infravermelho próximo varia entre 780-2500 nm e 4000-12820 cm^{-1} (Blanco & Villarroya, 2002; Metrohm, 2013). Nessa região, a absorção da radiação pela amostra ocorre exatamente em comprimentos de onda específicos que correspondam à diferença energética existente entre diferentes níveis vibracionais das moléculas.

Devido aos efeitos da anarmonicidade (quebra de ligação e repulsão entre átomos), que contradizem o modelo clássico para o oscilador harmônico, surge a necessidade de modificar o modelo vibracional. Esta modificação é necessária para explicar os efeitos da interação da radiação com a matéria. Através do modelo do oscilador anarmônico

quântico é possível evidenciar alguns critérios e explicações para as bandas NIR. Nesse tipo de espectroscopia vibracional, a energia dos fótons incididos é relativamente baixa para promover excitações eletrônicas, porém é suficientemente elevada para promover transições vibracionais mais energéticas que a transição fundamental (Pasquini, 2003). São estas as chamadas transições de sobretom. Na prática, pode ser observado que a intensidade do sinal nas regiões de sobretom são diversas ordens de grandeza inferiores se comparada às intensidades da região referente às transições fundamentais, uma vez que estas são menos prováveis de ocorrerem (Blanco & Villarroya, 2002). Dentre outras ocorrências estão a possibilidade de surgimento de bandas de combinação, geradas pela combinação de modos vibracionais. É importante ressaltar que, para que um modo vibracional tenha atividade no NIR, deve haver uma variação no momento de dipolo durante a vibração (Blanco & Villarroya, 2002; Metrohm, 2013; Pasquini, 2003).

Ligações entre um átomo leve e outro mais pesado, como C-H, N-H, O-H e S-H apresentam grande anarmonicidade e são responsáveis pela maioria das bandas observadas nos espectros NIR (Blanco & Villarroya, 2002; Pasquini, 2003). Além disso, os espectros podem conter diferentes informações relacionadas não só a propriedades químicas, mas físicas das amostras como, por exemplo, densidade e viscosidade (Blanco & Villarroya, 2002). Na Figura 7 são identificadas algumas atribuições de bandas observadas na região do espectro NIR.

Apesar da complexidade de um espectro NIR, é possível, através de métodos quimiométricos, obter diversas informações de uso prático. Pode-se citar diversas vantagens tais como a rapidez da análise, praticidade de obtenção dos espectros e a possibilidade de análise tanto em amostras sólidas, líquidas ou gasosas de forma não destrutiva e não invasiva.

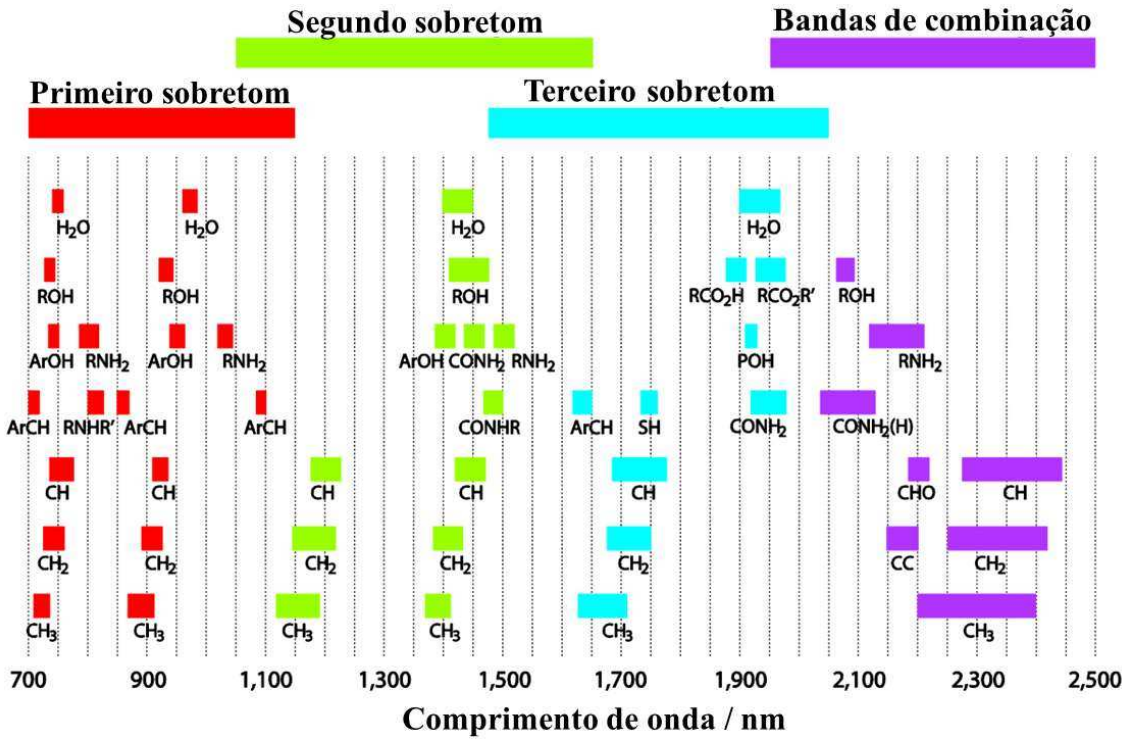

Figura 7. Atribuição de bandas características observadas na região do espectro NIR. Fonte: Adaptado de Metrohm (2013)

2.6. Quimiometria

A Quimiometria pode ser descrita como uma subárea da química que faz uso de métodos matemáticos, estatísticos e computacionais para planejar e otimizar experimentos, interpretar, classificar e realizar previsões a partir de dados de interesse químico (de Souza & Poppi, 2012; Ferreira, 2015). A origem do termo "Quimiometria" data de 1971, sendo relativamente recente mas, ainda assim, já bem estabelecido (Ferreira, 2015). Sua implementação definitiva nas áreas de química e correlatas só foi possível devido aos avanços computacionais ao longo dos anos e da necessidade de, cada vez mais, extrair informações relevantes de grandes quantidades de dados complexos multivariados.

2.6.1. Planejamento Experimental (DOE)

O planejamento experimental (DOE) é um tópico da Quimiometria que tem como principal objetivo identificar e interpretar os fatores (e.g., temperatura, pressão, pH, etc.) individuais e suas interações que influenciam na resposta experimental, além de otimizar a resposta e as condições experimentais buscando, para isso, realizar o menor número de experimentos (Hibbert, 2012). Ao contrário da forma tradicional univariada, na qual cada

fator é estudado individualmente, no DOE vários fatores são estudados e variados de forma simultânea e sistemática. Para isso, busca-se ajustar os dados a modelos lineares ou quadráticos que contenham informações sobre o sistema em estudo (Hibbert, 2012; Teófilo & Ferreira, 2006).

Em planejamentos fatoriais completos em que são investigados *k* fatores com apenas dois níveis para cada fator, é necessário a realização de pelo menos 2^k experimentos se não forem realizadas repetições. É comum utilizar as atribuições –1 e +1 para o nível inferior e superior, respectivamente. Por exemplo, para o fator pH, pode se atribuir -1 para pH 5 e +1 para pH 9. Na Tabela 2 dispõe-se como pode ser montado um planejamento para avaliação de três fatores, de forma que o planejamento seja ortogonal e os cálculos possam ser realizados corretamente. Neste caso também é levado em conta o efeito de possíveis relações sinérgicas ou antagônicas entre os fatores, evidenciadas nas colunas de interações.

Tabela 2. Disposição de um planejamento fatorial 2^3

	Fatores			**Interações**				
Média	**1**	**2**	**3**	**12**	**13**	**23**	**123**	**Resposta**
1	1	1	1	1	1	1	1	R1
1	1	1	-1	1	-1	-1	-1	R2
1	1	-1	1	-1	1	-1	-1	R3
1	1	-1	-1	-1	-1	1	1	R4
1	-1	1	1	-1	-1	1	-1	R5
1	-1	1	-1	-1	1	-1	1	R6
1	-1	-1	1	1	-1	-1	1	R7
1	-1	-1	-1	1	1	1	-1	R8

Separando os valores dispostos nas colunas média, fatores e interações em uma matriz **X**, e as respostas experimentais em um vetor **y**, o coeficiente de cada termo pode ser identificado através de uma abordagem polinomial, neste caso representada pela seguinte relação.

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3 + e$$

Dessa forma, o vetor **b** contendo os valores dos coeficientes que minimizam a soma quadrática dos erros das respostas pode ser encontrado empregando a Equação 2.

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \qquad \text{(Equação 2)}$$

em que **b** é o vetor de regressão, **X** é a matriz contendo a disposição do planejamento e **y** é o vetor contendo as respostas experimentais.

Da mesma forma, esse cálculo pode ser generalizado para outros tipos de DOE. A grande desvantagem dos planejamentos fatoriais completos é que quando muitos fatores são estudados a quantidade de experimentos necessários é elevada e a execução torna-se muitas vezes inviável (Analytical Methods Committee, 2013).

Parte desse problema é solucionado empregando os planejamentos fatoriais fracionados, uma vez que é possível estudar diversos fatores realizando apenas frações dos planejamentos fatoriais completos. Quando aplicados, os planejamentos fracionados incorporam os coeficientes atribuídos às interações de maior ordem aos coeficientes dos fatores principais e aos coeficientes de ordens menores. Uma vez que na maioria dos casos as interações de alta ordem não são significativas, a interpretação dos resultados se mantém, mas com o diferencial de terem sido realizados com um menor número de experimentos (Neto, Scarminio & Bruns, 2001).

2.6.2. Calibração Multivariada

A calibração multivariada pode ser dita como o cavalo de força da Quimiometria, por ser o tópico mais aplicado pelo setor produtivo. Nela são obtidos os chamados "*soft models*" ou modelos flexíveis, construídos a partir de dados empíricos, ao contrário dos "*hard models*" ou modelos rígidos, construídos com embasamento teórico a respeito do sistema de estudo (Ferreira, 2015; Manne, 1987).

O termo calibração refere-se à construção de modelos de regressão capazes de relacionar as respostas instrumentais obtidas da análise de determinadas amostras à propriedades conhecidas das mesmas, obtidas por algum método de referência (Valderrama, Braga & Poppi, 2009). O tipo de calibração pode ser classificado quanto a dimensão das respostas instrumentais obtidas para uma determinada amostra, podendo ser citada a de ordem zero, quando a resposta instrumental é um escalar, de primeira ordem, quando a resposta é um vetor, de segunda ordem, quando a resposta é uma matriz, e terceira ordem, quando a resposta é expressa na forma de um tensor (Valderrama et al., 2009).

Os modelos de calibração multivariada mais usados atualmente são os de primeira ordem. Neste tipo de calibração um vetor contendo centenas de respostas instrumentais é necessário para a construção do modelo (e.g., todos os comprimentos de onda de um espectro de absorbância). Neste contexto, as respostas instrumentais obtidas são usualmente tratadas como variáveis independentes e as propriedades de interesse são tratadas como variáveis dependentes, caracterizando uma regressão inversa.

A vantagem deste tipo de modelo de regressão, também chamada de vantagem de primeira ordem, seria a possibilidade de se analisar amostras complexas mesmo com a presença de diversos interferentes, desde que estes também estejam presentes durante a construção do modelo (Gholivand, Jalalvand, Goicoechea & Skov, 2014).

Uma vez que em muitos dos casos as respostas instrumentais estão altamente correlacionadas entre si e estão presentes em maior número que as amostras, o modelo de regressão linear múltipla (MLR) não pode mais ser aplicado. Dessa forma, para se obter o vetor **b,** que contém os coeficientes da equação, é necessário o uso de regressões mais sofisticadas como, por exemplo, a regressão por componentes principais (PCR) ou a regressão por mínimos quadrados parciais (PLS) (Valderrama et al., 2009).

As etapas de maior importância para se obter o modelo de regressão são:

1) Obter a propriedade química ou física de interesse através do método de referência e organizá-la em um vetor **y**;
2) Obter as respostas instrumentais através do método alternativo e organizá-las em uma matriz **X**;
3) Remover as amostras anômalas (*outliers*), se identificadas pelo analista;
4) Aplicar tratamentos matemáticos ao conjunto de dados multivariado de forma a evidenciar a informação relevante e remover o ruído presente neste;
5) Separar as amostras em um conjunto de treinamento, usado para construção do modelo, e em um conjunto de previsão para validação do modelo construído;
6) Usar métodos de regressão (e.g., PCR ou PLS) para construir o melhor modelo com base nos parâmetros estatísticos obtidos a partir da validação cruzada;
7) Construir o modelo com o método de regressão escolhido utilizando o conjunto de treinamento;
8) Avaliar o modelo com base nos parâmetros estatísticos obtidos e das previsões das amostras do conjunto externo.

Por fim, é interessante para aspectos práticos e um bom funcionamento do modelo, que todas as amostras utilizadas na construção e validação do modelo tenham características físicas e químicas similares às amostras das quais se desejam realizar previsões futuramente (ASTM, 2000). Uma vez que o método de regressão PLS tem sido o mais utilizado e fornece em geral melhores resultados, este será descrito em maiores detalhes (Martins, Teófilo & Ferreira, 2010).

2.6.2.1. Regressão por Quadrados Mínimos Parciais (PLS)

O PLS refere-se a um método de regressão derivado dos estudos realizados no ano de 1975 por Herman Wold na área de econometria (Lorber, Wangen & Kowalski, 1987; Manne, 1987; Martins et al., 2010). Sua aplicação na quimiometria deve-se a seu filho, Svante, alguns anos depois (Ferreira, 2015; Geladi & Kowalski, 1986). O grande diferencial deste método deve-se à projeção dos dados originais em um subespaço de dimensão reduzida, de forma a eliminar a informação irrelevante (Ferreira, 2015). Diferente do método PCR, o PLS leva em conta a variância da variável dependente para a geração do subespaço a partir das novas componentes, também chamadas de variáveis latentes (Ferreira, 2015; Sekulic et al., 1993). Apesar de ser muito utilizado para resolução de diversos problemas, e fornecer em geral melhores resultados que outros métodos de regressão, o PLS ainda não é totalmente entendido e ainda é alvo de opiniões diversas (Bro & Eldén, 2009; Lorber et al., 1987; Wold et al., 2009). Diversos algoritmos PLS estão disponíveis na literatura podendo ser citados o NIPALS, Kernel, SIMPLS, Krylov, PLS bidiagonal, dentre outros (Andersson, 2009; Indahl, 2014; Martins et al., 2010). Em um estudo realizado por Martins et. al. (2010), os autores mostraram que o algoritmo PLS bidiagonal é computacionalmente mais eficiente que os demais, de forma que este foi utilizado neste trabalho.

O método PLS bidiagonal consiste na decomposição da matriz de variáveis independentes $\mathbf{X}(I\,x\,J)$ de forma que $\mathbf{X} = \mathbf{U}\mathbf{R}\mathbf{V}^{\mathrm{T}}$, onde $\mathbf{U}(I\,x\,J)$ e $\mathbf{V}(I\,x\,J)$ são ortonormais e $\mathbf{R}(J\,x\,J)$ é uma matriz bidiagonal (Martins et al., 2010; Teófilo, Martins & Ferreira, 2009). Segundo Teófilo et. al. (2009), este algoritmo pode ser resumido nas seguintes etapas:

1) Inicialize o algoritmo para a primeira componente

$$\mathbf{v}_1 = \mathbf{X}^{\mathbf{T}}\mathbf{y} / \| \mathbf{X}^{\mathbf{T}}\mathbf{y} \| \;;\; \alpha_1 \mathbf{u}_1 = \mathbf{X}\mathbf{v}_1$$

2) Calcule os seguintes valores para $n = 2, ..., h$ variáveis latentes

$$\gamma_{n\text{-}1}\mathbf{v}_n = \mathbf{X}^{\mathbf{T}}\mathbf{u}_{n\text{-}1} - \alpha_{n\text{-}1}\mathbf{v}_{n\text{-}1}$$

$$\alpha_n \mathbf{u}_n = \mathbf{X}\mathbf{v}_n - \gamma_{n-1}\mathbf{u}_{n\text{-}1}$$

$$\text{Com } \mathbf{V}_h = \left(\mathbf{v}_1, ..., \mathbf{v}_h\right) \text{e } \mathbf{U}_h = \left(\mathbf{u}_1, ..., \mathbf{u}_h\right)$$

$$\mathbf{R}_h = \begin{pmatrix} \alpha_1 & \gamma_1 & 0 & 0 & 0 \\ 0 & \alpha_2 & \gamma_2 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & \alpha_{h\text{-}1} & \gamma_{h\text{-}1} \\ 0 & 0 & 0 & 0 & \alpha_h \end{pmatrix}$$

De forma que, uma vez efetuada a truncagem em *h* variáveis latentes, é possível obter o vetor de regressão através da pseudoinversa de Moore-Penrose de **X** por:

$$\mathbf{y} = \mathbf{Xb} \rightarrow \mathbf{y} = \mathbf{U}_h\mathbf{R}_h\mathbf{V}_h^{\mathrm{T}}\mathbf{b} \rightarrow \mathbf{b} = \mathbf{V}_h\mathbf{R}_h^{-1}\mathbf{U}_h^{\mathrm{T}}\mathbf{y}$$

Para se determinar o número de variáveis latentes ideal (*h*) a ser utilizado na obtenção do vetor **b**, deve-se tomar cuidado durante a truncagem para que não ocorra falta de ajuste, tampouco um sobre ajuste do modelo (Ferreira, 2015; Pell, Ramos & Manne, 2007; Sekulic et al., 1993). Para isso utiliza-se o método da validação cruzada, onde o modelo é avaliado por diversos parâmetros estatísticos de forma que o valor de *h* seja escolhido frente a esses resultados. Neste trabalho, para a escolha de *h* foram avaliados os parâmetros raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*), coeficiente de correlação de validação cruzada (*RCV*) e índice de desempenho do desvio (*RPD*), os quais serão descritos na próxima seção. Existem diversas formas de se separar os conjuntos para a validação cruzada, tais como os métodos *Leave One Out*, *Contiguous Blocks*, *Venetian Blinds*, *Random Subsets*, dentre outros, muitas vezes já implementados em *softwares* como o PLS_Toolbox da empresa Eigenvector, porém estes não serão abordados em mais detalhes.

2.6.2.2. Figuras de mérito em calibração multivariada

Para se avaliar a qualidade de um modelo de calibração multivariada ou de qualquer outro método proposto, deve ser realizada a validação do mesmo. Para isso existem diversos parâmetros que podem ser utilizados afim de se obter informações quantitativas neste quesito, parâmetros estes denominados figuras de mérito. Dentre as diversas figuras de mérito existentes para avaliação de modelos de calibração de primeira ordem pode-se citar a raiz quadrada do erro quadrático médio (*RMSE*) e coeficiente de correlação (*R*), os quais são calculados através das Equações 3 e 4, respectivamente.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{I}} \qquad \text{(Equação 3)}$$

$$R = \frac{\sum_{i=1}^{I} (\hat{y}_i - \overline{\hat{y}})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{I} (\hat{y}_i - \overline{\hat{y}})^2 \sum_{i=1}^{I} (y_i - \overline{y})^2}} \qquad \text{(Equação 4)}$$

sendo $\hat{y}$ e $\overline{\hat{y}}$ o valor estimado e médio estimado, y e $\overline{y}$ os valores obtidos e valores médios obtidos. Quando a validação cruzada interna (*CV*) é usada, *I* representa o número de amostras no conjunto de calibração, e o erro e coeficiente de correlação são chamados de raiz quadrada do erro quadrático médio de validação cruzada (*RMSECV*) e coeficiente de correlação de validação cruzada (*RCV*), respectivamente. Quando a validação externa é usada, *I* representa o número de amostras de previsão (*P*) e, neste caso, os coeficientes de correlação e o erro são nomeados coeficiente de correlação de previsão (*RP*) e raiz quadrada do erro quadrático médio de previsão (*RMSEP*), respectivamente.

O modelo também pode ser avaliado em termos do índice de desempenho do desvio (*RPD*), descrito na Equação 5. Este representa a razão entre o desvio padrão (*SD*) dos valores da propriedade analisada e a raiz quadrada do erro quadrático médio de validação cruzada. Esta estatística fornece uma base de padronização do erro padrão da predição (Williams & Sobering, 1993).

$$RPD = \frac{SD}{RMSECV} \qquad \text{(Equação 5)}$$

Os parâmetros sensibilidade (*SEN*), seletividade (*SEL*), sensibilidade analítica (γ) e limite de detecção (*LOD*) também são usados como figuras de mérito e são descritos segundo as Equações 6 a 9, respectivamente.

A sensibilidade refere-se à fração do sinal responsável pelo acréscimo de uma unidade de concentração à propriedade de interesse e é apresentada na Equação 6 (Valderrama et al., 2009).

$$SEN = \frac{1}{\| \mathbf{b} \|} \qquad \text{(Equação 6)}$$

em que $\| \mathbf{b} \|$ é a norma euclidiana do vetor de coeficientes de regressão do modelo de calibração (Ferreira, 2015; Valderrama et al., 2009).

A partir desse ponto, é necessário introduzir o conceito de sinal analítico líquido (NAS) para realizar o cálculo das figuras de mérito. O NAS representa o sinal instrumental referente ao componente de interesse de forma que esse seja ortogonal ao sinal dos interferentes (Ferreira, 2015; Valderrama et al., 2009).

A seletividade, descrita na Equação 7, pode ser entendida como a fração do sinal instrumental referente ao componente de interesse que está sobreposto ao sinal dos interferentes, sendo a fração do sinal útil para a previsão (Ferreira, 2015; Valderrama et al., 2009).

$$SEL = \frac{nas_{k,i}}{\| \mathbf{x}_{k,i} \|} \quad \text{(Equação 7)}$$

em que $nas_{k,i}$ é o valor escalar do sinal analítico líquido para a amostra *i* e $\| \mathbf{x}_{k,i} \|$ representa a norma euclidiana do vetor de resposta instrumental para a amostra *i* (Rambo, Amorim & Ferreira, 2013; Valderrama et al., 2009). Uma vez que é obtido um valor de seletividade por amostra, é comum expressar esta como a média dos resultados.

A sensibilidade analítica é outra figura de mérito presente em diversos trabalhos e é descrita na Equação 8.

$$\gamma = \frac{SEN}{\| \partial_x \|} \quad \text{(Equação 8)}$$

em que *SEN* é a sensibilidade e $\| \partial_x \|$ é a norma euclidiana do desvio padrão do sinal de referência (Ferreira, 2015; Valderrama et al., 2009). Uma vez que esta é apresentada na unidade "concetração^{-1}", o inverso da sensibilidade analítica pode ser analisado de forma mais intuitiva, representando a menor variação na variável dependente que pode ser distinguida pelo método proposto (Ferreira, 2015).

Diferentes formas para o cálculo do limite de detecção são descritas na literatura (Allegrini & Olivieri, 2014; Ortiz et al., 2003; Valderrama et al., 2009). Neste trabalho optou-se pelo método desenvolvido por Ortiz et. al. (2003), uma vez que este atende às demandas da comunidade cientifica em termos da implementação dos erros tipo I (α) e tipo II (β) no cálculo das figuras de mérito (Ortiz et al., 2003).

$$LOD = \frac{\Delta(\alpha, \beta) w_{y_0} \sigma}{\hat{a}} \quad \text{(Equação 9)}$$

em que σ é o desvio padrão do resíduo da regressão, w_{y_0} é uma função da posição dos padrões e $\hat{a}$ é o coeficiente da regressão linear do método de Ortiz (Ortiz et al., 2003).

2.6.2.3. Seleção de variáveis

Apesar de métodos de compressão de dados como PCR e PLS serem eficientes na construção de modelos de calibração multivariada a partir de um grande número de

variáveis independentes, grande parte da informação modelada pode ser irrelevante ou constituída de ruído experimental (Teófilo et al., 2009). Dessa forma, para a construção de modelos mais robustos e de mais fácil interpretação, é importante a aplicação de métodos para selecionar as variáveis. Há diversos métodos de seleção de variáveis disponíveis na literatura e, neste trabalho, dois métodos que estão sendo muito usados atualmente serão descritos.

2.6.2.3.1. Algoritmo Genético (GA)

Com crescente popularidade na comunidade científica, o Algoritmo Genético (GA) é um algoritmo que foi proposto no início dos anos 1960 como uma alternativa para se solucionar problemas de otimização (Cong & Li, 1994; Niazi & Leardi, 2012). Para isso, são aplicados os fundamentos da teoria evolucionaria de Darwin no sentido que, fazendo uma analogia à uma classe de seres vivos, as variáveis mais aptas a realizar previsões têm seu "código genético" passado para as próximas gerações de forma que as variáveis de maior relevância têm maior chance de sobreviver ao processo evolucionário (Cong & Li, 1994; Riccardo Leardi, 2001; Riccardo Leardi & Lupiáñez González, 1998). Segundo Niazi e Leardi (2012), o GA pode ser resumido em cinco principais etapas:

1) Codificação das variáveis: Consiste em traduzir a informação contida nas variáveis independentes de modo que cada variável corresponderia a um gene e, um conjunto de variáveis à um cromossomo. Isso pode ser realizado convertendo estes valores para código binário, por exemplo;
2) Inicialização da População: Determina-se o tamanho *N* da população inicial e a estrutura de cada cromossomo é determinada de uma forma totalmente aleatória;
3) Avaliação da Resposta: Para cada cromossomo, a sequência de genes é avaliada e classificada;
4) Reprodução: É criada uma nova população de *N* cromossomos com base na combinação dos cromossomos com melhor avaliação. A cada par combinado é dada origem a um par de descendentes com material genético combinado dos progenitores;
5) Mutação: Consiste na modificação de um bit de um gene de forma aleatória conforme uma taxa pré-estabelecida.

A Figura 8 apresenta um esquema proposto por Niazi (2012) de forma a ilustrar didaticamente as principais etapas envolvidas nos GAs. Para realização deste trabalho, o GA proposto por Leardi et. al. (1992) foi utilizado.

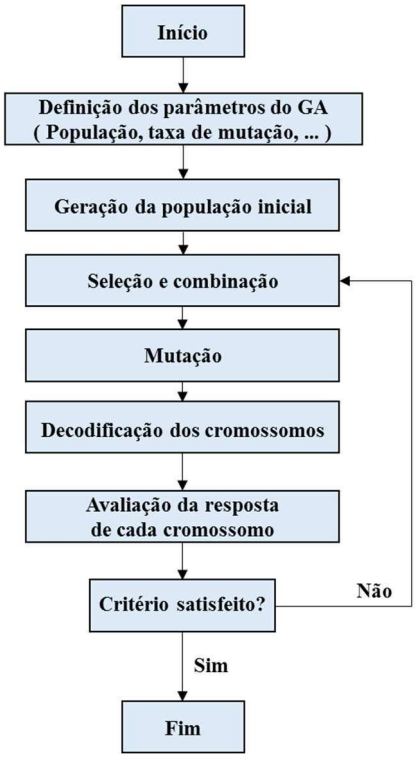

Figura 8. Esquema geral do Algoritmo Genético (GA). Fonte: Adaptado de Niazi (2012)

2.6.2.3.2. Seleção dos Preditores Ordenados (OPS)

Recentemente, Teófilo et. al. (2009) apresentaram um novo método de seleção de variáveis com base no uso de um vetor informativo para auxiliar na escolha das variáveis de maior importância. O método seleção dos preditores ordenados (OPS) utiliza diferentes vetores obtidos através de cálculos usando as variáveis independentes e dependentes e pode ser resumido nas seguintes etapas, ilustradas na Figura 9:

1) Escolhe-se um vetor ou combinação de vetores para ser usado como vetor informativo;
2) As variáveis originais são destacadas segundo os valores absolutos dos elementos do vetor informativo;
3) As variáveis destacadas são ordenadas de forma decrescente de importância;
4) Uma janela inicial e incrementos são definidos de forma que diversos modelos são construídos e cross-validados. Esse processo continua até que todas as variáveis ou uma porcentagem definida destas tenha sido incluída para a construção dos modelos.

5) O conjunto de variáveis que fornecer os melhores parâmetros estatísticos durante a validação cruzada representa o conjunto com melhor capacidade preditiva e é escolhido.

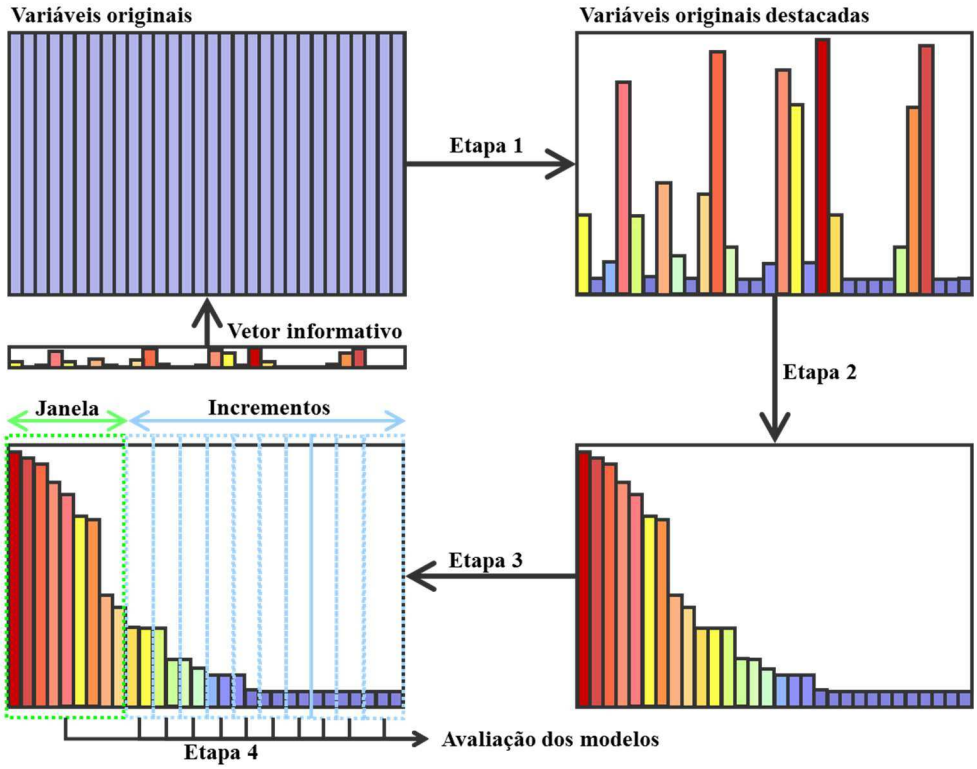

Figura 9. Esquema do algoritmo OPS. Fonte: Teófilo (2009)

É importante ressaltar que o número ideal de variáveis latentes para a construção do vetor de regressão com maior caráter preditivo pode não ser o mesmo daquele para a construção do vetor com maior caráter informativo. Dessa forma, é necessário definir um número de variáveis latentes para a construção do modelo (*h*Mod) e outro para a construção do vetor informativo (*h*OPS) (Teófilo et al., 2009). Para o estudo do *h*OPS ideal, é fixado um valor máximo (*n*) desejado para *h*OPS e são construídos diversos vetores informativos utilizando desde *h*Mod até *n* variáveis latentes. Para cada vetor informativo construído, as variáveis são destacadas, ordenadas e o estudo de conjuntos de variáveis é realizado usando o vetor construído com *h*Mod para realizar as previsões. Para cada valor de *h*OPS estudado, armazena-se o menor valor de *RMSECV* obtido no estudo dos conjuntos. O valor de *h*OPS ideal é aquele que fornecer o menor valor de *RMSECV*.

O algoritmo OPS encontra-se disponível no endereço http://www.deq.ufv.br/chemometrics.

3. Referências Bibliográficas

Allegrini, F. & Olivieri, A. C. (2014). IUPAC-consistent approach to the limit of detection in partial least-squares calibration. Analytical Chemistry, 86(15), 7858–7866. http://doi.org/10.1021/ac501786u

Analytical Methods Committee, A. N. 55. (2013). Experimental design and optimisation (4): Plackett–Burman designs. Analytical Methods, 5(8), 1901. http://doi.org/10.1039/c3ay90020g

Andersson, M. (2009). A comparison of nine PLS1 algorithms. Journal of Chemometrics, 23(10), 518–529. http://doi.org/10.1002/cem.1248

ASTM. (2000). E 1655 - Standard Practices for Infrared Multivariate Quantitative Analysis. http://doi.org/10.1520/E1655-05R12.2

Barbosa, L. C. A., Maltha, C. R. A., Silva, V. L. & Colodette, J. L. (2008). Determinação da relação siringila/guaiacila da lignina em madeiras de eucalipto por pirólise acoplada à cromatografia gasosa e espectrometria de massas (PI CG/EM). Química Nova, 31(8), 2035–2041. http://doi.org/10.1590/S0100-40422008000800023

Batalha, L. A. R., Han, Q., Jameel, H., Chang, H., Colodette, J. L. & Borges Gomes, F. J. (2015). Production of fermentable sugars from sugarcane bagasse by enzymatic hydrolysis after autohydrolysis and mechanical refining. Bioresource Technology, 180, 97–105. http://doi.org/10.1016/j.biortech.2014.12.060

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

Boerjan, W., Ralph, J. & Baucher, M. (2003). Lignin biosynthesis. Annu Rev Plant Biol, 54(1), 519–546. http://doi.org/10.1146/annurev.arplant.54.031902.134938

Bragg, W. H. & Bragg, W. L. (1913). The Reflection of X-rays by Crystals. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 88(605), 428–438. http://doi.org/10.1098/rspa.1913.0040

Bro, R. & Eldén, L. (2009). PLS works. Journal of Chemometrics, 23(2), 69–71. http://doi.org/10.1002/cem.1177

Canilha, L., Santos, V. T. O., Rocha, G. J. M., Almeida e Silva, J. B., Giulietti, M., Silva, S. S., … Carvalho, W. (2011). A study on the pretreatment of a sugarcane bagasse sample with dilute sulfuric acid. Journal of Industrial Microbiology & Biotechnology, 38(9), 1467–1475. http://doi.org/10.1007/s10295-010-0931-2

Cardona, C. A., Quintero, J. A. & Paz, I. C. (2010). Production of bioethanol from sugarcane bagasse: Status and perspectives. Bioresource Technology, 101(13), 4754–4766. http://doi.org/10.1016/j.biortech.2009.10.097

Chen, S.-F., Danao, M.-G. C., Singh, V. & Brown, P. J. (2014). Determining sucrose and glucose levels in dual-purpose sorghum stalks by Fourier transform near infrared (FT-NIR) spectroscopy. Journal of the Science of Food and Agriculture, 94(12), 2569–2576. http://doi.org/10.1002/jsfa.6606

Colares, C. J. G., Pastore, T. C. M., Coradin, V. T. R., Camargos, J. A. A., Moreira, A. C. O., Rubim, J. C. & Braga, J. W. B. (2015). Exploratory analysis of the distribution of lignin and cellulose in woods by Raman imaging and chemometrics. Journal of the Brazilian Chemical Society, 26(6), 1297–1305. http://doi.org/10.5935/0103-5053.20150096

Collins, C. H. (2009). Michael Tswett e o "nascimento" da Cromatografia. Scientia Chromatographica, 1(1), 20.

CONAB - Companhia Nacional de Abastecimento. (2016). Acompanhamento da safra brasileira de cana-de-açúcar. Safra 2015/16 - Quarto levantamento. Retrieved from http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16_04_14_09_06_31_boletim_cana_portugues_-_4o_lev_-_15-16.pdf

Cong, P. & Li, T. (1994). Numeric genetic algorithm Part I. Theory, algorithm and simulated experiments. Analytica Chimica Acta, 293(1–2), 191–203. http://doi.org/10.1016/0003-2670(94)00079-4

Santos, F., Colodette, J. L. & Queiroz, J. H. de. (2013). Bioenergia e Biorrefinaria: Cana-de-Açúcar e Espécies Florestais. Viçosa, MG.

de Souza, A. M. & Poppi, R. J. (2012). Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: Um tutorial, parte I. Química Nova, 35(1), 223–229. http://doi.org/10.1590/S0100-40422012000100039

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

de Vasconcelos, S. M., Santos, A. M. P., Rocha, G. J. M. & Souto-Maior, A. M. (2013). Diluted phosphoric acid pretreatment for production of fermentable sugars in a

sugarcane-based biorefinery. Bioresource Technology, 135, 46–52. http://doi.org/10.1016/j.biortech.2012.10.083

Driemeier, C. & Calligaris, G. A. (2011). Theoretical and experimental developments for accurate determination of crystallinity of cellulose I materials. Journal of Applied Crystallography, 44(1), 184–192. http://doi.org/10.1107/S0021889810043955

Eisberg, R. & Resnick, R. (1994). Física Quântica - Átomos, Moléculas, Sólidos, Núcleos e Partículas (9th ed.).

Ferreira, M. M. C. (2015). Quimiometria - Conceitos, Métodos e Aplicações. (E. da Unicamp, Ed.). Campinas, SP.

Geladi, P. & Kowalski, B. (1986). Partial Least-Squares Regression - a Tutorial. Analytica Chimica Acta, 185, 1–17. http://doi.org/10.1016/0003-2670(86)80028-9

Gholivand, M.-B., Jalalvand, A. R., Goicoechea, H. C. & Skov, T. (2014). Chemometrics-assisted simultaneous voltammetric determination of ascorbic acid, uric acid, dopamine and nitrite: Application of non-bilinear voltammetric data for exploiting first-order advantage. Talanta, 119, 553–563. http://doi.org/10.1016/j.talanta.2013.11.028

Gírio, F. M., Fonseca, C., Carvalheiro, F., Duarte, L. C., Marques, S. & Bogel-Łukasik, R. (2010). Hemicelluloses for fuel ethanol: A review. Bioresource Technology, 101(13), 4775–4800. http://doi.org/10.1016/j.biortech.2010.01.088

Goldemberg, J., Coelho, S. T. & Guardabassi, P. (2008). The sustainability of ethanol production from sugarcane. Energy Policy, 36(6), 2086–2097. http://doi.org/10.1016/j.enpol.2008.02.028

Gonzalez, M. H., Bianchi, S. R., Pereira, C. D., Cassiano, N. M. & Cass, Q. B. (2011). Detector evaporativo com espalhamento de luz: princípios de operação e aplicações em cromatografia líquida de alta eficiência. Scientia Chromatographica, 3(4), 315–325. http://doi.org/10.4322/sc.2011.019

Gutiérrez-Rojas, I., Moreno-Sarmiento, N. & Montoya, D. (2015). Mecanismos y regulación de la hidrólisis enzimática de celulosa en hongos filamentosos: casos clásicos y nuevos modelos. Revista Iberoamericana de Micología, 32(1), 1–12. http://doi.org/10.1016/j.riam.2013.10.009

Hendriks, A. T. W. M. & Zeeman, G. (2009). Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology, 100(1), 10–18. http://doi.org/10.1016/j.biortech.2008.05.027

Henke, B. L., Gullikson, E. M. & Davis, J. C. (1993). X-Ray Interactions: Photoabsorption, Scattering, Transmission, and Reflection at E = 50-30,000 eV, Z = 1-92. Atomic Data and Nuclear Data Tables, 54(2), 181–342. http://doi.org/10.1006/adnd.1993.1013

Hibbert, D. B. (2012). Experimental design in chromatography: A tutorial review. Journal of Chromatography B, 910, 2–13. http://doi.org/10.1016/j.jchromb.2012.01.020

Himmel, M. E., Ding, S.-Y., Johnson, D. K., Adney, W. S., Nimlos, M. R., Brady, J. W. & Foust, T. D. (2007). Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. Science, 315(5813), 804–807. http://doi.org/10.1126/science.1137016

Huang, Q., Yan, Q., Fu, J., Lv, X., Xiong, C., Lin, J. & Liu, Z. (2016). Comparative study of different alcoholate pretreatments for enhanced enzymatic hydrolysis of sugarcane bagasse. Bioresource Technology, 211, 464–471. http://doi.org/10.1016/j.biortech.2016.03.067

Indahl, U. G. (2014). The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modelling. Journal of Chemometrics, 28(3), 168–180. http://doi.org/10.1002/cem.2589

Jiang, Z. H., Yang, Z., So, C. L. & Hse, C. Y. (2007). Rapid prediction of wood crystallinity in Pinus elliotii plantation wood by near-infrared spectroscopy. Journal of Wood Science, 53(5), 449–453. http://doi.org/10.1007/s10086-007-0883-y

Ju, X., Bowden, M., Brown, E. E. & Zhang, X. (2015). An improved X-ray diffraction method for cellulose crystallinity measurement. Carbohydrate Polymers, 123, 476–481. http://doi.org/10.1016/j.carbpol.2014.12.071

Kim, J. S., Lee, Y. Y. & Kim, T. H. (2015). A review on alkaline pretreatment technology for bioconversion of lignocellulosic biomass. Bioresource Technology, 199, 42–48. http://doi.org/10.1016/j.biortech.2015.08.085

Klemm, D., Heublein, B., Fink, H. P. & Bohn, A. (2005). Cellulose: Fascinating biopolymer and sustainable raw material. Angewandte Chemie - International Edition, 44(22), 3358–3393. http://doi.org/10.1002/anie.200460587

Kumar, P., Barrett, D. M., Delwiche, M. J. & Stroeve, P. (2009). Methods for Pretreatment of Lignocellulosic Biomass for Efficient Hydrolysis and Biofuel Production. Industrial & Engineering Chemistry Research, 48(8), 3713–3729. http://doi.org/10.1021/ie801542g

Lavarack, B. P., Griffin, G. J. & Rodman, D. (2002). The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. Biomass and Bioenergy, 23(5), 367–380. http://doi.org/10.1016/S0961-9534(02)00066-1

Leardi, R. (2001). Genetic algorithms in chemometrics and chemistry: A review. Journal of Chemometrics, 15(7), 559–569. http://doi.org/10.1002/cem.651

Leardi, R., Boggia, R. & Terrile, M. (1992). Genetic Algorithms As A Strategy for Feature-Selection. J.Chemomet., 6(July), 267–281. http://doi.org/10.1002/cem.1180060506

Leardi, R. & Lupiáñez González, A. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. Chemometrics and Intelligent Laboratory Systems, 41(2), 195–207. http://doi.org/10.1016/S0169-7439(98)00051-3

Lorber, A., Wangen, L. E. & Kowalski, B. R. (1987). A theoretical foundation for the PLS algorithm. Journal of Chemometrics, 1(1), 19–31. http://doi.org/10.1002/cem.1180010105

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

Manne, R. (1987). Analysis of two partial-least-squares algorithms for multivariate calibration. Chemometrics and Intelligent Laboratory Systems, 2(1–3), 187–197. http://doi.org/10.1016/0169-7439(87)80096-5

MAPA. (2016). Cana-de-açúcar. Retrieved August 1, 2016, from http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar

Martins, J. P. A., Teófilo, R. F. & Ferreira, M. M. C. (2010). Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. Journal of Chemometrics, 24(6), n/a-n/a. http://doi.org/10.1002/cem.1309

Metrohm. (2013). NIR Spectroscopy - A guide to near-infrared spectroscopic analysis of industrial manufacturing processes. Herisau, Switzerland.

Mohanty, A. K., Misra, M. & Hinrichsen, G. (2000). Biofibres, biodegradable polymers and biocomposites: An overview. Macromolecular Materials and Engineering, 276–277(1), 1–24. http://doi.org/10.1002/(SICI)1439-2054(20000301)276:1<1::AID-MAME1>3.0.CO;2-W

Monti, A., Di Virgilio, N. & Venturi, G. (2008). Mineral composition and ash content of six major energy crops. Biomass and Bioenergy, 32(3), 216–223. http://doi.org/10.1016/j.biombioe.2007.09.012

Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y. Y., Holtzapple, M. & Ladisch, M. (2005). Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresource Technology, 96(6), 673–686. http://doi.org/10.1016/j.biortech.2004.06.025

Neto, B. de B., Scarminio, I. S. & Bruns, R. E. (2001). COMO FAZER EXPERIMENTOS - PESQUISA E DESENVOLVIMENTO NA CIÊNCIA E NA INDÚSTRIA. Campinas, SP, Brasil: Editora da Unicamp.

Neves, P. V., Pitarelo, A. P. & Ramos, L. P. (2016). Production of cellulosic ethanol from sugarcane bagasse by steam explosion: Effect of extractives content, acid catalysis and different fermentation technologies. Bioresource Technology, 208, 184–194. http://doi.org/10.1016/j.biortech.2016.02.085

Niazi, A. & Leardi, R. (2012). Genetic algorithms in chemometrics. Journal of Chemometrics, 26(6), 345–351. http://doi.org/10.1002/cem.2426

Obernberger, I., Biedermann, F., Widmann, W. & Riedl, R. (1997). Concentrations of inorganic elements in biomass fuels and recovery in the different ash fractions. Biomass and Bioenergy, 12(3), 211–224. http://doi.org/10.1016/S0961-9534(96)00051-7

Ogeda, T. L. & Petri, D. F. S. (2010). Hidrólise Enzimática de Biomassa. Química Nova, 33(7), 1549–1558. http://doi.org/10.1590/S0100-40422010000700023

Ortiz, M. ., Sarabia, L. ., Herrero, A., Sánchez, M. ., Sanz, M. ., Rueda, M. ., … Meléndez, M. . (2003). Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares. Chemometrics and Intelligent Laboratory Systems, 69(1–2), 21–33. http://doi.org/10.1016/S0169-7439(03)00110-2

Park, S., Baker, J. O., Himmel, M. E., Parilla, P. A. & Johnson, D. K. (2010). Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance. Biotechnology for Biofuels, 3(1), 10. http://doi.org/10.1186/1754-6834-3-10

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

Pell, R. J., Ramos, L. S. & Manne, R. (2007). The model space in partial least squares regression. Journal of Chemometrics, 21(3–4), 165–172. http://doi.org/10.1002/cem.1067

Peng, Y., Gardner, D. J., Han, Y., Kiziltas, A., Cai, Z. & Tshabalala, M. A. (2013). Influence of drying method on the material properties of nanocellulose I: Thermostability and crystallinity. Cellulose, 20(5), 2379–2392. http://doi.org/10.1007/s10570-013-0019-z

Pereira, L. G., Dias, M. O. S., MacLean, H. L. & Bonomi, A. (2015). Investigation of uncertainties associated with the production of n-butanol through ethanol catalysis in sugarcane biorefineries. Bioresource Technology, 190, 242–250. http://doi.org/10.1016/j.biortech.2015.04.095

Rabelo, S. C. (2010). Avaliação e otimização de pré-tratamentos e hidrólise enzimatica do bagaço de cana-de-açúcar para a produção de etanol de segunda geração. UNICAMP.

Ragauskas, A. J. (2014). Materials for Biofuels. Retrieved from https://books.google.com.br/books?id=2p66CgAAQBAJ&dq=crystallinity+of+cellulose+as+determined+by+cp/mas+nmr+and+xrd+methods&hl=pt-BR&source=gbs_navlinks_s

Rambo, M. K. D., Amorim, E. P. & Ferreira, M. M. C. (2013). Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. Analytica Chimica Acta, 775, 41–49. http://doi.org/10.1016/j.aca.2013.03.015

Rambo, M. K. D. & Ferreira, M. M. C. (2015). Determination of Cellulose Crystallinity of Banana Residues Using Near Infrared Spectroscopy and Multivariate Analysis. Journal of the Brazilian Chemical Society, 26(7), 1491–1499. http://doi.org/10.5935/0103-5053.20150118

Raveendran, K., Ganesh, A. & Khilar, K. C. (1995). Influence of mineral matter on biomass pyrolysis characteristics. Fuel, 74(12), 1812–1822. http://doi.org/10.1016/0016-2361(95)80013-8

Rico, J. A. P., Mercedes, S. S. P. & Sauer, I. L. (2010). Genesis and consolidation of the Brazilian bioethanol: A review of policies and incentive mechanisms. Renewable and Sustainable Energy Reviews, 14(7), 1874–1887. http://doi.org/10.1016/j.rser.2010.03.041

Salehi Jouzani, G. & Taherzadeh, M. J. (2015). Advances in consolidated bioprocessing systems for bioethanol and butanol production from biomass: a comprehensive review. Biofuel Research Journal, 5(5), 152–195.

Santos, F. A., Queiróz, J. H. de, Colodette, J. L., Fernandes, S. A., Guimarães, V. M. & Rezende, S. T. (2012). Potencial da palha de cana-de-açúcar para produção de etanol. Química Nova, 35(5), 1004–1010. http://doi.org/10.1590/S0100-40422012000500025

Segal, L., Creely, J. J., Martin, A. E. & Conrad, C. M. (1959). An Empirical Method for Estimating the Degree of Crystallinity of Native Cellulose Using the X-Ray Diffractometer. Textile Research Journal, 29(10), 786–794. http://doi.org/10.1177/004051755902901003

Sekulic, S., Seasholtz, M. B., Wang, Z., Kowalski, B. R., Lee, S. E. & Holt, B. R. (1993). Nonlinear multivariate calibration methods in analytical chemistry. Analytical Chemistry, 65(19), 835A–845A. http://doi.org/10.1021/ac00067a001

Silva, N. L. C. (2010). Produção de bioetanol de segunda geração a partir de biomassa residual da indústria de celulose.

Silva, R., Haraguchi, S. K., Muniz, E. C. & Rubira, A. F. (2009). Aplicações de Fibras Lignocelulósicas na química de polímeros e em compósitos. Química Nova, 32(3), 661–671. http://doi.org/10.1590/S0100-40422009000300010

Sun, S., Sun, S., Cao, X. & Sun, R. (2015). The role of pretreatment in improving the enzymatic hydrolysis of lignocellulosic materials. Bioresource Technology, 199, 49–58. http://doi.org/10.1016/j.biortech.2015.08.061

Sun, Y. & Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: A review. Bioresource Technology, 83(1), 1–11. http://doi.org/10.1016/S0960-8524(01)00212-7

Teeaar, R., Serimaa, R. & Paakkarl, T. (1987). Crystallinity of cellulose, as determined by CP/MAS NMR and XRD methods. Polymer Bulletin, 17(3), 231–237. http://doi.org/10.1007/BF00285355

Teófilo, R. F. & Ferreira, M. M. C. (2006). Quimiometria II: Planilhas eletrônicas para cálculus de planejamentos experimentais, um tutorial. Quimica Nova, 29(2), 338–350. http://doi.org/10.1590/S0100-40422006000200026

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

Terinte, N., Ibbett, R. & Schuster, K. C. (2011). Overview on Native Cellulose and Microcrystalline Cellulose I Structure Studied By X-Ray Diffraction (Waxd): Comparison Between Measurement Techniques. Lenzinger Berichte, 89, 118–131. http://doi.org/10.1163/156856198X00740

Timung, R., Mohan, M., Chilukoti, B., Sasmal, S., Banerjee, T. & Goud, V. V. (2015). Optimization of dilute acid and hot water pretreatment of different lignocellulosic biomass: A comparative study. Biomass and Bioenergy, 81, 9–18. http://doi.org/10.1016/j.biombioe.2015.05.006

Tong, D. S., Xia, X., Luo, X. P., Wu, L. M., Lin, C. X., Yu, W. H., … Zhong, Z. K. (2013). Catalytic hydrolysis of cellulose to reducing sugar over acid-activated montmorillonite catalysts. Applied Clay Science, 74, 147–153. http://doi.org/10.1016/j.clay.2012.09.002

Valderrama, P., Braga, J. W. B. & Poppi, R. J. (2009). Estado da arte de figuras de mérito em calibração multivariada. Quimica Nova, 32(5), 1278–1287. http://doi.org/10.1590/S0100-40422009000500034

Williams, P. & Sobering, D. (1993). Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. Journal of Near Infrared Spectroscopy, 1(1), 25–32. http://doi.org/10.1255/jnirs.3

Wold, S., Høy, M., Martens, H., Trygg, J., Westad, F., MacGregor, J. & Wise, B. M. (2009). The PLS model space revisited. Journal of Chemometrics, 23(2), 67–68. http://doi.org/10.1002/cem.1171

Zanuncio, A. J. V., Carvalho, A. G., Trugilho, P. F. & Monteiro, T. C. (2014). Extractives and energetic properties of wood and charcoal. Revista Árvore, 38(2), 369–374. http://doi.org/10.1590/S0100-67622014000200018

CAPÍTULO 2

Carbohydrate Polymers 158 (2017) 20–28

Contents lists available at ScienceDirect

Carbohydrate Polymers

journal homepage: www.elsevier.com/locate/carbpol

Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods

Ítalo P. Caliari[a], Márcio H.P. Barbosa[b], Sukarno O. Ferreira[c], Reinaldo F. Teófilo[a,*]

[a] Departament of Chemistry, Universidade Federal de Viçosa, 36570-900 Viçosa, Minas Gerais, Brazil
[b] Departament of Plant Science, Universidade Federal de Viçosa, 36570-900 Viçosa, Minas Gerais, Brazil
[c] Departament of Physics, Universidade Federal de Viçosa, 36570-900 Viçosa, Minas Gerais, Brazil

DOI: 0.1016/j.carbpol.2016.12.005

Abstract

A method for estimation of sugarcane (*Saccharum* spp.) biomass crystallinity using near infrared spectroscopy (NIR) and partial least squares regression (PLS) as an alternative to the standard method using X-ray diffractometry (XRD) is proposed. Crystallinity was obtained using XRD from sugarcane bagasse. NIR spectra were obtained of the same material. PLS models were built using the NIR and crystallinity values. Cellulose crystallinity ranged from 50 to 81%. Two variable selection algorithms were applied to improve the predictive ability of models, i.e. (a) Ordered Predictors Selection (OPS) and (b) Genetic Algorithm. The best model, obtained with the OPS algorithm, presented values of correlation coefficient of prediction, root mean squared error of prediction and ratio of performance deviation equals to 0.92, 3.01 and 1.71, respectively. A scatter matrix among lignin, α-cellulose, hemicellulose, ash and crystallinity was built that showed that there was no correlation among these properties for the samples studied.

Keywords: Crystallinity, Sugarcane, PLS, NIR, XRD, OPS

1. Introduction

Cellulose is considered the most abundant biopolymer on our planet, making up about 50% of the dry weight of plants (Chen, Zhang, Xu & Cao, 2015; Gutiérrez-Rojas, Moreno-Sarmiento & Montoya, 2015; Silva, Haraguchi, Muniz & Rubira, 2009;

Timhadjelt, Serier, Belgacem & Bras, 2015). Thus, by being present in different sources of biomass, cellulose is the subject of several studies as raw material in the production of biofuels and chemicals of industrial interest. This relevance is justified by the need to reduce oil use and emissions of greenhouse gases (Liu and Han, 2015, Mosier et al., 2005, Sun and Cheng, 2002 and Timhadjelt et al., 2015; Yang, Yan, Chen, Lee & Zheng, 2007).

Cellulose is composed of D-glucose monomers joined by β-1,4 glycosidic linkages which form polymer chains held together by van der Waals force and intermolecular hydrogen bonding between hydroxyl groups present at positions C-2, C-3 and C-6 (Mohanty, Misra & Hinrichsen, 2000; Peng et al., 2013 and Tong et al., 2013). Due to this arrangement, there will be areas of high ordering, called crystalline regions. There are also areas with low or no ordering, also said to be amorphous regions (Gutiérrez-Rojas et al., 2015; Klemm, Heublein, Fink & Bohn, 2005; Peng et al., 2013). Among its many allomorphs, I_α (triclinic) and I_β (monoclinic) cellulose forms are the most abundant, being found concomitantly in lignocellulosic materials, differing in proportion according to biomass source (Klemm et al., 2005; Park, Baker, Himmel, Parilla & Johnson, 2010). Nevertheless, recent research has reported that the native state of wood cellulose may not be crystalline (Agarwal, Ralph, Reiner & Baez, 2016). Under the model for cellulose proposed by Agarwal et al. (2016), although the cellulose chains are highly organized ~50% of them can be accessed by water.

Crystallinity index, which is defined as the fraction of crystalline matter within a sample, is related to various physical properties of materials, such as Young's modulus, dimensional stability, density, hardness, chemical reactivity, among others (Jiang, Yang, So & Hse, 2007; Terinte, Ibbett & Schuster, 2011). Consequently, once hydrolysis reactions occur preferentially in amorphous regions, due to recalcitrance of crystalline form, the knowledge of such index provide an estimation of energy output involved in these steps and thus the viability of the process itself (Hendriks and Zeeman, 2009, Rambo and Ferreira, 2015 and Sun and Cheng, 2002).

The standard method for determining the crystallinity index is through X-ray diffractometry (XRD), using the powder method (Jiang et al., 2007, Peng et al., 2013, Timhadjelt et al., 2015 and Zidan et al., 2012). Generally, despite the good results, this method slows the analysis as diffractogram acquisition is a relatively lengthy process and reliant on specific sample preparation, besides the high instrument and maintenance costs. Moreover, its field application is virtually impossible. In literature, there are several studies on near infrared spectroscopy (NIR) associated with multivariate regression for

crystallinity index prediction of various biomass sources, drugs and related materials. This technique offers information on hydrogen bonds, which is decisive in crystalline structure formation (Jiang et al., 2007; Monrroy, Garcia, Troncoso & Freer, 2015; Nørgaard, Hahn, Knudsen, Farhat & Engelsen, 2005; Zidan et al., 2012). The NIR spectroscopy is advantageous due to rapidity, being performed in few seconds with minimum sample manipulation and reduced instrument cost, which are essentially important factors in routine analysis. As another great benefit, NIR analyses require no sample preparation or manipulation with hazardous chemicals, solvents, or reagents (Workman & Weyer, 2008). Besides, the analysis can be performed in situ, which would facilitate plant selection in bio-refineries, such as sugarcane breeding programs for energy production. However, NIR absorption spectra are often complex and normally possess broad overlapping NIR absorption bands that require special procedures for data analysis. Therefore, quantification is done by using multivariate calibration (chemometrics) (Workman & Weyer, 2008).

Multivariate calibration provides local models and, therefore, for each type of biomass it is necessary to build a model. This is required because each type of biomass has specific compounds, which need to be modeled in order to ensure reliable predictions. Although Jiang et al. (2007) have published a method for prediction of wood crystallinity in *Pinus elliottii* by NIR spectroscopy, such a model is specific to this type of biomass and cannot be considered for sugarcane biomass. In addition, sample preparation for obtaining the spectra is also very important for this kind of method.

Brazilian sugarcane production is broadly intended for sugar and ethanol industry, being considered one of the largest worldwide. The estimated production of sugarcane for the 2015/16 season increased by 4.9% compared to the previous harvest, reaching 665.6 million tons (CONAB, 2015/16). However, a large part of the biomass produced is not properly utilized, becoming industrial waste (Silva et al., 2009). Thus, seeking feasible alternatives for its application represents both economic and environmental progress. However, the control of properties such as crystallinity index through genetic improvement research would allow greater economic viability of hydrolysis processes, thus providing greater competitiveness to fuel production from that biomass source.

To the best of our knowledge, among the studies on crystallinity prediction using NIR and multivariate regression methods, none was found involving the method for sugarcane biomass analysis. Therefore, the objective of this study was to build a model for the crystallinity prediction of sugarcane (*Saccharum* spp.) biomass, using NIR

spectroscopy associated with methods of multivariate analysis based on partial least squares (PLS) regression and the methods of variables selection such as the ordered predictors selection (OPS) and the genetic algorithm (GA).

2. Material and methods

2.1. Preparation of biomass samples

Two hundred and twenty-one experimental sugarcane (*Saccharum* spp.) genotypes were supplied by the germplasm bank of the *Universidade Federal de Viçosa*, Viçosa, Minas Gerais, Brazil. The plantation was set in May 2014 using rows 5 m in length in an experimental field in Viçosa, MG, Brazil (20°44′37″ latitude south, 42°50′38″ longitude west).

For the preparation of each biomass sample, the stalks were crushed, pressed for juice extraction, dried in an oven at 65 °C for 72 h, ground and sieved between 20 and 80 mesh. The fraction retained on the 80 mesh sieve (−20/+80 mesh fraction) was retained for crystallinity and compositional analysis and stored at room temperature in sealed containers (Hames et al., 2008). The material retained in the solid catch pan (−80 mesh) was separated for ash analysis.

2.2. Compositional analysis

The oven-dried samples were incinerated at 580 °C for 24 h to determine ash content on the non-extracted samples (Sluiter et al., 2008a). The extractive samples (2 g dry weight) were extracted successively in water and ethanol in a Soxhlet extraction unit for a period of 10 and 5 h, respectively. For these samples, Klason lignin content was determined gravimetrically by using 72% sulfuric acid using the procedure established by the United States National Renewable Energy Laboratory (NREL) (Sluiter et al., 2008b).

α-cellulose and hemicellulose were determined gravimetrically according to Zhou, Jiang, Via, Fasina, and Han (2015). Holocellulose was quantified with sodium chlorite treatment and α-cellulose content was determined by extraction with 17.5% aqueous sodium hydroxide of the holocellulose powder. Hemicellulose content was evaluated by difference between holocellulose and α-cellulose contents.

2.3. XRD analysis

Approximately 3.0 g of biomass sample was mixed with analytical grade isopropanol. This mixture was pressed against a cylindrical support of polymethyl methacrylate (PMMA) in disc form with 2 mm depth and 25 mm diameter. Isopropanol ensured biomass fixation onto the support. X-ray diffraction measurement was made after a minimum period of 24 h at room temperature, for complete solvent evaporation. An X-ray diffractometer D8 Discover by Bruker equipped with Cu radiation source with wavelength 1.5418 Å (voltage of 40 kV and current of 40 mA) and 0.2 mm slits was used. The θ-2θ scans were performed in the range of 10–40° with 0.1° steps and 2 s per step. The time for obtaining the diffractogram was approximately 20 min per sample.

2.4. Determination of the crystallinity index

After baseline correction, crystallinity values were determined according to the method of Segal, Creely, Martin, and Conrad (1959), according to Eq. (1).

$$CrI(\%) = \frac{(I_{200} - I_{am})}{I_{200}} \times 100 \qquad \text{Equation (1)}$$

where $CrI(\%)$ is the crystallinity index calculated as a percentage, I_{200} is the diffraction intensity associated with crystalline cellulose (maximum diffraction between $20° < 2\theta < 25°$), and I_{am} is the intensity associated to amorphous cellulose (minimum diffraction between $15° < 2\theta < 20°$). Crystallinity values were defined as dependent variables for the building of the multivariate regression model.

2.5. NIR spectroscopic analysis

The biomass samples were also analyzed using the FT-NIR Varian spectrometer Model 660 IR, equipped with a diffuse reflectance accessory. Readings were taken in the wavenumber range of 4000–10000 cm^{-1}. The spectra were obtained with a resolution of 4 cm^{-1}, 32 scans per analysis and using a gold reference for background. In each spectrum, 1038 wavenumbers were collected. The signal was given in log(1/Reflectance). The sample spectra were obtained in duplicate and the mean spectra was used for data analysis. These data are the independent variables.

2.6. Data analysis

The NIR spectra were used for building a spectra matrix $\mathbf{X}_{(m,n)}$ with *m* samples and *n* observations, and the crystallinity values were arranged in a crystallinities vector $\mathbf{y}_{(m,1)}$. The **X** matrix was transformed by applying the first derivative with the Savitzky-Golay algorithm with 5-point window (Savitzky & Golay, 1964). Both **X** matrix and **y** vector were mean centered. The regression was performed with the PLS bidiagonal algorithm written in the laboratory. The OPS algorithm was applied using OPS_Toolbox package, available at www.deq.ufv.br/chemometrics. The genetic algorithm was applied using the package PLS_Toolbox 6.7 (Eigenvector INC), all calculations were performed using the Matlab R2016a software.

2.7. Figures of merit

The quality of the models was evaluated by the parameters root mean square error (RMSE) and correlation coefficient (R) which were calculated by Eqs. (2) and (3), respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{I} (y_i - \hat{y}_i)^2}{I}} \qquad \text{Equation (2)}$$

$$R = \frac{\sum_{i=1}^{I} (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{I} (\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{I} (y_i - \bar{y})^2}} \qquad \text{Equation (3)}$$

where $\hat{y}$ and $\bar{\hat{y}}$ being the estimated value and mean of estimated values, respectively, y and $\bar{y}$ the values and mean values obtained, respectively. In the case where the internal cross-validation (*CV*) is used, *I* represents the number of samples in the cross-validation set, and the error and correlation coefficient are called root mean square error of cross validation (*RMSECV*) and correlation coefficient of cross-validation (*RCV*), respectively. For the external validation, *I* represents the number of prediction samples (*P*) and, in this case, the error and correlation coefficients are named correlation root mean square error of prediction (*RMSEP*) and correlation coefficients of prediction (*RP*), respectively. The Kennard-Stone algorithm was used to select samples of the calibration and prediction sets (Kennard & Stone, 1996).

The model was also assessed in terms of Ratio of Performance Deviation (*RPD*), described in Eq. (4). This is the ratio of the standard deviation (*SD*) of the values of the property analyzed and *RMSECV*. This statistic provides a basis for standardization of the standard prediction error (Williams & Sobering, 1993).

$$RPD = \frac{SD}{RMSECV} \qquad \text{Equation (4)}$$

The sensitivity (*SEN*), selectivity (*SEL*), analytical sensitivity (γ) and limit of detection (*LOD*) were calculated according to Eqs. (5) through 8, respectively.

$$SEN = \frac{1}{\| \mathbf{b} \|} \qquad \text{Equation (5)}$$

where ||**b**|| is the Euclidean norm of the vector of regression coefficients of the PLS model (Valderrama, Braga & Poppi, 2009).

$$SEL = \frac{nas_{k,i}}{\| \mathbf{x}_{k,i} \|} \qquad \text{Equation (6)}$$

where nas_i is the scalar value of the net analytical signal for sample *i* and $\| \mathbf{x}_i \|$ represents the Euclidean norm of the instrumental response vector for sample *i* (Rambo, Amorim & Ferreira, 2013; Valderrama et al., 2009).

$$\gamma = \frac{SEN}{\| \partial_x \|} \qquad \text{Equation (7)}$$

where *SEN* is the sensitivity and $\|\partial_x\|$ is the Euclidean norm of the reference signal standard deviation (Valderrama et al., 2009).

$$LOD = \frac{\Delta(\alpha, \beta) w_{y_0} \sigma}{\hat{a}} \qquad \text{Equation (8)}$$

where σ is the standard deviation of the regression residue, w_{y0} is a function of the standards position and $\hat{a}$ is the linear regression coefficient of the Ortiz method (Ortiz et al., 2003).

3. Results and discussion

3.1. XRD results

The maximum diffraction can be seen in the region of $20° < 2\theta < 25°$ and the minimum between $15° < 2\theta < 20°$; these results are consistent with those obtained in literature for cellulose in lignocellulosic biomass (Jiang et al., 2007, Rambo and Ferreira, 2015 and Terinte et al., 2011). Fig. 1A shows the crystallinity distribution in the 221 analyzed samples and Fig. 1B shows the typical X-ray diffractograms of sugarcane biomass from samples with different crystallinities. The crystallinity values found for the analyzed samples ranged between 50% and 81%. These results are according to what is expected when compared to experiments performed for other sources of lignocellulosic biomass (Jiang et al., 2007, Park et al., 2010, Rambo and Ferreira, 2015 and Terinte et al., 2011). According to Jiang et al. (2007), cellulose crystallinity values for biomass of *Pinus elliottii* ranged between 35% and 61%. Yet Rambo and Ferreira (2015) found crystallinity values ranging from 37% to 56% for banana biomass.

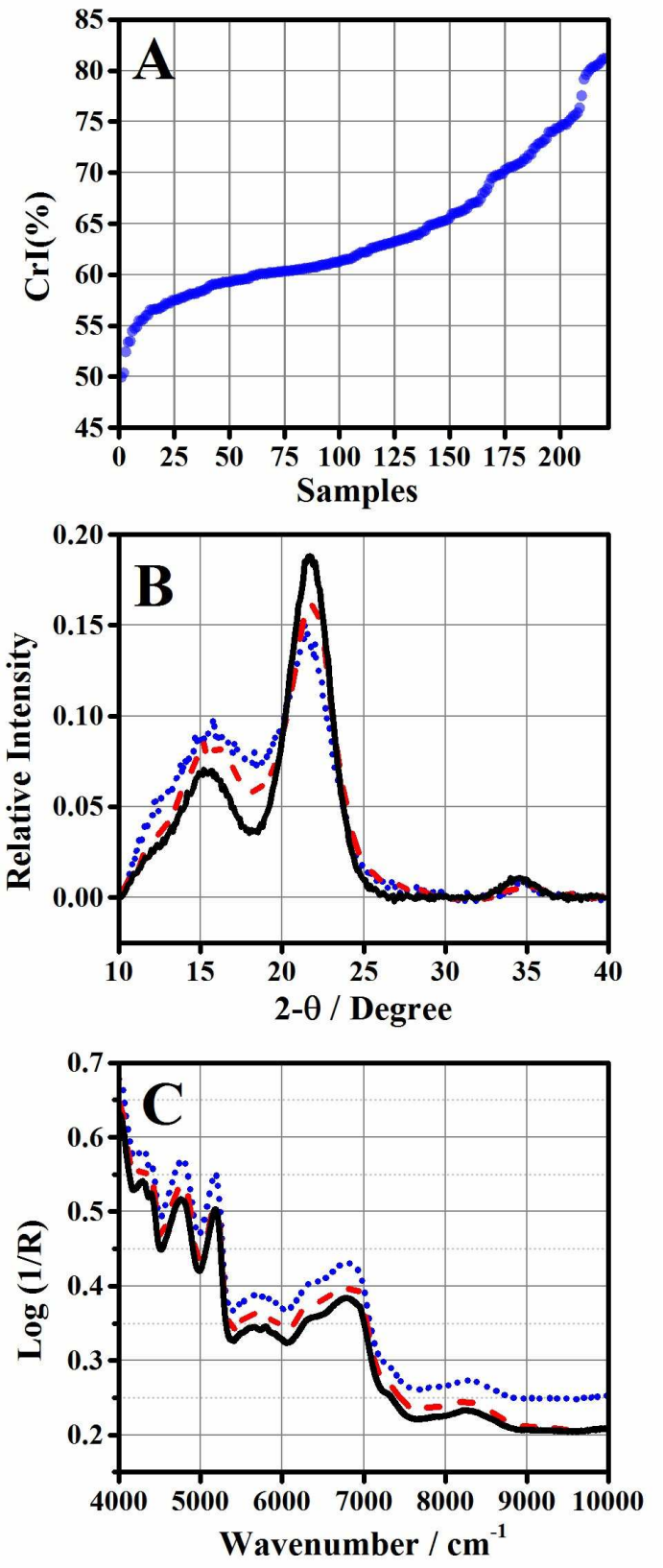

Fig. 1. (A) Cellulose crystallinity distribution in sugarcane samples, (B) X-ray diffractograms obtained from samples with different crystallinities, (C) NIR spectra obtained from samples with different crystallinities. Key: (•••) *CrI*: 50%; (---) *CrI*: 65%; (—) *CrI*: 80%.

3.2. NIR spectroscopy results

The NIR spectra of the biomass samples with minimum, mean and maximum crystallinity values are shown in Fig. 1C, wherein the main bands are located between 4000 and 5500 cm^{-1} and 6000–7100 cm^{-1}.

From the spectra, the following bands can be attributed: stretching of first overtone of cellulose polymeric OH at 6897 cm^{-1}, combination of stretching of second overtone of cellulose O-H and C-O at 5495 cm^{-1}, stretching and bending of O-H in polysaccharides at 5208 cm^{-1}, O-H stretching and cellulose C-H deformation at 4762 cm^{-1}, combination of C-H stretching and CH_2 deformation at 4283–4286 $cm^{-1}$and combination of C-H and C-C stretching at 4010 cm^{-1} (Rambo and Ferreira, 2015 and Workman and Weyer, 2008). Similar spectra were obtained by Morgano, Faria, Ferrão, Bragagnolo, and Ferreira (2005) studying coffee, by Rambo and Ferreira (2015) in banana biomass and Jiang et al. (2007) studying a pine species.

3.3. Obtaining the multivariate calibration models

3.3.1 Choosing the optimal spectra treatment

To choose the treatment that provided the model with best fit, several transformations were applied and then the data were mean centered. It was investigated the use of the raw data, the first (D1) and second (D2) derivatives, the multiplicative signal correction (MSC), the standard normal variance (SNV) and the baseline correction (Wise, Gallagher, Shaver, Windig & Koch, 2006). For D1 and D2, a preliminary study was carried on the size of the optimal window to be used in the Savitzky-Golay smoothing algorithm. The statistical parameters *RMSECV*, *RCV* and *RPD* were compared in each case (Fig. 2), thus, an optimal choice would be to use the first derivative through the Savitzky-Golay algorithm with window 5. Thus, the optimum number of latent variables for the model (*h*Mod) in this conditions was 5.

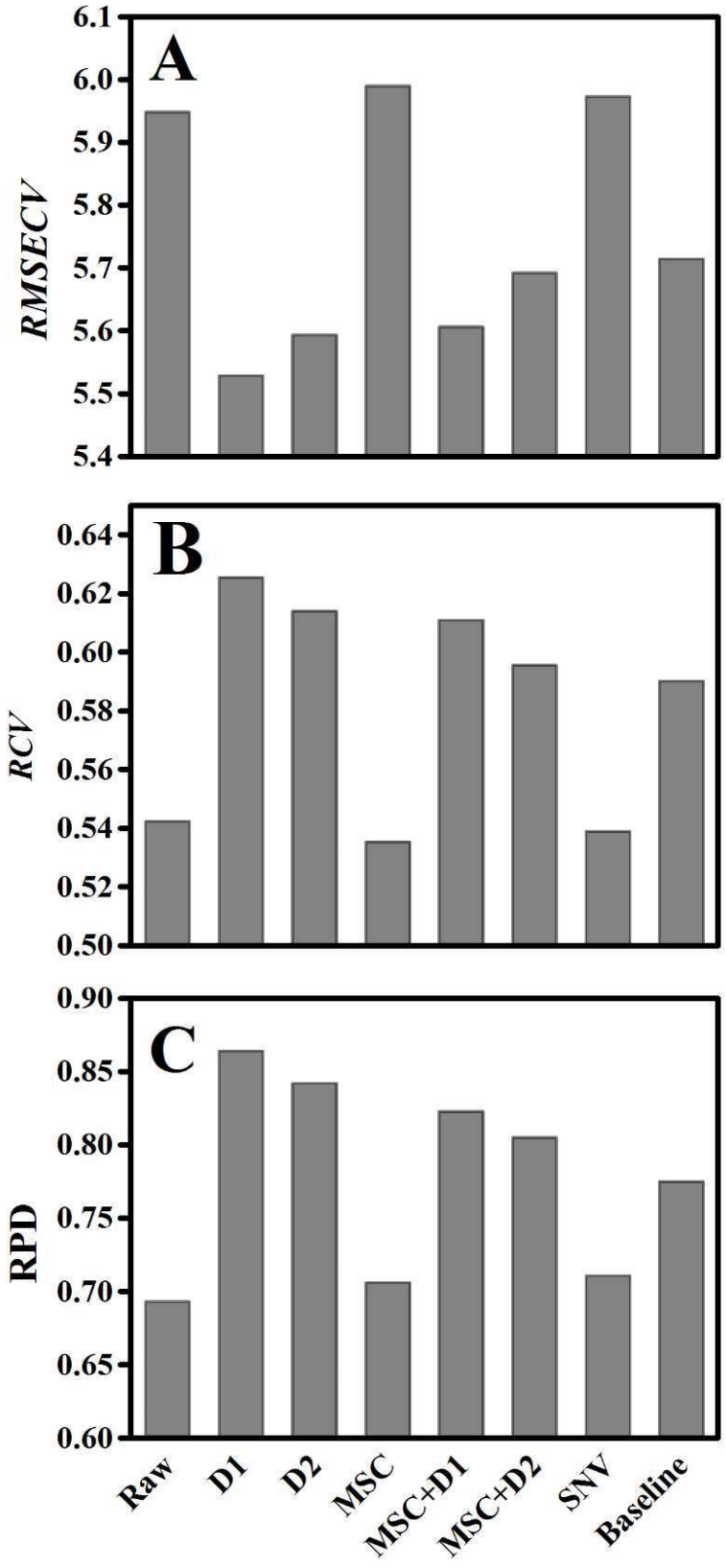

Fig. 2. Values for the models built with different treatments (A) *RMSECV*, (B) *RCV* and (C) *RPD*. Key: Raw data, first (D1) and second (D2) derivatives, multiplicative signal correction (MSC), standard normal variance (SNV) and baseline correction. All data were mean centered.

3.3.2. Variable selection with OPS algorithm

After investigating the number of latent variables (*h*) to build the optimal regression vector to be used in OPS as informative vector, was found a number *h* equal to 9, named *h*OPS (Teófilo, Martins & Ferreira, 2009). This one was chosen once the *h*OPS equal to 9 provided the lowest *RMSECV* value and highest *RCV* value, as shown in Fig. 3A and B, respectively.

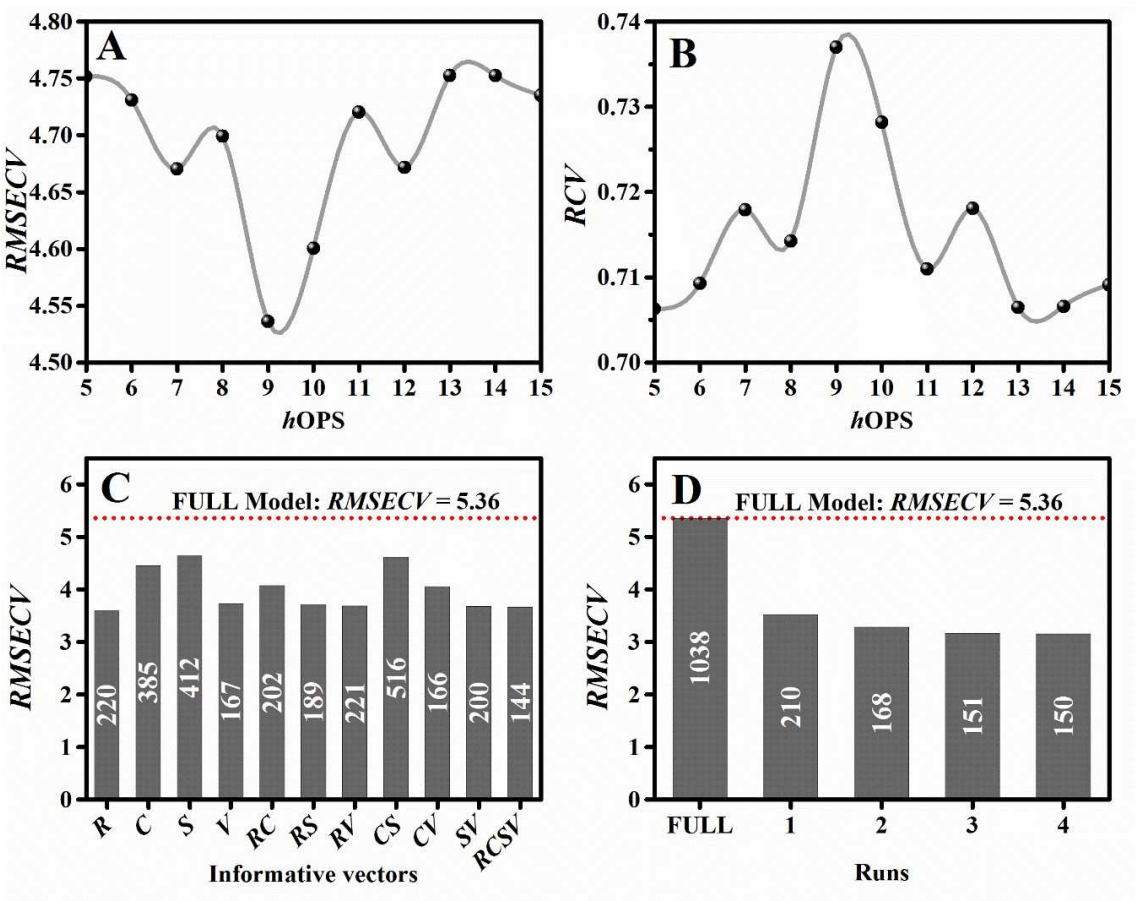

Fig. 3. (A) *RMSECV* values for different *h*OPS values; (B) *RCV* values for different *h*OPS values; (C) *RMSECV* values obtained using different informative vectors and (D) *RMSECV* values for the OPS sequential runs.

The results of the study of informative vectors such as (1) regression vectors (R) built with *h*OPS equal to 9; (2) correlation between the columns of **X** and **y** (C); (3) Residues (S) and (4) NAS vector (V) as well as their combinations (Teófilo et al., 2009) are shown in Fig. 3C. The performance of each vector was compared in terms of *RMSECV* (Fig. 3C). From the results, it is possible to identify the Regression vector as the informative vector which provided best results, once it obtained the lowest *RMSECV*

value, being the chosen vector. Using these conditions, OPS algorithm was applied several times on the selected variables until the *RMSECV* had no significant decreased values, as shown in Fig. 3D. Inside each column in Fig. 3C and D is presented the numbers of variables selected to be used in the models.

Results presented in Fig. 3 indicate a significant improvement in the model with the variables selected, showing *RMSECV* values of 5.36–3.15 when the OPS algorithm was applied. Furthermore, the number of variables used was reduced from 1038 to only 150.

3.3.3. Variables selection with GA algorithm

The GA input parameters were optimized using the Plackett-Burman experimental design (Plackett & Burman, 1946) in order to get the best prediction result possible. The response used to decide on the best parameters were the *RMSECV* and *RCV*. The parameters and their respective optimal values were: population size 52; generations 150; mutation rate 0.001; window size 1; convergence 70; initial terms 20; crossover 2; splits 3; replications 3 and interactions 1.

3.3.4. Comparison of results

Results shown in Fig. 4 indicate that there was better data distribution around the line of 45°, and a decrease of relative errors and a narrowing of the distribution curve of these values, in the case where model was built with variables selected by the OPS algorithm.

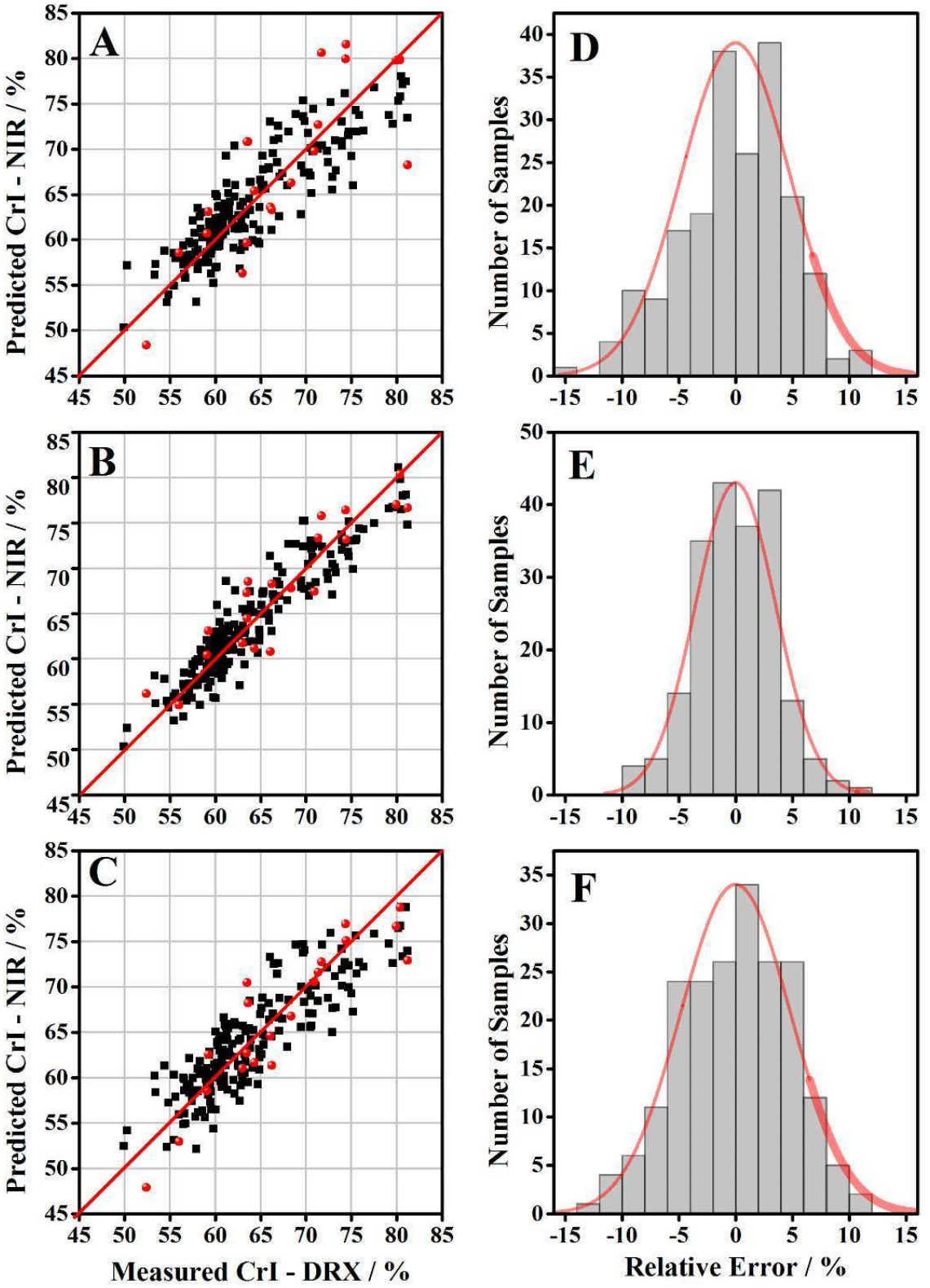

Fig. 4. Relation between *CrI* values measured and predicted for the models: (A) with all variables; (B) OPS; (C) GA (■ represents the calibration sample set and ● the prediction sample set); Histogram with model Relative Errors: (D) with all variables; (E) OPS; (F) GA.

A comparison between the crystallinity values predicted by the models for the prediction sample set is shown in Table 1. Note that, in most cases, there was a higher approximation to the reference values by the predictions made with the variables selected, such that OPS algorithm presented slightly better results than GA.

Table 1. Measured, predicted and relative errors (RE) values for the external set samples.

	Regression Models					
	FULL		OPS		GA	
Measured[a]	Predicted[a]	RE[a]	Predicted[a]	RE[a]	Predicted[a]	RE[a]
62.97	56.34	10.53	61.77	1.91	61.04	3.07
74.43	81.55	−9.56	73.15	1.72	75.12	−0.93
63.48	59.64	6.05	67.24	−5.93	62.72	1.19
66.08	63.60	3.75	60.81	7.98	64.60	2.24
71.35	72.69	−1.88	73.33	−2.78	71.63	−0.40
59.11	60.68	−2.66	60.44	−2.25	58.54	0.96
74.33	79.97	−7.59	76.40	−2.78	76.92	−3.48
81.18	68.24	15.94	76.64	5.59	72.91	10.18
80.34	79.87	0.58	80.37	−0.04	78.77	1.96
52.43	48.40	7.69	56.15	−7.10	47.91	8.61
66.25	63.32	4.43	68.28	−3.06	61.34	7.41
79.93	79.79	0.18	77.00	3.67	76.67	4.08
71.72	80.62	−12.41	75.75	−5.61	72.76	−1.45
63.64	70.83	−11.30	68.49	−7.62	68.21	−7.18
56.02	58.54	−4.50	54.94	1.93	52.95	5.47
68.37	66.27	3.06	67.81	0.81	66.77	2.34
64.30	65.38	−1.68	61.18	4.86	61.68	4.07
63.50	70.81	−11.51	64.43	−1.47	70.49	−11.01
70.94	69.76	1.66	67.41	4.97	70.56	0.53
59.22	63.11	−6.57	63.09	−6.54	62.57	−5.65

RE: Relative error.

[a] Values expressed as percentage.

From the data shown in Table 2, it is possible to verify that the variable selection methods provided better results for most parameters evaluated. The model built with the selection made by the OPS algorithm clearly showed better results for *RMSECV*, *RCV*, *RMSEP*, *RP*, *RPD*, γ^{-1} and *LOD*.

Table 2. Statistics obtained for the built models.

	Regression Models		
	FULL	OPS	GA
Vector	–	Reg	–
*h*OPS	–	9	–
*h*Mod	5	5	5
*n*Vars	1038	150	201
RMSECV	5.36	3.31	4.46
RCV	0.58	0.86	0.73
RMSEP	5.28	3.01	3.47
RP	0.81	0.92	0.91
RPD	0.86	1.71	1.21
SEN	3.77×10^{-5}	1.77×10^{-5}	2.34×10^{-5}
SEL	0.07	0.09	0.10
γ	3.90	4.07	1.83
γ^{-1}	0.26	0.25	0.55
LOD	10.65	6.65	11.95

R: Regression vector, used as an informative vector in the OPS algorithm; *h*OPS and *h*Mod. Number of latent variables for the OPS and for the model, respectively; *n*Vars: Number of independent variables used in the model building; *RMSECV*: Root mean square error of cross-validation; *RMSEP*: Root mean square error of prediction, *RCV*: correlation coefficient of cross-validation; *RP*: correlation coefficient of prediction; *RPD*: Ratio of performance deviation. *SEN*: Sensitivity. *SEL*: Selectivity. γ: Analytical sensitivity. γ^{-1}: Inverse of analytical sensitivity. *LOD*: limit of detection.

For the OPS model, the *RCV* and *RP* values were higher than 0.85 and considered excellent for the aim of this work. Furthermore, once the OPS model presented the highest *RPD* value in relation to full and GA models, it is possible to say that OPS showed better model fit and quality predictions.

The low *SEN* values obtained for all the models were expected and can be assigned to the derivative treatment applied to the NIR spectra. Similarly, low *SEL* values are also in accordance with the expected, since the lignocellulosic biomass has many interfering compounds that have no relation to the crystallinity.

The γ^{-1} value for the OPS model indicates that it is possible to identify a variation of 0.25 units in the crystallinity index from the NIR spectra. Finally, another quality of the built model is in the fact that *LOD* is significantly below in comparison to the lowest experimental value of cellulose crystallinity found for sugarcane. This indicates that the

model is able to detect crystallinity sugarcane cellulosic biomass even if there is essentially amorphous cellulose.

Jiang et al. (2007) and Rambo and Ferreira (2015) presented models using NIR spectroscopy and PLS to predict crystallinity using Segal method in *Pinus elliottii* and banana residues, respectively. In a comparison among the OPS model and the models obtained by Jiang et al. (2007) and Rambo and Ferreira (2015) it can be observed that, the number of latent variables were respectively 5, 5 and 6. These values indicate that the complexity of the information analyzed is similar. Essentially, the results in Table 2 are very close to those presented by Jiang et al. (2007) and Rambo and Ferreira (2015) in terms of *RMSECV*, *RCV*, *RMSEP* and *RP*, even though they were built from different biomasses. Analyzing the *SEL*, *SEN* and γ^{-1} values calculated by Rambo and Ferreira (2015), again the values are close to those presented in this work.

The Segal method is useful for comparing the relative differences among samples (Park et al., 2010), in the same way, as the method presented in this work is not absolute and use the Segal method as reference, it will also be useful for relative comparisons. Furthermore, Park et al. (2010) suggested the overestimation of the crystallinity measures attributed to the Segal method due to (1) minimum position between the 002 and the 101 peaks; (2) neglected peaks in the calculations; and (3) neglected variation in peak width. Therefore, despite the NIR spectrum is not sensitive to the effects related to the preferred orientation of the sample, all uncertainties in the prediction attributed to the Segal method, were included in the built model using NIR.

The variables selected, as shown in Fig. 5, reveals that both algorithms selected wavenumbers well distributed around almost the whole spectral region studied. Besides, a greater selectivity was achieved with the OPS algorithm relating to the number of relevant variables selected between 5000 to approximately 6500 cm^{-1}, when compared to the other wavenumbers analyzed, (Fig. 5A). This can be explained since the NIR signal in the region near 5868 cm^{-1} is typical of cellulose amorphous contribution (Rambo & Ferreira, 2015) and probably was not relevant in the building of models.

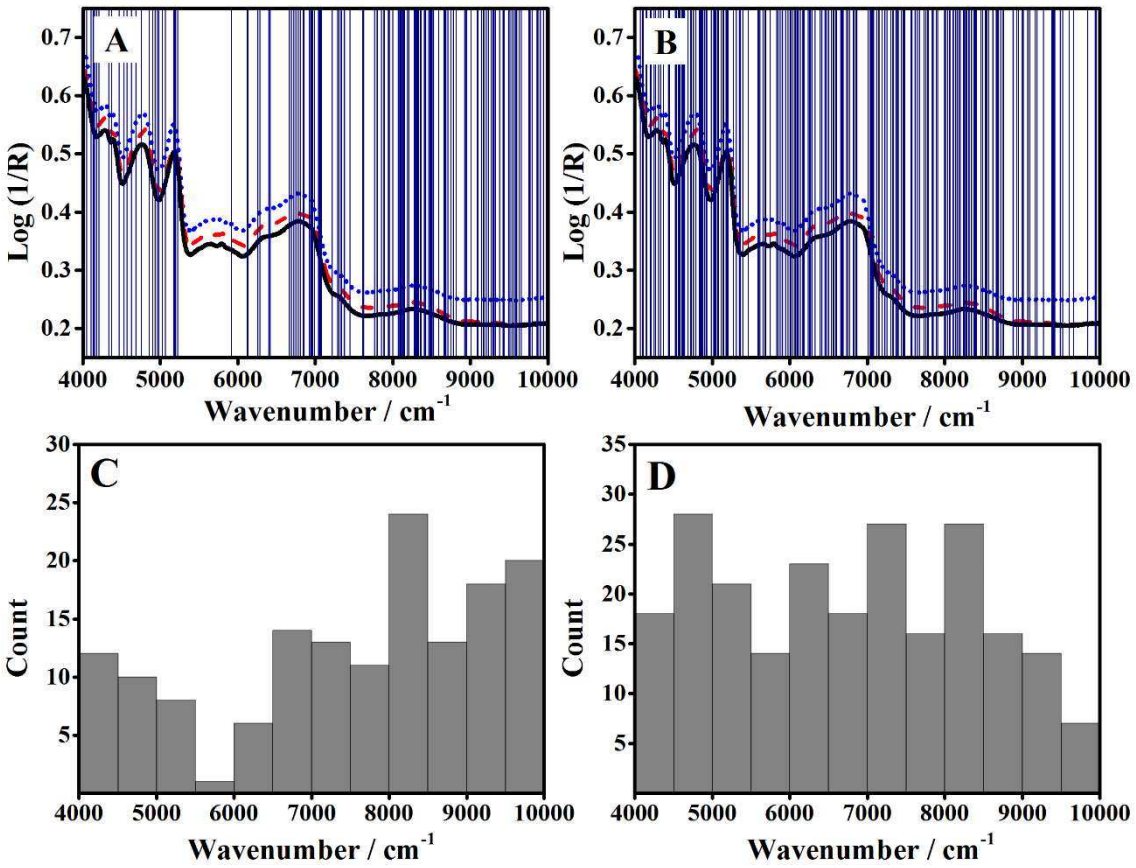

Fig. 5. Vertical lines represent the variables selected by the algorithms: (A) OPS and (B) GA. Histogram with the counts of selected variables in different spectral ranges for (C) OPS and (D) GA.

On the other hand, the GA algorithm was less specific in the choice of variables, selecting wavenumbers throughout the spectrum, leading to a small decrease in selections only between the short range of 9500 and 10000 cm^{-1} (Fig. 5B). The region that was the most commonly selected between both algorithms was in the range of 8000–8500 cm^{-1}.

3.3.5. Correlations between chemical composition and crystallinity

Information on the behavior of the chemical and physical components of biomass are important to help understanding the formation of the components and the property of the material studied. The Fig. 6 shows the correlations among lignin, α-celullose, hemicellulose, ash and crystallinity of the sugarcane biomass. The results show that there was no correlation between the chemical components and crystallinity for the studied samples.

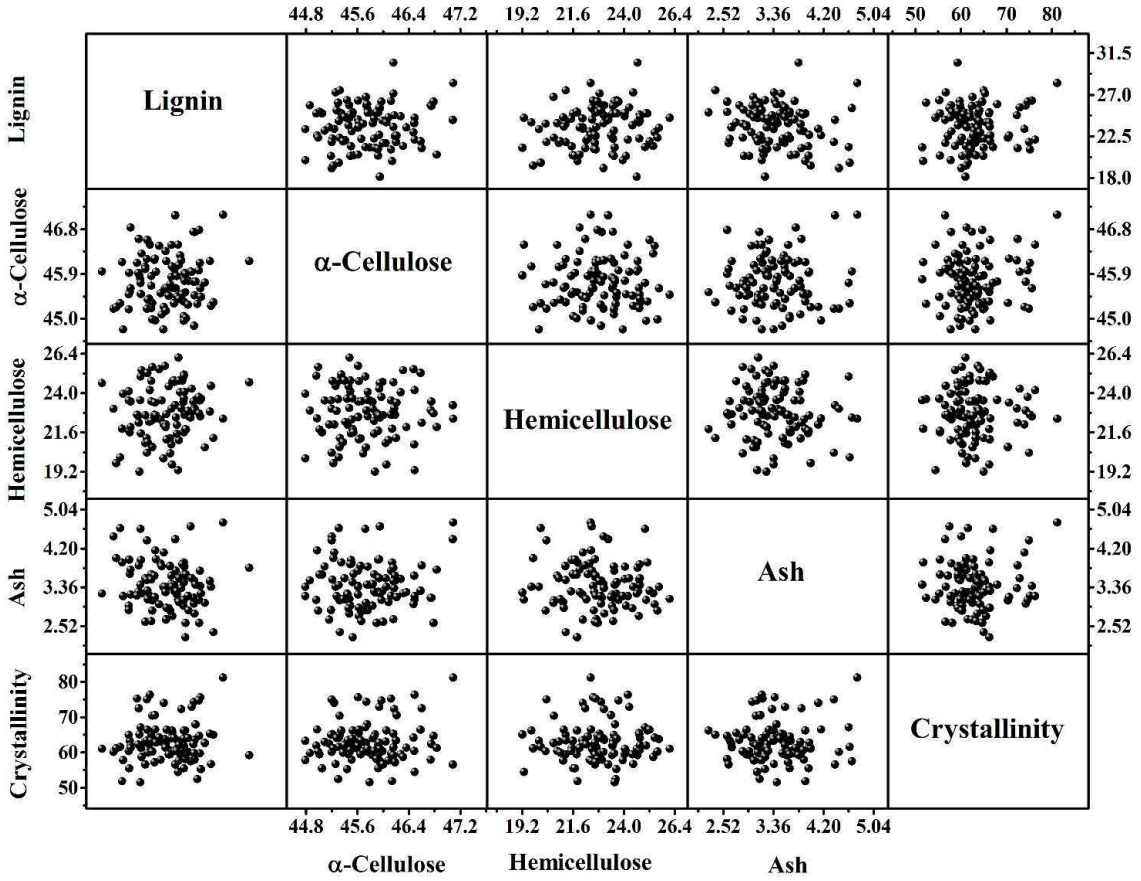

Fig. 6. Scatter matrix of correlation between chemical composition and crystallinity.

Andersson, Serimaa, Paakkari, Saranpää, and Pesonen (2003), studying the relation between crystallinity of wood and the size of cellulose crystallites in Norway spruce (*Picea abies*), showed that the mass fraction of crystalline cellulose in the cell wall changes with the ring number and distance from the pith. However, in a new study of crystallinity of Scots pine and Norway spruce cellulose, Andersson, Wikberg, Pesonen, Maunu, and Serimaa (2004) observed that the crystallinity of cellulose and the thickness of cellulose crystallites were the same from the pith to the bark for both Norway spruce and Scots pine. Thus, they concluded that the degree of order of cellulose microfibrils do not vary with the distance from the pith even thought their orientation varies.

Although no studies of Andersson et al., 2003 and Andersson et al., 2004 have explored the correlation presented in this work, the results indicate that crystallinity is an independent property for the samples studied. These results justify the need to estimate the crystallinity of wood and biomass, especially when the interest is the production of chemicals and biofuels.

4. Conclusion

The results highlighted the feasibility of crystallinity index prediction in sugarcane biomass though near infrared spectroscopy and multivariate analysis methods. The models were successfully built using a wide range of crystallinity between 50 and 81% according to Segal method. The best treatment of spectra was the first derivative followed by mean centering. Among the built models, the one obtained by OPS algorithm showed an excellent prediction ability, besides, the OPS was carried out in a short time, unlike the GA. Thus, the OPS algorithm proved to be very efficient, overcoming GA, one of the variable selection methods most widespread in the literature. The advantage of this kind of prediction is the timesaving and convenience of obtaining a NIR spectrum in relation to an X-ray diffractogram, since NIR spectroscopy measurements can be performed on-line in an easier way. Thus, the proposed method is suitable for reliable routine analyses when faster results are needed. Finally, the scatter matrix plot between chemical composition and crystallinity shows that there was no correlation among lignin, α-cellulose, hemicellulose, ash and crystallinity for the samples studied.

5. Acknowledgements

The authors gratefully acknowledge the financial support of Brazilian research funding agencies CNPq, CAPES and FAPEMIG and the SisNano-UFV X-ray laboratory for support with the XRD measurements. This work is a collaboration research project of members of the Rede Mineira de Química (RQ-MG) supported by the FAPEMIG [grant numbers REDE-113/10 and CEX-RED-00010-14] and also supported by the FAPEMIG [grant number CEX-APQ-01424-13].

6. References

Agarwal, U. P., Ralph, S. A., Reiner, R. S. & Baez, C. (2016). Probing crystallinity of never-dried wood cellulose with Raman spectroscopy. Cellulose, 23(1), 125–144. http://dx.doi.org/10.1007/s10570-015-0788-7

Andersson, S., Serimaa, R., Paakkari, T., Saranpää, P. & Pesonen, E. (2003). Crystallinity of wood and the size of cellulose crystallites in Norway spruce (Picea abies). Journal of Wood Science, 49(6), 531–537. http://dx.doi.org/10. 1007/s10086-003-0518-x

Andersson, S., Wikberg, H., Pesonen, E., Maunu, S. L. & Serimaa, R. (2004). Studies of crystallinity of Scots pine and Norway spruce cellulose. Trees, 18(3), 346–353. http://dx.doi.org/10.1007/s00468-003-0312-9

CONAB—Companhia Nacional de Abastecimento. (2016). Acompanhamento da safra brasileira de cana-de-ac¸ úcar. Safra 2015/16—Quarto levantamento. Retrieved from. http://www.conab.gov.br/OlalaCMS/uploads/arquivos/16 04 14 09 06 31 boletim cana portugues - 4o lev - 15-16.pdf

Chen, Y., Zhang, Y., Xu, C. & Cao, X. (2015). Cellulose nanocrystals reinforced foamed nitrile rubber nanocomposites. Carbohydrate Polymers, 130, 149–154. http://dx.doi.org/10.1016/j.carbpol.2015.05.017

Gutiérrez-Rojas, I., Moreno-Sarmiento, N. & Montoya, D. (2015). Mecanismos y regulación de la hidrólisis enzimática de celulosa en hongos filamentosos: casos clásicos y nuevos modelos. Revista Iberoamericana de Micología, 32(1), 1–12. http://dx.doi.org/10.1016/j.riam.2013.10.009

Hames, B., Ruiz, R., Scarlata, C., Sluiter, A., Sluiter, J. & Templeton, D. (2008). Preparation of samples for compositional analysis: Laboratory analytical procedure (LAP). Golden, CO, USA: National Renewable Energy Laboratory. Technical report NREL/TP-510-42620.

Hendriks, A. T. W. M. & Zeeman, G. (2009). Pretreatments to enhance the digestibility of lignocellulosic biomass. Bioresource Technology, 100(1), 10–18. http://dx.doi.org/10.1016/j.biortech.2008.05.027

Jiang, Z. H., Yang, Z., So, C. L. & Hse, C. Y. (2007). Rapid prediction of wood crystallinity in Pinus elliotii plantation wood by near-infrared spectroscopy. Journal of Wood Science, 53(5), 449–453. http://dx.doi.org/10.1007/s10086- 007-0883-y

Kennard, R. W. & Stone, L. A. (1996). Computer aided design of experiments. Technometrics, 11(1), 137–148. http://dx.doi.org/10.1080/00401706.1969. 10490666

Klemm, D., Heublein, B., Fink, H. P. & Bohn, A. (2005). Cellulose: Fascinating biopolymer and sustainable raw material. Angewandte Chemie— International Edition, 44(22), 3358–3393. http://dx.doi.org/10.1002/anie.200460587

Liu, Z. & Han, G. (2015). Production of solid fuel biochar from waste biomass by low temperature pyrolysis. Fuel, 158, 159–165. http://dx.doi.org/10.1016/j.fuel. 2015.05.032

Mohanty, A. K., Misra, M. & Hinrichsen, G. (2000). Biofibres, biodegradable polymers and biocomposites: An overview. Macromolecular Materials and Engineering, 276–277(1), 1–24. http://dx.doi.org/10.1002/(SICI)1439-2054(20000301)276:13.0.CO;2-W

Monrroy, M., Garcia, J. R., Troncoso, E. & Freer, J. (2015). Fourier transformed near infrared (FT-NIR) spectroscopy for the estimation of parameters in pretreated lignocellulosic materials for bioethanol production. Journal of Chemical Technology and Biotechnology, 90(7), 1281–1289. http://dx.doi.org/10.1002/jctb.4427

Morgano, M. A., Faria, C. G., Ferrão, M. F., Bragagnolo, N. & Ferreira, M. M. C. (2005). Determinacão de proteína em café cru por espectroscopia NIR e regressão PLS. Ciência E Tecnologia de Alimentos, 25(1), 25–31. http://dx.doi.org/10.1590/S0101-20612005000100005

Mosier, N., Wyman, C., Dale, B., Elander, R., Lee, Y. Y., Holtzapple, M. & Ladisch, M. (2005). Features of promising technologies for pretreatment of lignocellulosic biomass. Bioresource Technology, 96(6), 673–686. http://dx.doi.org/10.1016/j.biortech.2004.06.025

Nørgaard, L., Hahn, M. T., Knudsen, L. B., Farhat, I. A. & Engelsen, S. B. (2005). Multivariate near-infrared and Raman spectroscopic quantifications of the crystallinity of lactose in whey permeate powder. International Dairy Journal, 15(12), 1261–1270. http://dx.doi.org/10.1016/j.idairyj.2004.12.009

Ortiz, M., Sarabia, L., Herrero, A., Sánchez, M., Sanz, M., Rueda, M., . . . & Meléndez, M. (2003). Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares. Chemometrics and Intelligent Laboratory Systems, 69(1–2), 21–33. http://dx.doi. org/10.1016/S0169-7439(03)00110-2

Park, S., Baker, J. O., Himmel, M. E., Parilla, P. A. & Johnson, D. K. (2010). Cellulose crystallinity index: Measurement techniques and their impact on interpreting cellulase performance. Biotechnology for Biofuels, 3(1), 10. http://dx.doi.org/10.1186/1754-6834-3-10

Peng, Y., Gardner, D. J., Han, Y., Kiziltas, A., Cai, Z. & Tshabalala, M. A. (2013). Influence of drying method on the material properties of nanocellulose I: Thermostability and crystallinity. Cellulose, 20(5), 2379–2392. http://dx.doi.org/10.1007/s10570-013-0019-z

Plackett, R. L. & Burman, J. P. (1946). The design of optimum multifactorial experiments. Biometrika, 33(4), 305. http://dx.doi.org/10.2307/2332195

Rambo, M. K. D. & Ferreira, M. M. C. (2015). Determination of cellulose crystallinity of banana residues using near infrared spectroscopy and multivariate analysis. Journal of the Brazilian Chemical Society, 26(7), 1491–1499. http://dx.doi.org/10.5935/0103-5053.20150118

Rambo, M. K. D., Amorim, E. P. & Ferreira, M. M. C. (2013). Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. Analytica Chimica Acta, 775, 41–49. http://dx.doi.org/10.1016/j.aca.2013.03.015

Savitzky, A. & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical Chemistry, 36(8), 1627–1639. http://dx.doi.org/10.1021/ac60214a047

Segal, L., Creely, J. J., Martin, A. E. & Conrad, C. M. (1959). An empirical method for estimating the degree of crystallinity of native cellulose using the X-ray diffractometer. Textile Research Journal, 29(10), 786–794. http://dx.doi.org/10.1177/004051755902901003

Silva, R., Haraguchi, S. K., Muniz, E. C. & Rubira, A. F. (2009). Aplicac¸ ões de Fibras Lignocelulósicas na química de polímeros e em compósitos. Química Nova, 32(3), 661–671. http://dx.doi.org/10.1590/S0100-40422009000300010

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J. & Templeton, D. (2008). Determination of ash in biomass: Laboratory analytical procedure (LAP). Golden, CO, USA: National Renewable Energy Laboratory. Technical report NREL/TP-510-42622.

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D. & Crocker, D. (2008). Determination of structural carbohydrates and lignin in biomass: Laboratory analytical procedure (LAP). Golden, CO, USA: National Renewable Energy Laboratory. Technical Report NREL/TP-510-42618.

Sun, Y. & Cheng, J. (2002). Hydrolysis of lignocellulosic materials for ethanol production: A review. Bioresource Technology, 83(1), 1–11. http://dx.doi.org/10.1016/S0960-8524(01)00212-7

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

Terinte, N., Ibbett, R. & Schuster, K. C. (2011). Overview on native cellulose and microcrystalline cellulose I structure studied by X-ray diffraction (Waxd): Comparison between measurement techniques. Lenzinger Berichte, 89, 118–131. http://dx.doi.org/10.1163/156856198X00740

Timhadjelt, L., Serier, A., Belgacem, M. N. & Bras, J. (2015). Elaboration of cellulose based nanobiocomposite: Effect of cellulose nanocrystals surface treatment and interface melting. Industrial Crops and Products, 72, 7–15. http://dx.doi.org/10.1016/j.indcrop.2015.02.040

Tong, D. S., Xia, X., Luo, X. P., Wu, L. M., Lin, C. X., Yu, W. H., . . . & Zhong, Z. K. (2013). Catalytic hydrolysis of cellulose to reducing sugar over acid-activated montmorillonite catalysts. Applied Clay Science, 74, 147–153. http://dx.doi.org/10.1016/j.clay.2012.09.002

Valderrama, P., Braga, J. W. B. & Poppi, R. J. (2009). Estado da arte de figuras de mérito em calibrac¸ ão multivariada. Quimica Nova, 32(5), 1278–1287. http://dx.doi.org/10.1590/S0100-40422009000500034

Williams, P. & Sobering, D. (1993). Comparison of commercial near infrared transmittance and reflectance instruments for analysis of whole grains and seeds. Journal of Near Infrared Spectroscopy, 1(1), 25–32. http://dx.doi.org/10.1255/jnirs.3

Wise, B. M., Gallagher, N. B., Shaver, J. M., Windig, W. & Koch, R. S. (2006). PLS Toolbox version 4.0 for use with MATLABTM. Wenatchee, WA: Eigenvector Research.

Workman, J. J. & Weyer, L. (2008). In T. F. Group (Ed.), Practical guide to interpretive near-infrared spectroscopy. Boca Raton, FL: CRC Press.

Yang, H., Yan, R., Chen, H., Lee, D. H. & Zheng, C. (2007). Characteristics of hemicellulose, cellulose and lignin pyrolysis. Fuel, 86(12–13), 1781–1788. http://dx.doi.org/10.1016/j.fuel.2006.12.013

Zhou, C., Jiang, W., Via, B. K., Fasina, O. & Han, G. (2015). Prediction of mixed hardwood lignin and carbohydrate content using ATR-FTIR and FT-NIR. Carbohydrate Polymers, 121, 336–341. http://dx.doi.org/10.1016/j.carbpol.2014.11.062

Zidan, A. S., Rahman, Z., Sayeed, V., Raw, A., Yu, L. & Khan, M. A. (2012). Crystallinity evaluation of tacrolimus solid dispersions by chemometric analysis.

International Journal of Pharmaceutics, 423(2), 341–350. http://dx.doi.org/10.1016/j.ijpharm.2011.11.003

CAPÍTULO 3

Previsão de carboidratos estruturais no hidrolisado de biomassa de cana-de-açúcar usando espectroscopia NIR, PLS e métodos de seleção de variáveis

Abstract

Um método para previsão dos teores de glicanas e xilanas diretamente da biomassa de cana-de-açúcar (*Saccharum* spp.) foi proposto neste trabalho usando espectroscopia de infravermelho próximo (NIR) e regressão por mínimos quadrados parciais (PLS). O método proposto é uma alternativa à hidrólise do material seguido da análise por cromatografia liquida de alta eficiência (HPLC). Foram utilizados dois métodos de seleção de variáveis para melhorar a capacidade preditiva e interpretativa dos modelos, i.e, *i*) Seleção dos preditores ordenados (OPS) e *ii*) Algoritmo Genético (GA). Os espectros NIR foram obtidos da biomassa moída, seca, peneirada e sem extrativos. Os teores de glicanas e xilanas foram obtidos da análise do hidrolisado desse mesmo material pelo método de referência empregando HPLC. Os teores de glicanas e xilanas para os diferentes genótipos experimentais avaliados variaram entre 20-43% e 15-30%, respectivamente. Os melhores modelos foram obtidos com o uso do método de seleção de variáveis OPS. Os parâmetros estatísticos raiz quadrada do erro quadrático médio de previsão, coeficiente de correlação de previsão e índice de desempenho do desvio para o melhor modelo de glicanas foram 1.81, 0.95 e 2.00 e para xilanas foram 1.32, 0.94 e 1.79, respectivamente.

Palavras-chave: Glicanas, Xilanas, HPLC, OPS, GA

---

1. Introdução

A cana-de-açúcar é uma das culturas mais bem estabelecidas no Brasil e é alvo de diversos estudos como matéria prima para produção de biocombustíveis (Goldemberg et al., 2008). Além da possibilidade de conversão do caldo, pesquisas usando a fração de biomassa surgem como uma forma de agregar valor e reduzir o impacto ambiental causado pelo descarte desse resíduo (Lavarack et al., 2002). Alguns fatores que tornam esse tipo de material valorizado são seus elevados teores de glicanas e xilanas, possibilitando elevados rendimentos na conversão da biomassa ao produto final. Além dos elevados teores de açúcares, a presença de reduzidas quantidades de materiais

inorgânicos também é uma vantagem e está associada a menos problemas com o maquinário industrial (Cardona et al., 2010).

Ainda assim, o monitoramento dos teores de glicanas e xilanas é importante para que durante o processo de conversão não ocorram gastos com excesso de reagentes ou investimentos com materiais que não sejam adequados. O monitoramento de tais propriedades também é fundamental para que programas de melhoramento genético tenham as informações necessárias para uma tomada de decisão razoável (Via, Zhou, Acquah, Jiang & Eckhardt, 2014).

O método de referência para a determinação de glicanas e xilanas é através da hidrólise da biomassa seguida pela análise do hidrolisado por cromatografia líquida de alta eficiência (HPLC). Alguns problemas relacionados a esta prática são os elevados custos dos equipamentos e reagentes, o considerável consumo de tempo e a inviabilidade de se realizar a análise em campo. Estes problemas podem ser solucionados se essa determinação for realizada usando espectroscopia de infravermelho próximo (NIR). Além de ser uma técnica que pode ser realizada em menos de 30 segundos, esta é não destrutiva e não invasiva, necessitando de um mínimo manuseio de amostra. Apesar de um espectro NIR apresentar diversas informações relacionadas a propriedades físicas e químicas da amostra, estas informações estão sobrepostas e são de difícil interpretação. Dessa forma, o uso da quimiometria é fundamental para se obter informações uteis a partir dos espectros NIR. Através da quimiometria utilizam-se métodos computacionais intensivos tais como a regressão por mínimos quadrados parciais (PLS) para obtenção de modelos para a previsão da propriedade de interesse a partir dos espectros NIR.

Ainda assim, em muitas das vezes grande parte das informações condidas no espectro pode não estar relacionada ao problema em questão. Dessa forma, são utilizados métodos de seleção de variáveis para remover a informação irrelevante e, assim, melhorar a capacidade preditiva dos modelos. Algum dos métodos mais usados atualmente para seleção de variáveis em dados espectroscópicos são seleção dos preditores ordenados (OPS) e o algoritmo genético (GA).

Este trabalho teve como objetivo a construção de modelos para previsão dos teores de glicanas e xilanas em biomassa de cana-de-açúcar (*Saccharum* spp.) a partir dos espectros NIR.

2. Materiais e Métodos

2.1. Preparo das amostras de biomassa

As amostras de biomassa seguiram o procedimento descrito nos itens 2.1 e 2.2 do Capítulo 2 para preparo de amostra e remoção de extrativos.

2.2. Obtenção dos espectros NIR

A obtenção dos espectros NIR foi realizada de acordo com o disposto no item 2.5 do Capítulo 2. Os espectros foram obtidos das amostras sem extrativos com o diferencial de serem monitorados 3113 comprimentos de onda por espectro.

2.3. Hidrólise da biomassa

Seguindo o procedimento do laboratório nacional de energia renovável dos Estados Unidos (NREL) (Sluiter et al., 2008), para a determinação de carboidratos estruturais em biomassa, pesou-se em balança analítica 0,3000 g de amostra livre de extrativos. Esta massa foi misturada a 3,00 mL de H_2SO_4 72% em tubos que foram vedados. Os tubos foram deixados em banho à 30 ºC por 120 minutos com agitação a cada 5 minutos sem a remoção dos mesmos do banho. Essa primeira etapa de hidrolise é útil para que ocorra o inchamento da celulose cristalina, diminuindo sua recalcitrância e facilitando sua total conversão em glicose. Após essa etapa, o ácido foi diluído para uma concentração de 4% pela adição de 84 mL de H_2O deionizada aos tubos que, após agitação, foram novamente lacrados e autoclavados à 121 ºC por 60 minutos. Após os tubos serem removidos da autoclave estes foram resfriados lentamente até temperatura ambiente. O hidrolisado obtido foi filtrado, ajustou-se o pH para 5 com adição de $CaCO_3$ e realizou-se novamente a filtração com filtro de papel. Foi realizada outra etapa de filtração com filtros de seringa de nylon com poros de 0,22 μm e o hidrolisado foi transferido para recipientes *vials*.

2.4. Análise por HPLC

A quantificação dos monômeros presentes do hidrolisado foi realizada utilizando um sistema HPLC Shimadzu. A coluna cromatográfica utilizada foi uma Rezex™ (Phenomenex) RFQ Fast Acid H (8%) de 100-7,8 mm precedida por uma pré-coluna de 2-7,8 mm, ambas operando à 60 °C. A eluição foi feita com uma solução de ácido acético

0,08% a uma vazão de 0,85 mL/min. A detecção do analito foi feita através de um detector ELSD.

As curvas de calibração utilizadas na determinação das concentrações de carboidratos no hidrolisado foram construídas com padrões de D(+)glicose. D(+)xilose e L(+)arabinose em uma faixa de concentração que variou de 0,1 até 4,0 mg/mL.

2.5. Cálculo dos teores de açúcares em massa seca livre de extrativos

Para corrigir uma possível degradação dos monômeros durante a etapa de hidrólise, uma mistura contendo uma massa conhecida de cada padrão de carboidrato foi submetida às etapas de hidrólise. Dessa forma, obteve-se um constante de recuperação a fim de corrigir este efeito. Para o cálculo da correção foram realizadas três repetições para cada carboidrato analisado. O cálculo adotado para obtenção da porcentagem constante de recuperação ($\%_R$) para cada açúcar é descrito abaixo.

$$\%_R = \frac{\textit{concentração após a hidrólise}}{\textit{concentração antes da hidrólise}} \times 100$$

Para as amostras analisadas, a concentração corrigida (C_{corr}) no hidrolisado pode ser obtida por:

$$C_{corr} = \frac{\textit{concentração detectada}}{\%_R / 100}$$

Uma vez que, na biomassa, os monômeros estão ligados entre si na forma de longas cadeias, o fator de hidratação (f_{anidro}) durante sua quebra deve ser considerado. Dessa forma, para pentoses (xilose e arabinose) foi adotado um fator de correção de 0,88 (132/150) e um fator de 0,90 (162/180) para hexoses (glicose), de maneira que:

$$C_{anidro} = c_{corr} \times f_{anidro}$$

Assim, finalmente, os teores de glicanas (*i.e.*, polímero de glicose), xilanas (*i.e.*, polímero de xilose) e arabinanas (*i.e.*, polímero de arabinose) em base livre de extrativos ($\%_{Pol}$) podem ser calculados através de:

$$\%_{Pol} = \frac{C_{Anidro} \times V}{M_{LE}} \times 100$$

em que V e M_{LE} representam o volume do hidrolisado (em mL) e a massa seca livre de extrativos (em mg), respectivamente. Esse procedimento foi realizado para a obtenção das variáveis dependentes.

2.6. Análise de dados

A análise de dados foi realizada de acordo com o disposto no item 2.6 do Capítulo 2. O diferencial desse estudo se dá uma vez que os valores de **y** foram compostos pelos teores estudados. Para a escolha do tratamento espectral que forneceria melhores modelos para a previsão dos teores, diversos tratamentos foram aplicados aos espectros e estes foram centrados na média. Primeiro foi feito um estudo a respeito das janelas ótimas a serem aplicadas aos tratamentos de derivadas pelo algoritmo Savitzky-Golay. Foram então investigados diferentes tratamentos tais como o uso dos dados crus (*Raw*), primeira (D1) e segunda (D2) derivadas, correção multiplicativa de sinal (MSC), variável normal padrão (SNV) e correção de linha de base (*baseline*) (Wise, Gallagher, Shaver, Windig & Koch, 2006).

2.7. Figuras de Mérito

As figuras de mérito utilizadas na avaliação dos modelos foram as mesmas já descritas no item 2.7 do Capítulo 2.

2.8. Parâmetros usados na seleção de variáveis GA

Os parâmetros utilizados no algoritmo GA foram escolhidos com base no resultado de um planejamento experimental Plackett-Burman (Plackett & Burman, 1946). Os melhores parâmetros de entrada do algoritmo foram definidos com base nos valores de *RMSECV* e *RCV* obtidos para os modelos construídos com os diferentes conjuntos de variáveis selecionados para cada condição experimental. Os parâmetros e seus respectivos valores ótimos adotados foram: tamanho da população 152; gerações 150; taxa de mutação 0.001; tamanho de janela 3; convergência 20; termos iniciais 20; crossover 1; divisões 3; replicatas 3 e interações 1.

3. Resultados e Discussão

3.1. Resultados do HPLC

Através dos resultados apresentados na Figura 1 observa-se a distribuição dos teores de glicanas, xilanas e arabinanas obtidos a partir da análise do hidrolisado de diferentes amostras analisadas por HPLC. Para glicanas esses teores variaram entre 20,22 a 43,24%, para xilanas de 15,47 a 30,20% e arabinanas de 4,85 a 6,34%. Estes resultados concordam

parcialmente com os encontrados na literatura para esse tipo de biomassa quando avaliada a fração de xilanas e arabinanas (de Souza et al., 2013; Gírio et al., 2010) porém estão abaixo do esperado para glicanas (de Souza et al., 2013; de Vasconcelos et al., 2013).

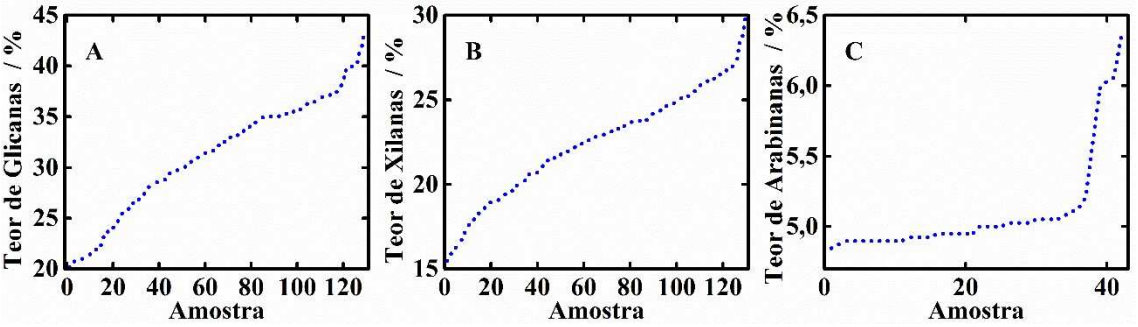

Figura 1. Distribuição dos teores de (A) glicanas, (B) xilanas e (C) arabinanas para as amostras estudadas.

3.2. Resultados NIR

Os espectros NIR obtidos para diferentes amostras de biomassa de cana-de-açúcar são apresentados na Figura 2. Nessa são apresentados os espectros das amostras com maiores, menores, e valores intermediários para os teores de glicanas, xilanas e arabinanas. Observa-se que mesmo com os teores significativamente diferentes os espectros são muito semelhantes para as diferentes amostras. Aparentemente, as regiões com maior variabilidade espectral estão entre os números de onda 4000 e 5000 cm^{-1} e entre 5300 e 7000 cm^{-1}. Os espectros NIR obtidos estão de acordo com os já reportados na literatura para outros tipos de materiais lignocelulósicos (Jiang et al., 2007; Rambo, Alves, Garcia & Ferreira, 2015; Rambo, Ferreira & Amorim, 2016; Rambo & Ferreira, 2015).

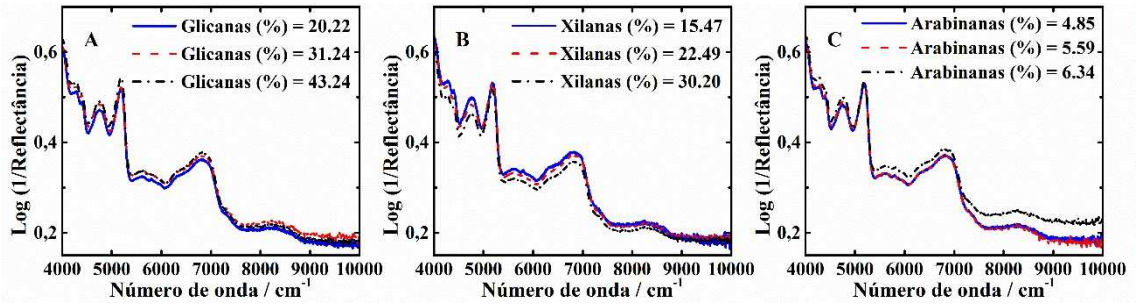

Figura 2. Espectros NIR das amostras que apresentaram maior diferenciação nos teores de (A) glicanas, (B) xilanas e (C) arabinanas.

3.3. Construção dos Modelos

Para a construção dos modelos, as amostras de biomassa de cana-de-açúcar foram separadas pelo algoritmo Kennard-Stone (Kennard & Stone, 1996) de forma que 75% dessas foram destinadas ao conjunto de calibração e 25% para o conjunto de previsão (Rambo et al., 2013). Para a construção e validação dos modelos de glicanas, xilanas e arabinanas foram utilizadas ao todo 129, 130 e 41 amostras, respectivamente. Não foi possível construir modelos satisfatórios para a previsão dos teores de arabinanas, de forma que os modelos obtidos para esta fração não serão apresentados.

3.3.1. Escolha do tratamento ideal aos espectros

Os parâmetros estatísticos *RMSECV, RCV* e *RPD* obtidos dos diferentes tratamentos espectrais estudados foram comparados em cada caso. Observou-se que o tratamento ótimo para o estudo de glicanas e xilanas foi o de segunda derivada com janela 5 pelo algoritmo Savitzky-Golay seguido da centragem na média das colunas da matriz **X**. O tratamento ideal foi escolhido com base no menor valor de *RMSECV* e maiores valores de *RCV* e *RPD*. Esses resultados são apresentados na Figura 3. Em ambos os casos o número ideal de variáveis latentes para o modelo (*h*Mod) foi igual a 2.

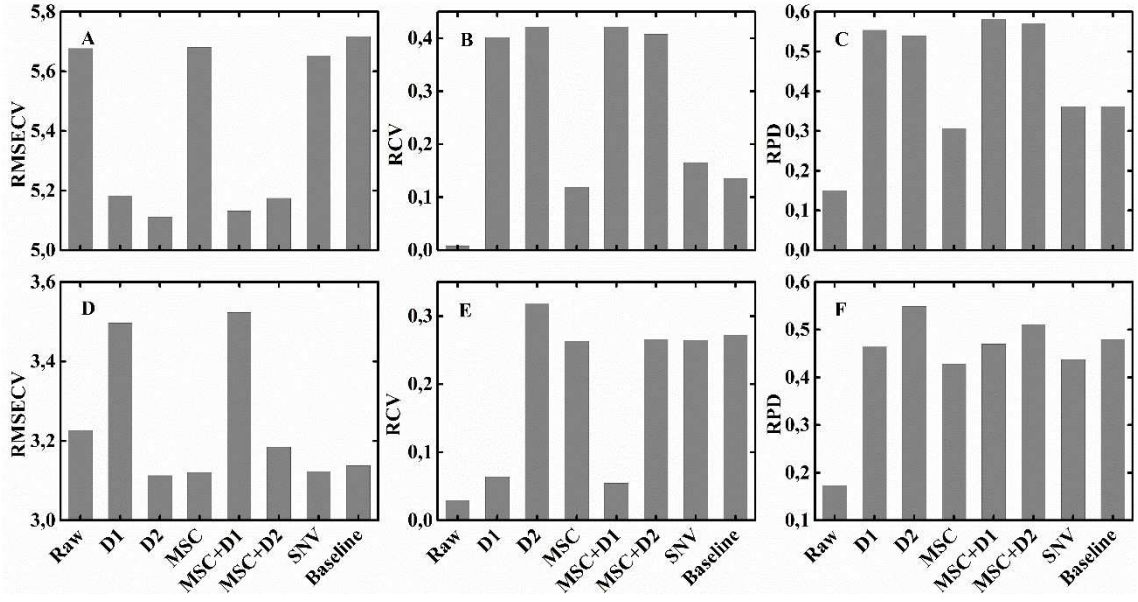

Figura 3. Parâmetros *RMSECV*, *RCV* e *RPD* obtidos para os modelos de glicanas (A, B e C) e xilanas (D, E e F) com diferentes tratamentos espectrais. Legenda: Dados crus (Raw), primeira (D1) e segunda (D2) derivadas, correção multiplicativa de sinal (MSC), variável normal padrão (SNV) e correção de linha de base (Baseline). Todos os dados foram centrados na média.

3.3.2. Seleção de variáveis com o algoritmo OPS

Após investigação do número de variáveis latentes usadas na construção do vetor informativo do algoritmo OPS, encontraram-se valores ideais de *h*OPS (Teófilo et al., 2009) iguais a 9 e 4, para glicanas e xilanas, respectivamente. Estes foram escolhidos com base nos resultados apresentados nas Figuras 4A, 4B, 4E e 4F onde observa-se que para estes valores de *h*OPS obtém-se os menores valores para *RMSECV* e os maiores valores para *RCV*. Os resultados obtidos no estudo de diferentes vetores informativos tais como (1) vetor de regressão (R) construído com *h*OPS ótimo; (2) correlação entre as colunas de **X** e **y** (C); (3) resíduos (S) e (4) vetor NAS (V) assim como suas combinações (Teófilo et al., 2009) são mostrados nas Figuras 4C e 4G. Através desses resultados identificou-se o vetor de combinações RCSV como sendo o mais indicado para uso no algoritmo OPS, uma vez que apresentou menores valores de *RMSECV*. Usando essas condições, o algoritmo OPS foi aplicado diversas vezes nas variáveis selecionadas até que os valores de *RMSECV* não fossem reduzidos de forma significativa, como mostrado nas Figuras 4D e 4H. O valor de *RMSECV* do modelo com todas as variáveis (FULL) é representado como uma linha horizontal. Dentro das colunas da Figura 4D e 4H são apresentados, respectivamente, o número de variáveis independentes selecionadas para uso nos modelos de glicanas e xilanas para os ciclos sequenciais do OPS.

Os resultados apresentados nas Figuras 4D e 4H indicam uma melhora significativa na capacidade preditiva dos modelos ao serem selecionadas as variáveis pelo algoritmo OPS. A partir do método OPS foi possível uma redução do *RMSECV* de 5,18 para 2,16 e de 3,13 para 1,27 para o modelo de glicanas e xilanas, respectivamente.

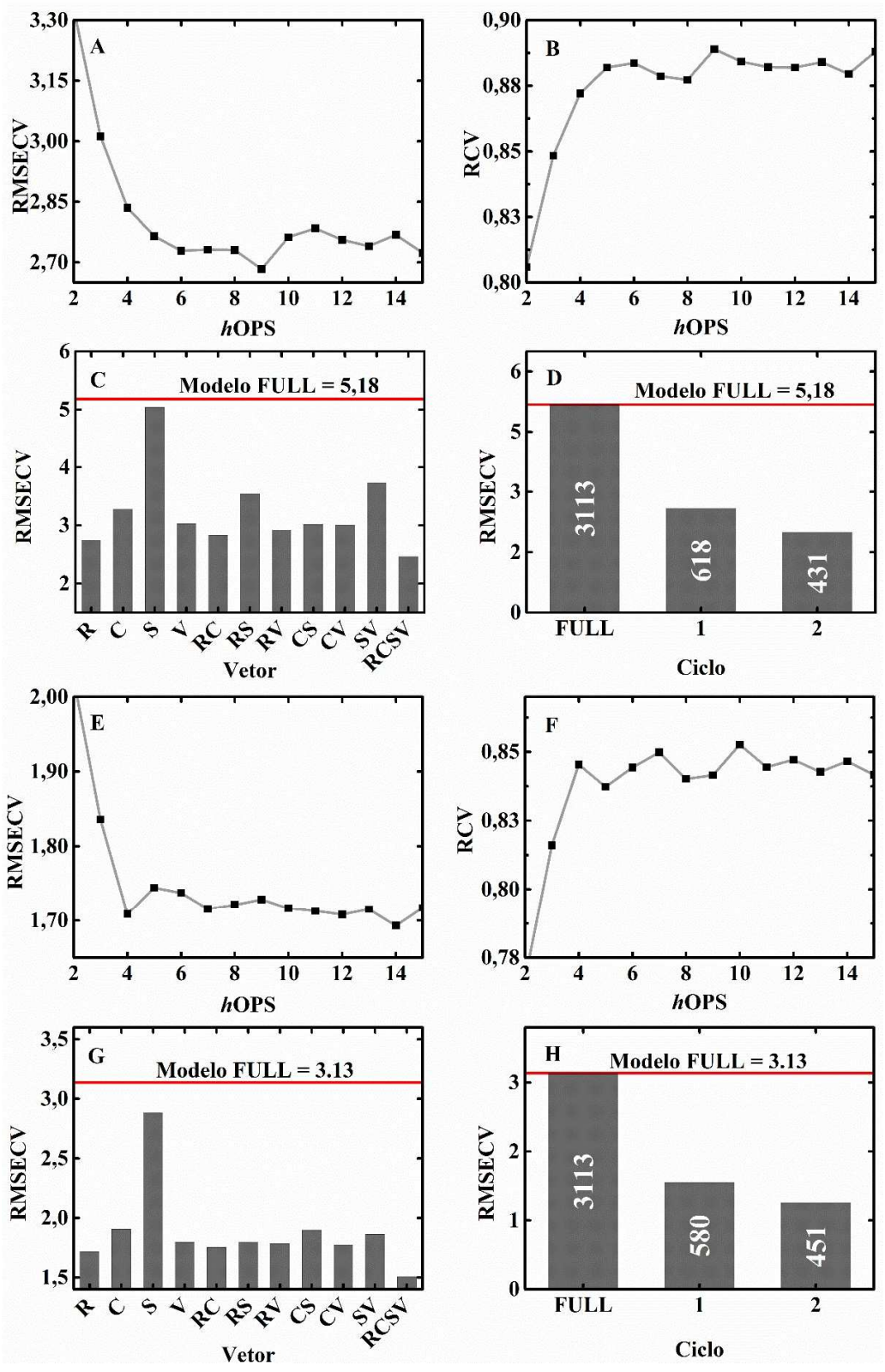

Figura 4. Resultados obtidos para os modelos de glicanas (A, B, C e D) e xilanas (E, F, G e H). (A e E) *RMSECV* para diferentes valores de *h*OPS; (B e F) *RCV* para diferentes valores de *h*OPS; (C e G) *RMSECV* para os diferentes vetores informativos e (D e H) *RMSECV* para os ciclos sequenciais do OPS; Legenda dos vetores informativos: Regressão (R), Correlação (C), Resíduos (S) e NAS (V).

3.3.3. Comparação dos resultados

Os resultados apresentados nas Figuras 5 e 6 indicam que houve uma melhor distribuição dos dados em torno da linha a 45º e um estreitamento da curva de distribuição dos erros relativos no caso onde os modelos foram construídos com as variáveis selecionadas pelo algoritmo OPS. Em todos os casos, o uso de métodos de seleção de variáveis forneceu resultados melhores do que os obtidos pelos modelos construídos com todas as variáveis (FULL).

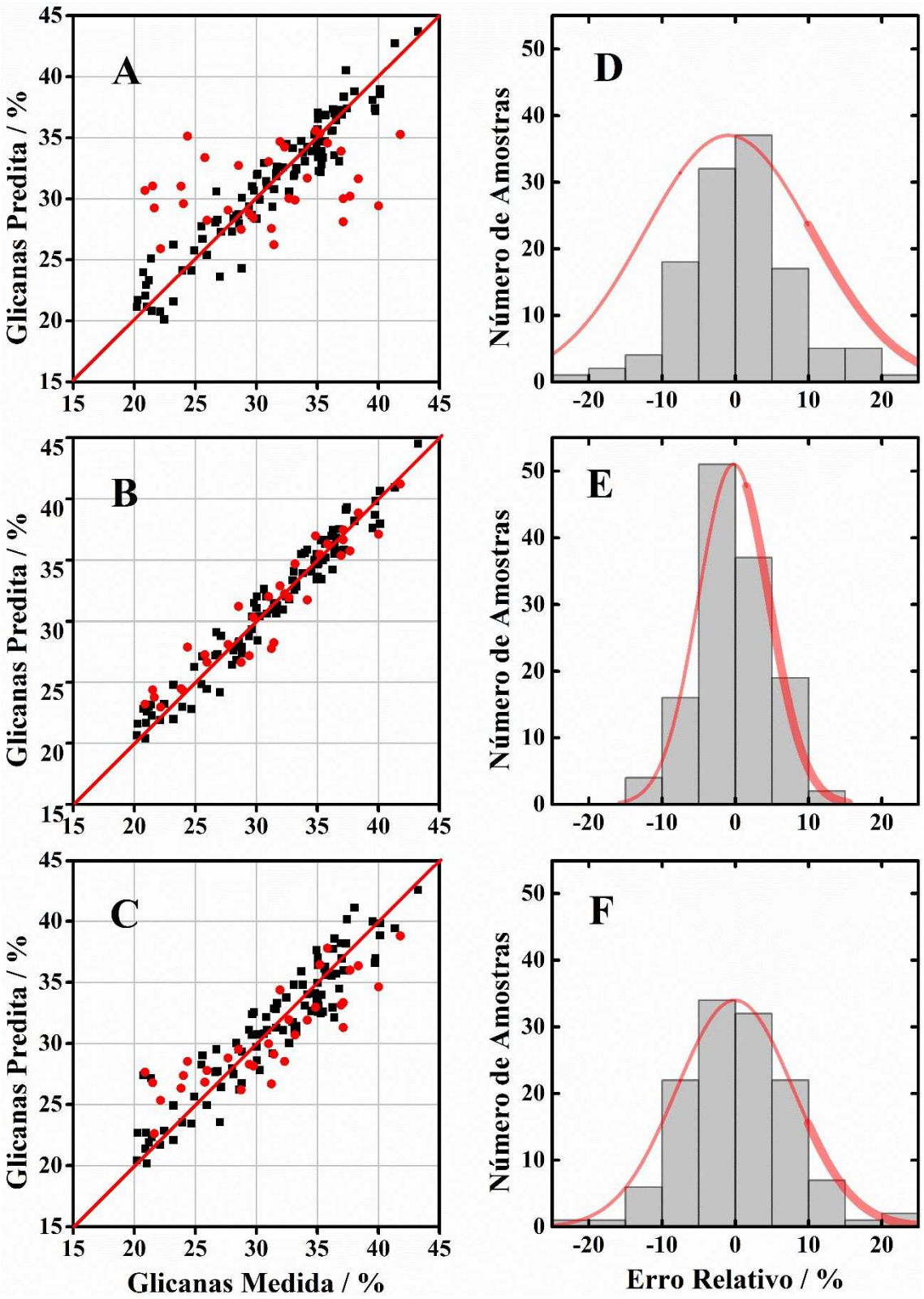

Figura 5. Relação entre teores de glicanas medidos e preditos para os modelos: (A) FULL; (B) OPS; (C) GA (■ representa as amostras do conjunto calibração e ● as amostras do conjunto previsão); Histogramas com os erros relativos para o modelo: (D) FULL; (E) OPS; (F) GA.

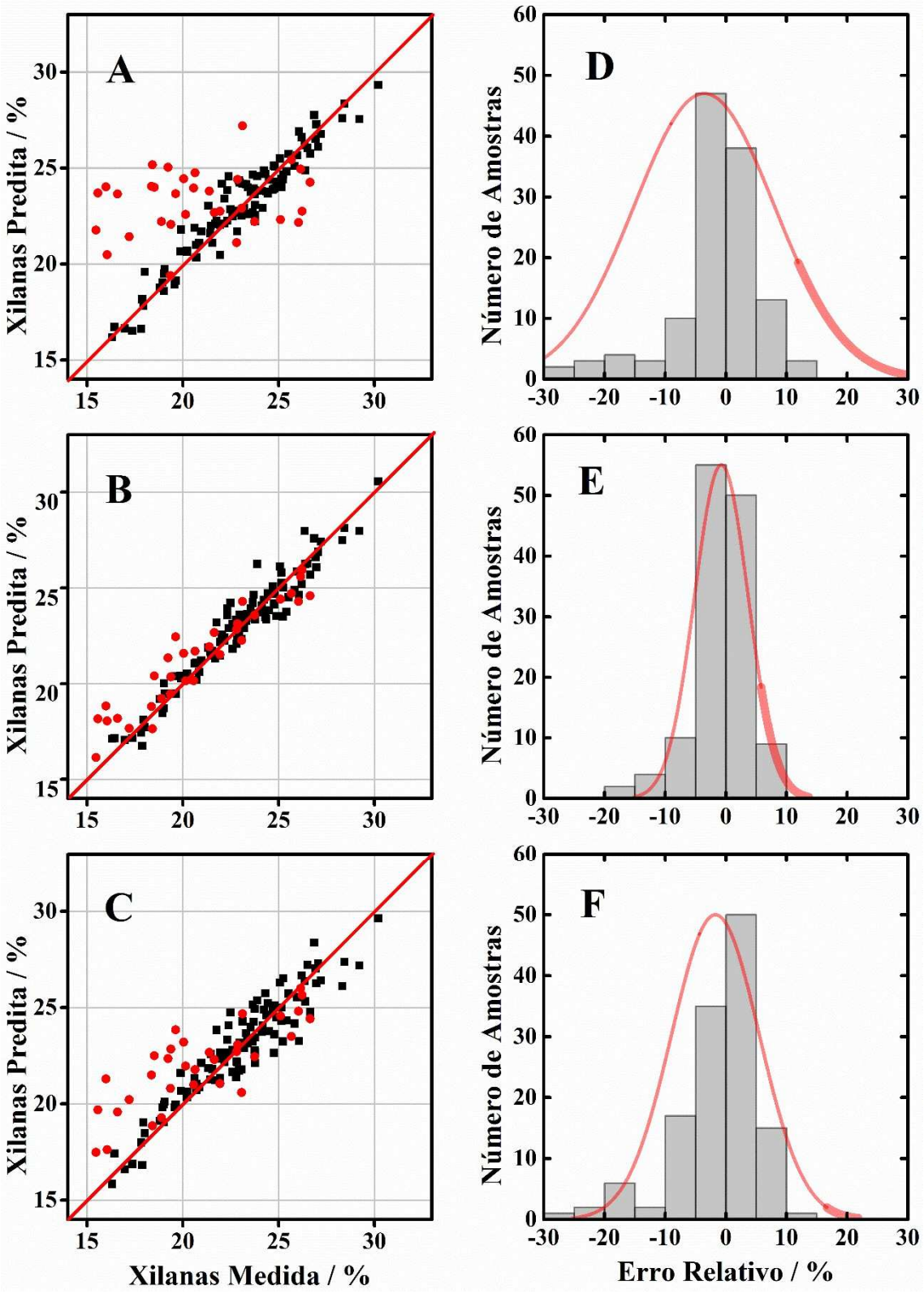

Figura 6. Relação entre teores de xilanas medidos e preditos para os modelos: (A) FULL; (B) OPS; (C) GA (■ representa as amostras do conjunto calibração e ● as amostras do conjunto previsão); Histogramas com os erros relativos para o modelo: (D) FULL; (E) OPS; (F) GA.

Analisando os dados apresentados na Tabela 1, é possível verificar que o uso dos métodos de seleção de variáveis forneceu modelos com melhor qualidade para a maioria das figuras de mérito avaliadas. Em especial, os modelos construídos com a seleção de variáveis OPS apresentaram melhores valores de *RMSECV*, *RCV*, *RMSEP*, *RP*, *RPD*, *SEL* e *LOD*.

Tabela 1. Parâmetros estatísticos para os modelos de glicanas.

	Modelos de Glicanas			**Modelos de Xilanas**		
	FULL	**OPS**	**GA**	**FULL**	**OPS**	**GA**
Vetor	-	RCSV	-	-	RCSV	-
***h*OPS**	-	9	-	-	4	-
***h*Mod**	2	2	2	2	2	2
***n*Vars**	3113	431	600	3113	451	621
RMSECV	5,18	2,16	3,74	3,04	1,27	2,11
RCV	0,37	0,94	0,74	0,17	0,91	0,69
RMSEP	5,65	1,81	3,12	4,19	1,32	2,35
RP	0,28	0,95	0,86	0,19	0,94	0,81
RPD	0,51	2,00	1,03	0,50	1,79	0,91
SEN	$5{,}36\times10^{-4}$	$2{,}72\times10^{-4}$	$3{,}32\times10^{-4}$	$9{,}87\times10^{-4}$	$5{,}27\times10^{-4}$	$5{,}32\times10^{-4}$
SEL	0,08	0,15	0,11	0,08	0,14	0,10
γ	2,60	2,37	2,97	4,26	2,16	6,54
***γ*⁻¹**	0,38	0,42	0,34	0,23	0,46	0,15
LOD	4,20	2,71	5,16	3,30	2,05	3,39

RCSV: Vetor de combinações usado como vetor informativo no algoritmo OPS; *h*OPS e *h*Mod. Número de variáveis latentes usado no OPS e no modelo, respectivamente; *n*Vars: Número de variáveis independentes usadas na construção do modelo; *RMSECV*: Raiz quadrática média do erro de validação cruzada; *RMSEP*: Raiz quadrática média do erro de previsão, *RCV*: Coeficiente de correlação da validação cruzada; *RP*: Coeficiente de correlação da previsão; *RPD*: Índice de desempenho do desvio. *SEN*: Sensibilidade *SEL*: Seletividade. γ: Sensibilidade analítica. γ^{-1}: Inverso da sensibilidade analítica. *LOD*: Limite de detecção.

Observa-se que o uso dos métodos de seleção de variáveis foi essencial para a obtenção de modelos úteis, uma vez que, enquanto os modelos FULL forneceram pobres previsões, os modelos OPS apresentaram elevada capacidade preditiva tanto para previsão dos teores de glicanas quanto de xilanas. Nos dois casos os modelos OPS

obtiveram os menores valores de *RMSECV* e *RMSEP* enquanto que os valores de *RCV* e *RP* foram superiores à 0.90, sendo, dessa forma, considerados excelentes.

Os baixos valores de *SEN* e *SEL* podem ser atribuídos ao tratamento aplicado aos espectros e à elevada quantidade de interferentes presente na biomassa, respectivamente, estando de acordo com o esperado e já apresentado em outros trabalhos (Caliari, Barbosa, Ferreira & Teófilo, 2017; Rambo & Ferreira, 2015). Apesar dos modelos OPS não terem apresentado os melhores valores para γ^{-1}, ainda assim, estes parâmetros estão muito próximo dos obtidos para os demais modelos e foram considerado satisfatórios. Em relação ao *LOD* dos modelos OPS, foram obtidos resultados satisfatórios, uma vez que estes estiveram consideravelmente abaixo dos menores valores experimentais para os teores de glicanas e xilanas estudados.

Considera-se que as diferentes formas de representação dos teores, i.e. em termos de glicose/glicanas ou xilose/xilanas, introduz uma variação nos resultados estatísticos mesmo que as amostras sejam essencialmente iguais. Dessa forma, os valores de *RMSECV*, *RMSEP*, *RCV* e *RP* obtidos neste trabalho são condizentes aos apresentados para os modelos obtidos nos trabalhos de Hayes (2012), Rambo et. al. (2016), Payne e Wolfrum (2015) e Rambo et. al. (2015) para diferentes tipos de biomassa. A maior diferença entre os modelos se dá em relação ao número de variáveis latentes usada no PLS. Enquanto que os modelos citados foram construídos com, respectivamente, pelo menos 8, 7, 9 e 3 variáveis latentes para glicose/glicanas e 13, 4, 9 e 3 para xilose/xilanas, neste trabalho foram utilizadas apenas 2 em ambos os casos.

Através dos resultados apresentados na Figura 7 é possível observar que ambos métodos de seleção de variáveis selecionaram números de onda sobre toda a região espectral estudada. Para o modelo de glicanas construído com OPS a região mais selecionada foi entre 4500 e 5000 cm^{-1} enquanto que houve uma rejeição de números de onda superiores a 9000 cm^{-1}. Ainda para o modelo de glicanas, o algoritmo GA atribuiu maior importância às regiões entre 4500-5000 cm^{-1} e 7500-8000 cm^{-1}. Comparando os resultados obtidos pelos dois métodos de seleção de variáveis, a região de 4500-5000 cm^{-1} certamente é responsável por grande parte da capacidade preditiva dos modelos. Workman (2008) atribui a banda em 4785 cm^{-1} ao estiramento de primeiro sobretom de O-H polimérico de celulose, de forma que a maior seleção de números de onda na região em 4500-5000 para o modelo de glicanas pode ser justificada.

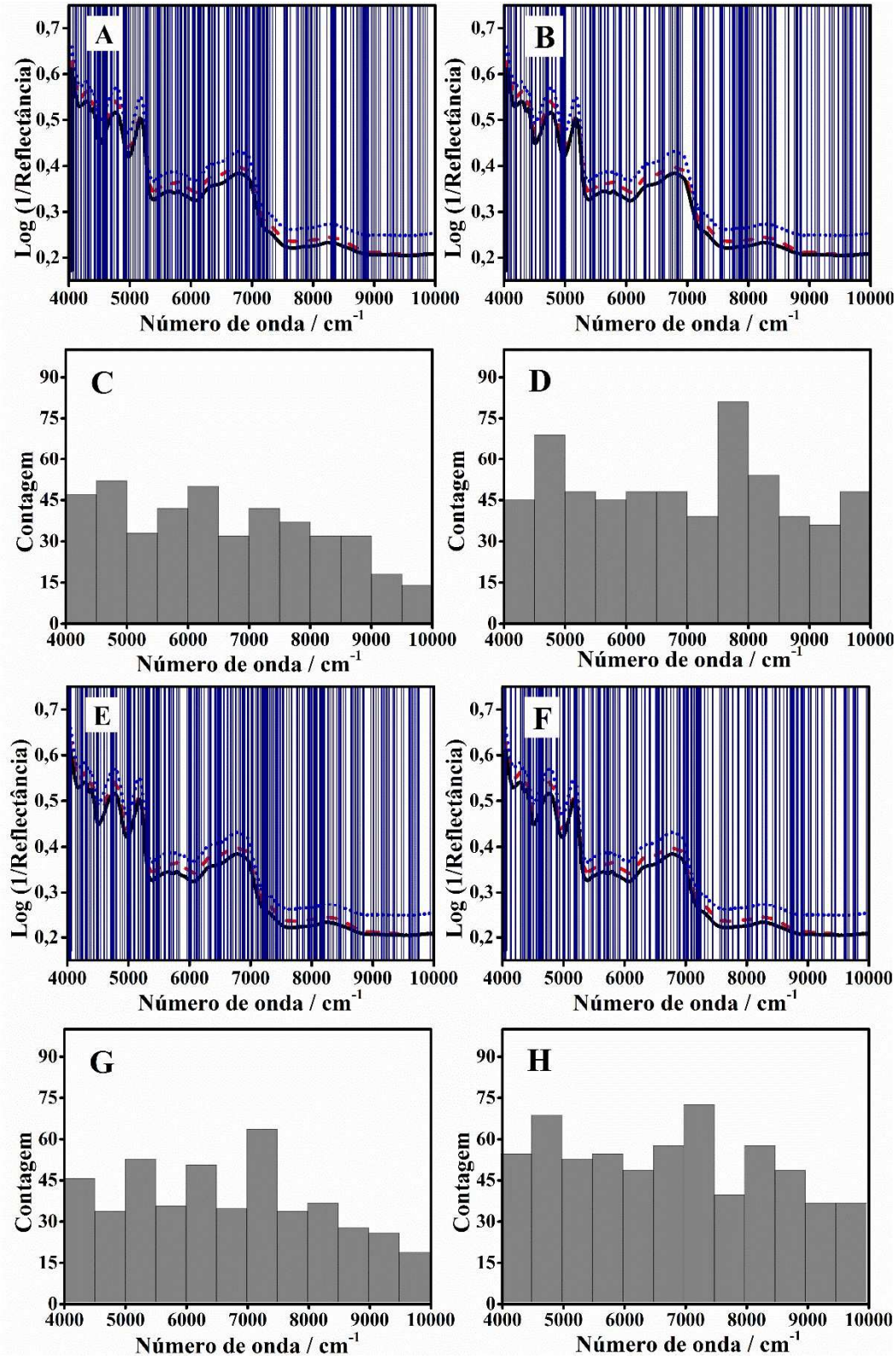

Figura 7. Estudo das variáveis selecionadas para os modelos de Glicanas (A, B, C e D) e Xilanas (E, F, G e H). (A, C, E e G) Seleção realizada com algoritmo OPS; (B, D, F e H) Seleção realizada com GA; Linhas verticais indicam as variáveis selecionadas e as colunas contem a contagem de variáveis selecionadas em um intervalo de números de onda.

Para o modelo de xilanas construído com a seleção OPS, observa-se uma maior seleção de números de onda entre 4000-4500, 5000-5500 6000-6500 e 7000-7500 e novamente uma rejeição a partir de 9000 cm^{-1} enquanto que o GA destacou em maior parte as regiões entre 4500-5000 e 7000-7500. Comparando os resultados de ambos os métodos de seleção, para os modelos de xilanas, foram selecionados em maior quantidade números de onda entre 7000 e 7500 cm^{-1} enquanto que números de onda superiores a 9000 cm^{-1} foram evitados.

4. Conclusão

Pode-se concluir que a previsão dos teores de glicanas e xilanas em biomassa de cana-de-açúcar foi possível e modelos com elevada capacidade preditiva foram construídos para este propósito. O uso de modelos de calibração multivariada para previsão dos teores de glicanas e xilanas na biomassa é útil para a obtenção de resultados mais rápidos, baratos e ainda assim confiáveis. Assim, o desenvolvimento desses métodos em substituição aos de referência mostraram-se viáveis e recomendados. O uso de métodos de seleção de variáveis foi indispensável para melhorar a capacidade preditiva e interpretativa dos modelos. Dentre os dois métodos de seleção de variáveis estudados o OPS teve destaque, superando o GA em termos da capacidade preditiva e do número de variáveis selecionadas.

5. Referências

Caliari, Í. P., Barbosa, M. H. P., Ferreira, S. O., & Teófilo, R. F. (2017). Estimation of cellulose crystallinity of sugarcane biomass using near infrared spectroscopy and multivariate analysis methods. *Carbohydrate Polymers*, *158*, 20–28. http://doi.org/10.1016/j.carbpol.2016.12.005

Cardona, C. A., Quintero, J. A., & Paz, I. C. (2010). Production of bioethanol from sugarcane bagasse: Status and perspectives. *Bioresource Technology*, *101*(13), 4754–4766. http://doi.org/10.1016/j.biortech.2009.10.097

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

de Vasconcelos, S. M., Santos, A. M. P., Rocha, G. J. M., & Souto-Maior, A. M. (2013). Diluted phosphoric acid pretreatment for production of fermentable sugars in a sugarcane-based biorefinery. *Bioresource Technology*, *135*, 46–52. http://doi.org/10.1016/j.biortech.2012.10.083

Gírio, F. M., Fonseca, C., Carvalheiro, F., Duarte, L. C., Marques, S., & Bogel-Łukasik, R. (2010). Hemicelluloses for fuel ethanol: A review. *Bioresource Technology*, *101*(13), 4775–4800. http://doi.org/10.1016/j.biortech.2010.01.088

Goldemberg, J., Coelho, S. T., & Guardabassi, P. (2008). The sustainability of ethanol production from sugarcane. *Energy Policy*, *36*(6), 2086–2097. http://doi.org/10.1016/j.enpol.2008.02.028

Hayes, D. J. M. (2012). Development of near infrared spectroscopy models for the quantitative prediction of the lignocellulosic components of wet Miscanthus samples. *Bioresource Technology*, *119*, 393–405. http://doi.org/10.1016/j.biortech.2012.05.137

Jiang, Z. H., Yang, Z., So, C. L., & Hse, C. Y. (2007). Rapid prediction of wood crystallinity in Pinus elliotii plantation wood by near-infrared spectroscopy. *Journal of Wood Science*, *53*(5), 449–453. http://doi.org/10.1007/s10086-007-0883-y

Lavarack, B. P., Griffin, G. J., & Rodman, D. (2002). The acid hydrolysis of sugarcane bagasse hemicellulose to produce xylose, arabinose, glucose and other products. *Biomass and Bioenergy*, *23*(5), 367–380. http://doi.org/10.1016/S0961-

9534(02)00066-1

Payne, C. E., & Wolfrum, E. J. (2015). Rapid analysis of composition and reactivity in cellulosic biomass feedstocks with near-infrared spectroscopy. *Biotechnology for Biofuels*, *8*(1), 43. http://doi.org/10.1186/s13068-015-0222-2

Plackett, R. L., & Burman, J. P. (1946). The Design of Optimum Multifactorial Experiments. *Biometrika*, *33*(4), 305. http://doi.org/10.2307/2332195

Rambo, M. K. D., Alves, A. R., Garcia, W. T., & Ferreira, M. M. C. (2015). Multivariate analysis of coconut residues by near infrared spectroscopy. *Talanta*, *138*, 263–272. http://doi.org/10.1016/j.talanta.2015.03.014

Rambo, M. K. D., Amorim, E. P., & Ferreira, M. M. C. (2013). Potential of visible-near infrared spectroscopy combined with chemometrics for analysis of some constituents of coffee and banana residues. *Analytica Chimica Acta*, *775*, 41–49. http://doi.org/10.1016/j.aca.2013.03.015

Rambo, M. K. D., & Ferreira, M. M. C. (2015). Determination of Cellulose Crystallinity of Banana Residues Using Near Infrared Spectroscopy and Multivariate Analysis. *Journal of the Brazilian Chemical Society*, *26*(7), 1491–1499. http://doi.org/10.5935/0103-5053.20150118

Rambo, M. K. D., Ferreira, M. M. C., & Amorim, E. P. (2016). Multi-product calibration models using NIR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*, *151*, 108–114. http://doi.org/10.1016/j.chemolab.2015.12.013

Sluiter, A., Hames, B., Ruiz, R., Scarlata, C., Sluiter, J., Templeton, D., & Crocker, D. (2008). *"Determination of Structural Carbohydrates and Lignin in Biomass: Laboratory Analytical Procedure (LAP)." Technical Report NREL/TP-510-42618*. Golden, CO, USA: National Renewable Energy Laboratory.

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

Via, B. K., Zhou, C., Acquah, G., Jiang, W., & Eckhardt, L. (2014). Near infrared spectroscopy calibration for wood chemistry: which chemometric technique is best for prediction and interpretation? *Sensors (Basel, Switzerland)*, *14*(8), 13532–13547. http://doi.org/10.3390/s140813532

Wise, B. M., Gallagher, N. B., Shaver, J. M., Windig, W., & Koch, R. S. (2006). *PLS_Toolbox Version 4.0 for use with MATLAB™*. Wenatchee, WA: Eigenvector Research.

Workman, J. J., & Weyer, L. (2008). *Practical Guide to Interpretive Near-Infrared Spectroscopy*. (T. & F. Group, Ed.). Boca Raton, FL: CRC Press. Retrieved from http://www.amazon.com/Practical-Guide-Interpretive-Near-Infrared-Spectroscopy/dp/157444784X

Conclusão Geral

Pode-se concluir que o uso da quimiometria é fundamental para um melhor aproveitamento de técnicas como a espectroscopia NIR. Através desta foi possível a construção de modelos para previsão de propriedades importantes da biomassa tais como a cristalinidade da celulose e os teores de glicanas e xilanas. A implementação de tais modelos é de extrema relevância uma vez que se obtém uma redução de custos ao mesmo tempo que resultados confiáveis passam a ser obtidos quase que instantaneamente. Para a obtenção dos modelos, o uso de métodos de seleção de variáveis foi fundamental. A aplicação de tais métodos de seleção foi responsável pela construção de modelos com elevada capacidade preditiva. Através destes também foi possível identificar regiões e bandas dos espectros NIR responsáveis pela previsão de determinadas propriedades, evidenciando, dessa forma, a função interpretativa de tais métodos de seleção. O método de seleção de variáveis OPS mostrou-se eficiente e foi superior ao mais difundido GA em todos os casos estudados. Finalmente, os modelos construídos serão uteis para uma tomada de decisão mais dinâmica em programas de melhoramento genético e na indústria.