UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Combinação de infravermelho próximo e ferramentas quimiométricas para análise de material desfibrado de cana-de-açúcar

Renato de Carvalho*

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre Profissional em Química, área de concentração: Química Tecnológica.

Orientador: Prof. Dr. Edenir Rodrigues Pereira Filho
*Empresa: Monsanto

São Carlos
Janeiro de 2015

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

C331ci

Carvalho, Renato de.
Combinação de infravermelho próximo e ferramentas quimiométricas para análise de material desfibrado de cana-de-açúcar / Renato de Carvalho. -- São Carlos : UFSCar, 2015.
57 f.

Dissertação (Mestrado profissional) -- Universidade Federal de São Carlos, 2015.

1. Quimiometria. 2. Cana-de-açúcar. 3. Infravermelho. 4. Melhoramento genético. 5. Calibração. I. Título.

CDD: 543.072 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Química

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Renato de Carvalho, realizada em 29/01/2015:

Prof. Dr. Edenir Rodrigues Pereira Filho
UFSCar

Prof. Dr. Paulo Cezar Vieira
UFSCar

Prof. Dr. Mauro Sérgio de Oliveira Leite
MONSANTO

Agradecimentos

Gostaria de agradecer primeiramente a Deus por me dar vida e saúde tornando possível este trabalho;

Ao Prof. Dr. Edenir Rodrigues Pereira Filho, por ter me orientado neste trabalho com muita paciência, vontade e dedicação.

A todos os professores e demais membros do GAIA pelas trocas de experiências durante as reuniões de grupo. Aos funcionários do Departamento de Química, especialmente aqueles da Secretaria de Pós Graduação por terem sido prestativos. Ao Programa de Mestrado Profissional da Universidade Federal de São Carlos pela oportunidade em desenvolver este trabalho.

À empresa Monsanto do Brasil pela oportunidade e a todos os seus colaboradores em especial aos meus companheiros de trabalho e amigos José Guilherme Scanavini, Tiago Serrani, Mauro Leite, Everaldo Coutinho, Hugo Rosa e Reginaldo Fragoso do time de Conchal, a todo time de funcionários da estação da Monsanto de Araçatuba e Mandaguaçu por terem me apoiado e contribuído para o meu projeto.

A minha família e amigos pessoais, que, graças a Deus, são tantos que não caberiam nesta página. Obrigado

Lista de Abreviaturas

CONSECANA	-	Conselho dos produtores de cana-de-açúcar, açúcar e álcool do estado de São Paulo
IR distante	-	Infravermelho distante
IR médio	-	Infravermelho médio
IR próximo	-	Infravermelho próximo
IR	-	Infravermelho
NIPALS	-	Nonlinear iterative partial least squares
NIR	-	Near Infrared
PBS	-	Peso do bagaço seco
PBU	-	Peso do bagaço úmido
PC	-	Principal Component
PCA	-	Principal component analysis
PCR	-	Principal component regression
PLS	-	Partial Least Squares
PROMETHEE	-	Preference ranking organization method for enrichment
RMSEC	-	Root mean square error of calibration
RMSECV	-	Root mean square error of cross validation
RMSEV	-	Root mean square error of validation
SIMCA	-	Soft independent modeling of class analogy
SVD	-	Singular value decomposition

Lista de Tabelas

TABELA 1.5 – VISÃO GERAL DA MODELAGEM 50

TABELA 2.5 – COMPARATIVO DOS ERROS APRESENTADOS PELAS ESTRATÉGIAS DE TRANSFERÊNCIA TESTADOS 52

Lista de Figuras

FIGURA 1.1 – ESQUEMA GLOBAL DE UM FLUXO DE PROGRAMA DE MELHORAMENTO GENÉTICO 3
FIGURA 2.3 – ESPECTROELETROMAGNÉTICO COM REGIÕES DO INFRAVERMELHO DEMARCADAS 8
FIGURA 3.3 – ESQUEMA GERAL DE FORMAÇÃO DOS COEFICIENTES DE REGRESSÃO PELO MÉTODO DE PLS 14
FIGURA 4.4 – ESQUEMA GERAL DA ORIGEM DE UMA AMOSTRA DE CANA-DE-AÇÚCAR 21
FIGURA 5.4 – GRÁFICO DE DIVISÃO DA COMPOSIÇÃO DA CANA-DE-AÇÚCAR COMPARADOS MATERIAIS EM CONDIÇÕES DE ÓTIMO E MATERIAIS DETERIORADOS PELA PERDA DE UMIDADE 22
FIGURA 6.4 – ESPECTOFOTOMETRO NIR ACOPLADO A ESTEIRA PARA ANÁLISE DE MATERIAL DESFIBRADO DE CANA DE AÇÚCAR 23
FIGURA 7.4 – MATERIAL DESFIBRADO DE CANA-DE-AÇÚCAR INSERIDO NA ESTEIRA PARA ANÁLISE 24
FIGURA 8.4 – ESPECTRO DE CANA-DE-AÇÚCAR 25
FIGURA 9.4 – INSTRUMENTAÇÃO UTILIZADA NA *VIA* ÚMIDA 26
FIGURA 10.4 – FLUXO DE TRABALHO PARA AQUISIÇÃO DO BANCO DE DADOS 27
FIGURA 11.4 – FLUXO DE AQUISIÇÃO DE DADOS PARA APLICAÇÃO DE TRANSFERÊNCIA DE CALIBRAÇÃO 37
FIGURA 12.5 – VARIÂNCIA EXPLICADA PARA O CONJUNTO DE DADOS DE ESPECTROS 41
FIGURA 13.5 – SCORES PARA O CONJUNTO DE DADOS DE ESPECTROS 42
FIGURA 14.5 – LOADINGS DA PRIMEIRA COMPONENTE PRINCIPAL PARA O CONJUNTO DE DADOS DE ESPECTROS 43
FIGURA 15.5 – LOADINGS DA SEGUNDA COMPONENTE PRINCIPAL PARA O CONJUNTO DE DADOS DE ESPECTROS 43
FIGURA 16.5 – ESPECTROS DO BANCO DE DADOS DA CLIBRAÇÃO COM AS VÁRIAVEIS INTERESSANTES SELECIONADAS 44
FIGURA 17.5 – ESPECTROS DO BANCO DE DADOS DA CALIBRAÇÃO COM AS VARIÁVEIS INTERESSANTES NORMALIZADOS 45
FIGURA 18.5 – EPECTROS DO BANCO DE DADOS DA CALIBRAÇÃO COM AS VARIÁVEIS INTERESSANTES NORMALIZADOS E DERIVADOS 45
FIGURA 19.5 – ESPECTROS DO BANCO DE DADOS DA CALIBRAÇÃO COM AS VARIÁVEIS INTERESSANTES NORMALIZADOS, DERIVADOS E CENTRADOS NA MÉDIA 46
FIGURA 20.5 – GRÁFICO DE RMSEC E RMSEV PARA O PARÂMETRO BRIX 47
FIGURA 21.5 – ESQUEMÁTICA DOS PONTOS DE ANGULAÇÃO PARA CÁLCULO DA MÁXIMA CURVATURA MODIFICADA 47

RESUMO

COMBINAÇÃO DE INFRAVERMELHO PRÓXIMO E FERRAMENTAS QUIMIOMÉTRICAS PARA ANÁLISE DE MATERIAL DESFIBRADO DE CANA-DE-AÇÚCAR

A empresa Monsanto apresenta em um dos seus segmentos o desenvolvimento de variedades de cana-de-açúcar que apresentam alta produtividade. A capacidade produtiva dessas variedades é aferida por meio de pelo menos 6 parâmetros analíticos: Grau brix, leitura de Pol, pureza, teor de fibra, Pol de caldo e Pol de cana. A determinação destes parâmetros é feita por medidas instrumentais que requerem intensa atenção do operador. Além disso, na Monsanto, o custo por análise é da ordem de R$ 7,50. Logicamente, não somente o custo por análise é o parâmetro principal, mas sim a rapidez na frequência de resultados para atestar se determinada variedade é ou não economicamente viável. Assim, a proposta desse projeto de mestrado profissional é utilizar um espectofotômetro de infravermelho próximo (NIR, *near, infrared*) combinado com ferramentas quimiométricas para propôr modelos de calibração. Esses modelos seriam utilizados para prever de forma mais rápida e com baixo custo os valores dos parâmetros analíticos reportados acima. Na realização desse projeto foram utilizados mais de 7000 espectros de infravermelho próximo que foram obtidos após o desfibramento de cana-de-açúcar. Os espectros foram obtidos e construídos modelos de calibração multivariada (6300 espectros) com a ferramenta quimiométrica PLS (*Partial least squares*). Os modelos foram criados no laboratório de Conchal e validados com 700 espectros que não fizeram parte do banco de dados da calibração e os erros de validação variaram de 0,57 (para grau brix) até 3,55 (para leitura de Pol). Além disso, os modelos de calibração criados foram empregados em duas outras estações experimentais da Monsanto. A primeira delas é localizada na cidade de Araçatuba (SP) e a segunda em Mandaguaçu (PR). Nessa etapa foram testadas 3 estratégias de transferência (*recalibration, model update e slope and intercept*) de calibração e a *slope and intercept* apresentou os menores erros de validação (cerca de 2 vezes mais baixos quando comparados sem transferência). O projeto de mestrado profissional representou até o presente momento uma economia de recursos financeiros da ordem de R$100.000,00 por safra.

ABSTRACT

COMBINATION BETWEEN NIR AND CHEMOMETRICS TOOLS TO ANALYSE FIBERED SAMPLES OF SUGARCANE

The Monsanto Company presents in one of its segments the development of varieties of sugacane which have high productivity. The productive capacity of these clones is measured using at lest six analytical parameters: brix degree, Pol, purity, fiber, juice pol and cane pol. The determination of these parameters is done by instrumental measurements that require intense attention of the operator. Beyond that, in Monsanto, the cost per analysis is the order to R$7.50. Of course, not only the cost per analysis is the main parameter, but how fast the frequency of results to verify if the variety is or isn't economically viable. The proposal of this professional master's degree project is to use the near infrared spectrophotometer (NIR) combined with chomometric tools to propose calibration models. These models would be used to provide faster and low costs of the analytical parameters reported above. In realization of this project were used over 7000 NIR spectra were obtained after defibration sugarcane. Spectra were obtained and constructed multivariate calibration models (6300 spectra) with the chemometric tool PLS (partial least squares). The models were created in the Conchal laboratory and validated with 700 spectra that not were part of the calibration database and the validation errors ranged from 0.57 (for brix degrees) to 3.55 (for pol). Furthmore, the models of calibration were used in two other experimental stations of Monsanto. The first is located in Araçatuba (SP) and the second in Mandaguaçu (PR) At this part were tested 3 strategies of transfer calibration (recalibration, model update and slope and intercept) and slope and intercept had the lowest validation errors (around twice lower when compared with no transfer). The professional master degree represented up to now a saving of financial resources around R$ 100.000,00 per harvest.

SUMÁRIO

1. - INTRODUÇÃO 2
2. - OBJETIVOS 6
3. - REVISÃO BIBLIOGRÁFICA 8
3.1. – ESPECTROFOTOMETRIA DE INFRAVERMELHO PRÓXIMO 8
3.2. ESPECTROFOTOMETRIA DE INFRAVERMELHO PRÓXIMO APLICADA A ANÁLISES DE MATERIAIS AGRÍCOLAS 9
3.3. FERRAMENTAS QUIMIOMÉTRICAS 11
3.3.1. *PCA* – ANÁLISE POR COMPONENTES PRINCIPAIS 12
3.3.2. *PLS* – REGRESSÃO POR MÍNIMOS QUADRADOS - CONSTRUÇÃO DE UM MODELO DE CALIBRAÇÃO 13
3.3.3. RMSEC – *ROOT MEAN SQUARE ERROR OF CALIBRATION* 16
3.3.4. RMSECV – *ROOT MEAN SQUARE ERROR OF CROSS VALIDATION* 16
3.3.5. RMSEV – *ROOT MEAN SQUARE ERROR OF VALIDATION* 17
3.4. ESTRATÉGIAS DE TRANSFERÊNCIA DE CALIBRAÇÃO 17
4. – MATERIAIS E MÉTODOS 21
4.1. – AMOSTRAS 21
4.1.1. – ANÁLISE PELO MÉTODO NIR 23
4.1.2. – ANÁLISE PELO MÉTODO *VIA* ÚMIDA 25
4.2. – AQUISIÇÃO DO BANCO DE DADOS 27
4.3. – PRÉ PROCESSAMENTO DOS DADOS 28
4.3.1. – NORMALIZAÇÃO INSERIR REFERENCIA 28
4.3.2. – PRIMEIRA DERIVADA 28
4.3.3. – CENTRAR NA MÉDIA 30
4.3.4. – AUTOESCALAMENTO 31
4.4. – ANÁLISE EXPLORATÓRIA 31
4.4.1. – PCA – ANÁLISE DE COMPONENTES PRINCIPAIS 31
4.5. – CONSTRUÇÃO DO MODELO DE CALIBRAÇÃO 32
4.5.1. – PLS – REGRESSÃO POR MÍNIMOS QUADRADOS 32
4.5.2. – PREVISÃO DE AMOSTRAS DESCONHECIDAS 35
4.5.3. – VALIDAÇÃO EXTERNA 35
4.6. – ESTRATÉGIAS DE TRANSFERÊNCIA DE CALIBRAÇÃO 36
4.6.1. – AQUISIÇÃO DE DADOS 36
4.6.2. – RECALIBRAÇÃO 37
4.6.3. – *MODEL UPDATE* 37
4.6.4. – CORREÇÃO POR *SLOPE AND INTERCEPT* 38
5. – RESULTADOS E DISCUSSÕES 40
5.1. – ÁNALISE EXPLORATÓRIA USANDO PCA 40
5.2. – PRÉ PROCESSAMENTO DOS DADOS 43
5.3. – CONSTRUÇÃO DO MODELO DE CALIBRAÇÃO 45
5.4. – RESULTADOS GERAIS DA CONSTRUÇÃO DO MODELO 49
5.5. – TRANSFERÊNCIA DE CALIBRAÇÃO 50
5.5.1. RECALIBRAÇÃO 50
5.5.2. MODEL UPDATE 50
5.5.3. CORREÇÃO POR AJUSTE DE *SLOPE AND INTERCEPT* 51
5.5.4. COMPARAÇÃO DOS ERROS APRESENTADOS 51
6. – CONCLUSÕES 54
7. – REFERÊNCIAS BIBLIOGRÁFICAS 56

INTRODUÇÃO

1. - INTRODUÇÃO

O Brasil é atualmente o maior produtor mundial de cana-de-açúcar, pois apresenta fatores edafoclimáticos favoráveis. A cana-de-açúcar possui como principal característica mecanismos fisiológicos altamente evoluídos para produção e acúmulo de sacarose. Considerando que a sacarose é a principal matéria prima energética, existe um interesse industrial e financeiro em grandes e produtivos cultivos.

A cana-de-açúcar pode ser dividida em três grupos: (i) A água que devido a fatores fisiológicos de crescimento e desenvolvimento é o maior constuinte da cana, (ii) Fibra, formada principalmente por lignina, celulose e hemicelulose funcionais para o crescimento da planta e (iii) Sólidos solúveis, onde encontra-se a sacarose, outros açúcares como frutose e glucose e demais impurezas. Os constituintes da cana representam, no geral, cerca de 68% de água, 19% de sólidos solúveis e de 13% de fibra. Esta concentração dos constituintes pode variar conforme o material plantado. Diferentes variedades de cana-de-açúcar apresentam distintas concentrações de cada um dos três fatores constituintes, bem como diferentes épocas de maturação alcançando máximos de produção em épocas diferentes de uma safra [1].

A necessidade de obtenção de altas quantidades de matéria prima energética impulsiona programas de melhoramento genético que tem como principal objetivo o desenvolvimento de materiais com maior capacidade produtora. Neste passo, o programa de melhoramento genético da Monsanto visa a obtenção de materiais com alta capacidade produtiva de sacarose em menores áreas.

Um programa de melhoramento genético de cana-de-açúcar possui peculiaridades que vão desde o cruzamento nas primeiras fases até o lançamento de uma variedade comercial no último estágio. Dentre estas especificidades, chamam a atenção a quantidade de indivíduos nas primeiras fases, o tempo de crescimento dos materiais para avaliação e a necessidade de examinar o comportamento dos materiais em diferentes tipos de solo e condições climáticas.

As primeiras fases de um programa de melhoramento genético apresentam um número elevado de indivíduos ocupando pequenas áreas e com poucas repetições. Ao passo que os materiais vão evoluindo para fases avançadas,

em uma disputa produtiva, diminui-se a quantidade de individuos e aumenta-se as repetições e áreas de ocupação.

Para obter informações suficientes do nível de produção e assim decidir se determinado material avança ou não de fase no melhoramento genético, são feitas análises biométricas em campo e tecnológicas em laboratório. Na Figura 1.1 pode-se verificar o afunilamento de um programa de melhoramento genético onde o tamanho da população (número de indivíduos) diminui com o passar das fases e aumenta as áreas de ocupação e repetições dos materiais selecionados no programa.

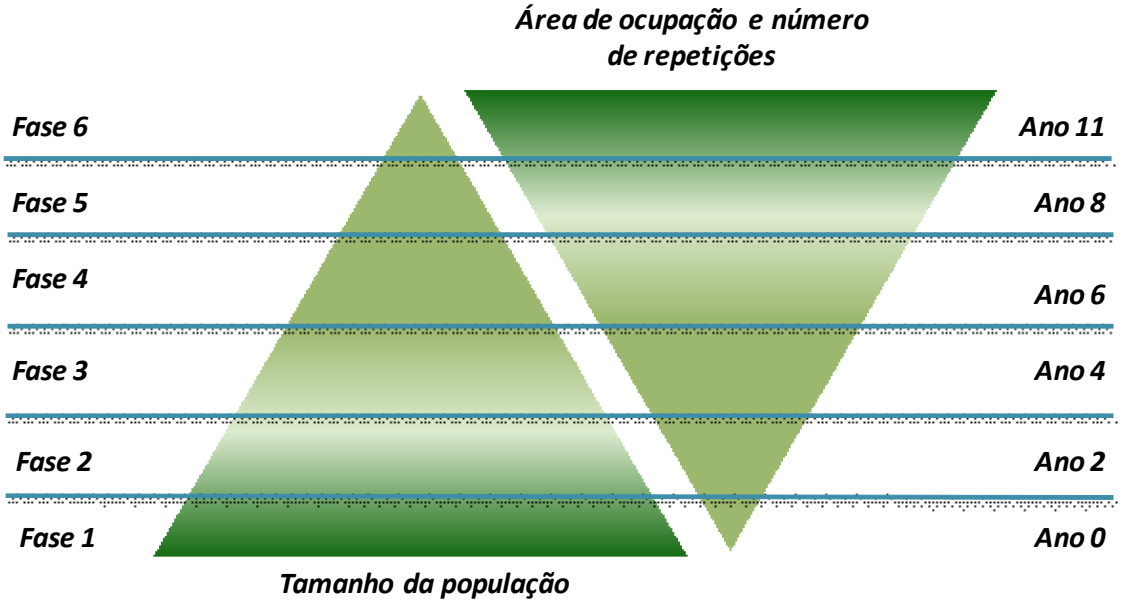

FIGURA 1.1: Esquema global de um fluxo de programa de melhoramento genético.

As análises realizadas em laboratório determinam os níveis gerais de sacarose e fibra de cada variedade utilizando seis parâmetros:

(i) grau brix, que quantifica o total de sólidos solúveis presentes no caldo da cana-de-açúcar,

(ii) leitura de pol, que quantifica a sacarose aparente do material,

(iii) pureza, a porcentagem de sacarose presente nos sólidos solúveis,

(iv) fibra, quantifica qual a porcentagem total de fibra no material,

(v) pol de caldo, parâmetro que determina a quantidade total de sacarose presente no caldo que poderá ser extraído e

(vi) pol de cana, quantidade de sacarose estimada a ser extraída do material uma vez que a porcentagem de fibra interfere na eficiência da extração.

A quantificação destes parâmetros pode ser realizada através do método convencional de *via* úmida. Este método foi padronizado pelo CONSECANA (Conselho dos produtores de cana-de-açúcar, açúcar e álcool do estado de São Paulo) onde equipamentos, reagentes e processos analíticos são descritos e padronizados [2]. Este método é eficiente e confiável, porém demanda tempo e custos elevados, impossibilitando o atendimento total do programa de melhoramento da Monsanto. Atualmente a demanda por este tipo de quantificação é da ordem de 20000 análises/safra. Além disto, por utilizar reagentes químicos nocivos a natureza, o método de análises por *via* úmida configura-se como um procedimento que apresenta deficiências em alguns pontos relacionados com a química verde.

Uma segunda forma de obter os seis parâmetros necessários é utilizando espectrofotometria de infravermelho próximo (*NIR – near infrared*). Este método analítico apresenta menores custos e tempo agregados além de demandar reduzido número de operadores envolvidos, diminuíndo também, o erro humano intrínseco a qualquer análise. Este método analítico necessita de um método de referência para a construção do modelo de calibração e retroalimentação do mesmo.

OBJETIVOS

2. - OBJETIVOS

Esta dissertação de mestrado profissional tem como objetivo otimizar o processo de análises quantitativas de cana-de-açúcar implementando tecnologia de infravermelho próximo (NIR) como principal método analítico. Os objetivos específicos são: (i) Construção de um modelo de calibração eficiente calculado através de respostas obtidas pelo método convencional de *via* úmida e espectros de NIR provenientes das mesmas amostras e (ii) transferir a calibração calculada a partir dos dados do laboratório mestre (Conchal – SP) para dois outros equipamentos escravos de estações experimentais diferentes da Monsanto (Araçatuba/SP e Mandaguaçu/PR).

REVISÃO BIBLIOGRÁFICA

3. - REVISÃO BIBLIOGRÁFICA

3.1. – Espectrofotometria de Infravermelho próximo

No espectro eletromagnético, o infravermelho (IR, *infrared*) faz limites com a região visível e região de microondas. A região espectral do infravermelho compreende a radiação com número de onda que varia de 12.800 a 10 cm^{-1}. Na Figura 2.3 pode-se perceber qual a divisão do espectro eletromagnético com demarcações nas faixas do Infravermelho próximo [3].

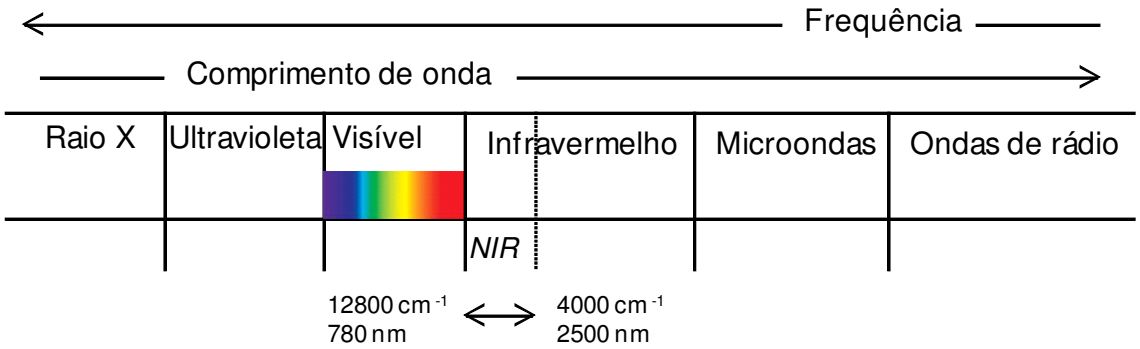

FIGURA 2.3**:** Espectroeletromagnético com regiões do infravermelho e infravermelho próximo demarcadas.

O espectro IR é usualmente subdividido em três regiões, denominadas IR-próximo, IR-médio e IR-distante, de acordo com o tipo de aplicação e instrumentação. O espectro NIR recebe este nome devido à sua proximidade do espectro visível. O infravermelho próximo corresponde à região onde pode-se observar bandas correspondentes as harmônicas ou modos de combinação de frequências fundamentais. As ligações envolvidas nesses modos de vibração normalmente são C-H, N-H, O-H e S-H [4].

Para que ocorra uma absorção da radiação infravermelha, é necessário que haja uma variação no momento de dipolo elétrico das moléculas que é resultante de seu movimento vibracional ou rotacional [4].

Este momento dipolo é determinado pela magnitude da diferença entre as cargas e a distância dos dois centros de carga.

A absorção da radiação infravermelha por um sistema é proveniente do momento dipolo sendo a frequência da radiação absorvida idêntica à da oscilação

do dipolo. Assim, quando há radiação eletromagnética incidente sobre uma componente com frequência correspondente a uma transição entre dois níveis vibracionais, pode-se dizer que há um espectro de absorção no infravermelho. Por esta definição, os espectros de absorção, emissão e reflexão no infravermelho de espécies moleculares podem ser racionalizados assumindo-se que todos originam de numerosas variações de energia produzida por transições de moléculas de um estado de energia vibracional ou rotacional para outro [3]. Diferentes moléculas apresentam diferentes frequências de transição moleculares, sejam elas vibracionais ou rotacionais.

As bandas espectrais no infravermelho próximo normalmente são largas e frequentemente aparecem sobrepostas. Raramente são encontradas bandas limpas que poderiam permitir uma correlação simples com uma concentração para um parâmetro de interesse. Para que haja esta correlação e interpretação do espectro, mais comumente são usadas técnicas de calibração multivariada em combinação com ferramentas quimiométricas [4].

3.2. Espectrofotometria de Infravermelho próximo aplicada a análises de materiais agrícolas

A utilização de métodos espectroscópicos apresenta várias vantagens, tais como a pequena quantidade de amostra necessária para o método analítico, onde muitas vezes este processo é não destrutivo e com alta frequência de amostragem.

Uma das principais utilizações de infravermelho próximo é a aplicação para análises quantitativas de materiais industriais e agrícolas. No campo agrícola, existem muitas demandas analíticas para o NIR, dependendo da cultura a qual se aplica. Dentro dos cultivos de cana-de-açúcar, pode-se destacar alguns trabalhos acadêmicos.

No ano de 2009, um grupo de pesquisadores australianos implementaram a tecnologia de infravermelho próximo em conjunto com ferramentas quimiométricas para análises de bagaço e açúcar bruto para melhoria de processo em indústrias de açúcar. O principal objetivo deste trabalho foi inicialmente mensurar teores de fibra em tempo real para programa de pagamento de fornecedor. Entretanto, após essa aplicação, desenvolveu-se modelos para análise de cinzas,

matéria seca pol, brix no caldo da cana e ainda açúcar final na cana comercial. Além disto, os dados destas calibrações foram ainda usados para auditar laboratórios, propor índices de qualidade, auxiliando na busca por soluções específicas para áreas com problemas de produtividade. Segundo os pesquisadores quase 30 sistemas já foram instalados em todo o mundo onde o processo de utilização contínua sempre em desenvolvimento pelos usuários [5].

No ano de 2007, pesquisadores utilizaram quimiometria para aplicação de NIR em um projeto para prever o desenvolvimento clonal de cana-de-açúcar. No trabalho utilizaram ferramentas importantes como PCA (Análise por componentes principais), PLS (Regressão por mínimos quadrados parciais) integradas a análises convencionais por cromatografia gasosa como método de referência. O principal objetivo deste trabalho era utilizar a espectrofotometria como alternativa para análises de forma confiável, rápida e não destrutiva como principal método analítico. Os autores utilizaram 65 clones de cana-de-açúcar de regiões geográficas muito diferentes. Os pesquisadores concluíram que uma análise por PCA pode mostrar de forma clara uma tendência evidente de similaridade para amostras com bom desempenho bem discriminadas das amostras com baixo desempenho. Além disso, evidenciaram que lignina parece estar ligada a clones com melhores desempenhos. Concluíram que a construção de um modelo através de PLS permite previsão satisfatória das classificações tradicionais de desempenho (análises quantitativas) e ainda que utilizando da ferramenta PROMETHEE (método de organização para avaliação de enriquecimento) demostraram que as amostras poderiam ser classificadas a partir dos espectros de NIR com a ordem de classificação estando de acordo com as classificações tradicionais [6].

Em 2010, pesquisadores argentinos buscavam o desenvolvimento de um modelo quimiométrico eficiente testando métodos de seleção de variáveis para determinação de brix diretamente no caldo de cana-de-açúcar. O principal objetivo deste trabalho foi encontrar um método que otimizasse a seleção de variáveis buscando uma forma de manter apenas os comprimentos de onda com informações importantes do ponto de vista químico resultando em melhores resultados. Os autores testaram dois métodos: (i) baseada nos coeficientes de regressão PLS e (ii) pesquisas de erros realizadas através de PLS de intervalos (iPLS) que incluíam a utilização de algorítimo genético nos quais, segundo os pesquisadores, apresentavam resposta mais adequada para a seleção de variáveis na calibração

multivariada. Os pesquisadores consideraram ainda que antes da seleção das variáveis, deve-se sempre discutir o número de variáveis latentes, a quantidade de comprimentos de onda selecionados, sendo a velocidade de operação do programa considerada a mais importante delas e a precisão da previsão de novas amostras. Concluiu-se também que para este tipo de trabalho, uma vez que fosse criado um algorítimo para seleção de variáveis, não há necessidade de retro alimentação, uma vez que as regiões selecionadas permanecem constantes durante a predição de todas as amostras futuras [7].

Ainda no ano de 2010, um grupo de pesquisadores brasileiros perceberam a necessidade de efetuar análises de matéria orgânica e fração de argila em solos e teores foliares de silício e nitrogênio em cana-de-açúcar com um método rápido e confiável. Assim, buscaram então efetuar estas análises com infravermelho próximo com o intuito de terem respostas rápidas, diminuição de mão-de-obra e reduzindo a utilização de reagentes químicos. Segundo os pesquisadores, aplicando-se a tecnologia de NIR para as quantificações de teores de argila e de matéria orgânica, os resultados não diferiam daqueles obtidos pelos métodos convencionais. Os autores obtiveram coeficientes de correlação da ordem de 0,77 e 0,70 para teor de argila e matéria orgânica, respectivamente. Estes valores indicam que o sistema de análises por infravermelho próximo pode substituir com vantagens os métodos convencionais sem comprometer a precisão e exatidão dos resultados. Os pesquisadores ainda concluem que a utilização de NIR possui grande potencial de uso nos laboratórios de solos e plantas para atender a demanda e as exigências do mercado atual [8].

3.3. Ferramentas quimiométricas

Quimiometria é o campo da química que utiliza ferramentas estatistícas e matemáticas para o planejamento e otimização das condições experimentais e para extração de informação química relevante de dados químicos multivariados.

A espectroscopia NIR e a Quimiometria convivem em simbiose. Uma vez que o espectro NIR possui bandas largas e sobrepostas, a quimiometria torna-se uma ferramenta valiosa para retirar as informações contidas nos espectros NIR permitindo a identificação e a quantificação de diversos parâmetros em diferentes matrizes [4].

A importante aplicação da quimiometria nos atuais laboratórios modernos de química evoluiu com a capacidade dos instrumentos analíticos de produzirem conjuntos de dados, cada vez maiores e mais complexos juntamente com a evolução dos computadores e da microeletrônica [9].

Grande parte dos métodos tecnológicos de análise química utilizando instrumentação moderna se dá apenas com o trabalho em conjunto com a quimiometria.

3.3.1. *PCA* – Análise por componentes principais

A PCA é uma das mais importantes ferramentas quimiométricas para uma análise exploratória multivariada de um banco de dados. Das principais relevâncias da exploração do banco de dados através de uma análise PCA, pode-se destacar a fácil visualização de amostras com perfis de *outliers*, relação entre variáveis medidas, relação entre agrupamento de amostras e determinação de partes com maior relevância, do ponto de vista químico de um grupo de dados. Além disto, derivados dos resultados de uma análise de PCA, atribui-se uma base onde se fundamentam a maioria dos outros métodos multivariados de análise de dados como a modelagem independente por analogia de classes (SIMCA, *soft independent modeling of class analogy*) e de calibração, como a regressão por componentes principais (PCR, *principal component regression*) e PLS [9,10].

De maneira geral, pode-se estabelecer que a PCA constitui a base para uma análise exploratória dos dados. Seus principais objetivos são visualizar a estrutura dos dados, encontrar similaridades entre amostras, detectar amostras anômalas e reduzir a dimensionalidade dos dados. Sua principal proposta é expressar informações significativas contidas nas variáveis originais em um pequeno grupo de novas variáveis, as então chamadas componentes principais [11]

A principal característica deste novo conjunto é a ortogonalidade, porém o mesmo é facilmente reconstruído a partir da combinação linear das variáveis originais. A diferença do novo conjunto de variáveis, gerado pela PCA, é a diminuição da dimensionalidade dos dados sem perda da informação química importante. Isto acontece porque o novo conjunto de variáveis concentra a maior parte da informação (variância) em poucas variáveis [10].

Existem disponíveis uma variedade de algorítimos que podem ser usados para calcular as componentes principais de um conjunto de dados. A decomposição por valores singulares (SVD, *singular value decomposition*) é um algorítimo rotineiramente empregado [12].

O cálculo de uma PCA consiste em fatorar uma matriz **X** (normalmente composta por espectros) de modo que $\mathbf{X}=\mathbf{TL}^T+\mathbf{E}$, onde tem-se **L** como a matriz de *loadings,* que trará consigo as coordenadas das variáveis; **T** como a matriz de *scores*, carregando as informações referentes as coordenadas das amostras; **E** sendo a matriz dos resíduos e T sobrescrito sendo o operador da transposição de matriz. A primeira componente principal é igual a $PC1=t_1l_1^T$, que é a melhor aproximação do posto 1 para **X** e corresponde a direção de maior variância no espaço multivariado. $\mathbf{E}_1=\mathbf{X}- t_1l_1^T$ é o resíduo de **X**, descontado PC1. A segunda componente principal é $PC2=t_2l_2^T$ que é a melhor aproximação de posto 1 para E_1 e corresponde a direção de maior variância no espaço multidimensional não modelado por PC1, ou seja, ortogonal a ela. $\mathbf{E}_2=\mathbf{E}_1-t_2l_2^T$ é o resíduo deixado por PC1 e PC2. As componentes subsequentes modelam sempre a direção de maior variância no espaço miltidimensional não modelado pelas PCs anteriores e são sempre ortogonais a elas [4,9,10].

Resumidamente, a primeira PC sempre explica a maior variância dos dados, a segunda PC é ortogonal à primeira e explica a variância dos dados que a primeira PC não explicou, a terceira componente principal, também ortogonal as outras, explica o máximo de variância que a primeira e a segunda PC's não explicaram e assim sucessivamente.

3.3.2. *PLS* – Regressão por mínimos quadrados - Construção de um modelo de calibração

Do ponto de vista da quimiometria, calibração refere-se a construção de um modelo que relaciona as respostas obtidas a partir de determinadas amostras e parâmetros conhecidos destas mesmas amostras. No geral, o principal objetivo do modelo construído é realizar previsões de concentrações a partir de análises de novas e desconhecidas amostras. Este conjunto de dados necessário para a calibração pode ser definido como amostras de referência (conjunto de calibração) para concentrações ou propriedades de interesse. As novas amostras, de onde

espera-se obter respostas previstas devem apresentar as mesmas características das amostras de referência.

Existem diferentes técnicas para efetuar esta correlação entre o conjunto de dados de espectros e os parâmetros de interesses destas amostras resultando em um modelo de calibração. Um dos métodos mais conhecidos e difundidos em quimiometria hoje em dia é o PLS [13].

De modo geral, pode-se afirmar que o PLS é um método de calibração multivariada que utiliza a técnica de análise de componentes principais para reduzir a dimensão do conjunto de dados para correlação dos espectros e as propriedades de interesse. O método de PLS vem ganhando muito espaço nos últimos anos devido aos seus beneficios e alta aplicabilidade. A Figura 3.3 mostra como é um esquema geral da correlação entre os espectros e os parâmetros de interesse para a formação dos coeficientes de regressão através do método de PLS.

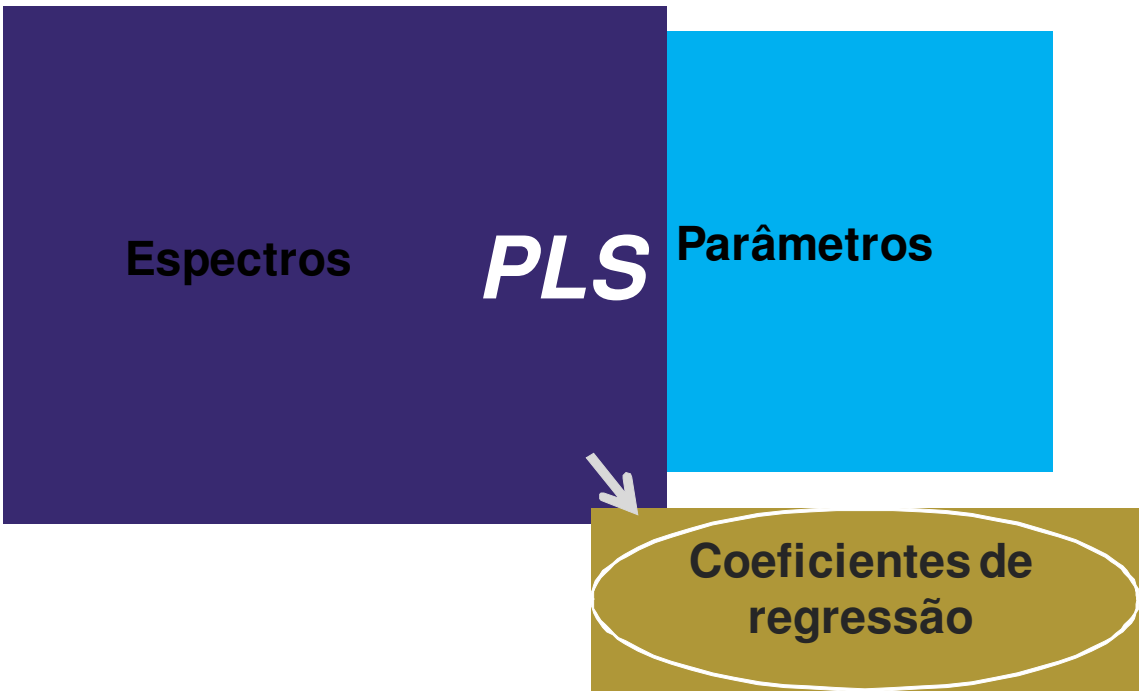

FIGURA 3.3**:** Esquema geral da formação dos coeficientes de regressão pelo método de PLS

O grande responsável pelo desenvolvimento do PLS foi Herman Wold que, por volta do de 1975 trabalhava com dados complicados da área de econometria [14].

Diferentemente a PCA, onde buscava-se uma correlação e assimilaridade entre os dados separadamente, na análise de PLS, busca-se um modelo onde as matrizes de resíduos das matrizes **X** e **Y** sejam as menores possíveis e, ao mesmo tempo consiga-se uma correlação linear entre as duas.

Em detalhes, podemos dizer que o PLS calcula as componentes principais para ambas as matrizes (**X** e **Y**) e para melhorar a correlação linear entre as duas matrizes, as componentes principais podem sofrer pequenas modificações em seus ângulos e estas modificações fazem com que estas componentes deixem de ser ortagonais passando a se chamar variáveis latentes.

Para realização de uma regressão multivariada por PLS, existem uma série de algorítimos que acabam chegando a resultados parecidos quando não aos mesmos. Um destes algorítimos, talvez o mais conhecido deles, é o algorítimo de NIPALS (*Nonlinear interative partial least squares)*. Este tipo de algorítimo busca uma relação linear entre as variáveis intrumentais (**X**) e as variáveis de interesse (**Y**) [15].

O grande diferencial da construção de um modelo de calibração pelo método de PLS é que as matrizes **X** e **Y** são decompostas simultaneamente em uma soma de *h* variáveis latentes, como na seguinte equação:

$$\mathbf{X = T\,P' + E = S\,t_h\,p'_h + E\quad Y = U\,Q' + F = S\,u_h\,q'_h + F}$$

Onde **T** e **P** são as matrizes de *scores* e loadings para **X**, respectivamente; **U** e **Q** são as respectivas matrizes de *scores* e *loadings* para a matriz **Y**. **E** e **F** são os resíduos. Uma correlação entre os blocos **X** e **Y** é simplesmente uma relação linear obtida pelo coeficiente de regressão linear, tal como:

$$\mathbf{U_h = b_h\,t_h}$$

Assumindo *h* de variáveis latentes, sendo que os valores de b_h são agrupados na matriz diagonal **B**, que contém os coeficientes de regressão entre a matriz de *scores* **U** de **Y** e a matriz de *scores* **T** de **X.** A relação linear melhor encontrada entre os *scores* dos dois blocos é calculada através de ajustes das variáveis latentes dos blocos **X** e **Y** [15].

3.3.3. RMSEC – *Root mean square error of calibration*

Os modelos de calibração construídos são avaliados usando o erro quadrático médio de calibração (RMSEC, *root mean square error of calibration*). Este erro é calculado quando o modelo prevê resultados de amostras que foram usadas na calibração. O RMSEC é calculado segundo a equação:

$$\sqrt{\frac{\sum_{i=1}^{Ncal}(y_i - \hat{y}_i)^2}{Ncal - k - 1}}$$

Onde y_i são os valores reais; $\hat{y}_i$ são os valores previstos; N_{cal} é o número de amostras da matriz de dados e; k é o número de variáveis latentes [16].

3.3.4. RMSECV – *Root mean square error of cross validation*

Para efetuar a validação cruzada de um modelo de calibração, uma amostra é retirada do banco de dados e calcula-se um modelo de calibração com as amostras restantes. Logo após, esta amostra retirada é predita com o modelo construído da qual ela não fez parte empregando um número crescente de variáveis latentes. Após esta predição, esta amostra volta ao banco de dados e uma segunda amostra é retirada e repete-se o procedimento. A esta técnica da-se o nome de *leave-one-out*, processo que se faz retirando apenas uma amostra. Existem ainda técnicas que na tentativa de otimizar o processo retiram duas ou mais amostras por vez.

Após este procedimento ser feito para todas as amostras, é calculado um erro médio para cada número de variáveis latentes empregados.

O procedimento de validação externa de um modelo de calibração é muito importante visto que o número de variaveis latentes a partir do qual não existe variação apreciável no valor de RMSECV é utilizado para predizer qual a melhor quantidade de variáveis latentes deve ser utilizado no modelo.

O cálculo do erro médio da validação cruzada é obtido segundo a equação:

$$\sqrt{\frac{\sum_{i=1}^{Ncal}(y_i - \hat{y}_i)^2}{Ncv}}$$

Onde y_i são os valores reais; $\hat{y}_i$ são os valores previstos; N_{cv} é o número de amostras usadas para a validação cruzada [16].

3.3.5. RMSEV *– Root mean square error of validation*

Após construção do modelo de calibração e definidos as melhores condições para valores de variáveis latentes, aplica-se o processo de validação externa do modelo. Nesta fase, um conjunto de dados de espectros que não participaram da calibração são previstos pelo modelo criado. Estas mesmas amostras devem ter seus parâmetros de interesse quantificados pelos processos convencionais. Logo após é calculado um erro quadrático médio de validação segundo a equação:

$$\sqrt{\frac{\sum_{i=1}^{Nval}(y_i - \hat{y}_i)^2}{Nv}}$$

Onde y_i são os valores reais; $\hat{y}_i$ são os valores previstos; N_v é o número de amostras de validação [16].

Os valores previstos por esta etapa devem apresentar-se concordantes com os valores reais.

3.4. Estratégias de transferência de calibração

Implementar um método de transferência de calibração implica em utilizar os coeficientes gerados por uma calibração calculada através de espectros captados por um equipamento 1 (mestre) em amostras desconhecidas captadas por um equipamento 2 (escravo) e obter respostas confiáveis.

Equipamentos analíticos de mesmos princípios instrumentais podem fornecer diferentes respostas para mesmas amostras por diversas razões.

De modo geral, pode-se dizer que estas diferenças são influenciadas basicamente por mudanças na constituição química e física da amostra, mudanças no ambiente do equipamento variando temperatura e umidade provocando deslocamentos de bandas de absorção [17].

Quando se analisa uma amostra com dois equipamentos, sejam eles de mesmo fabricante e mesmo modelo, pode-se destacar dois problemas principais:

deslocamento no comprimento de onda e mudanças nas respostas espectrais medidas. Se os equipamentos forem diferentes, a proporção das diferenças apresentadas nos dados serão evidentemente maiores [17].

Assim, pode-se dizer que se uma mesma amostra for analisada por diferentes equipamentos, os espectros apresentados terão o mesmo perfil, porém intensidades diferentes. Esta divergência na intensidade resultará em respostas diferentes e, considerando que elas são de mesmas amostras, os resultados deveriam ser similares [18,19].

A transferência de calibração impulsiona modificações estratégicas, sejam elas espectrais ou nos parâmetros previstos, ajustando os dados tornando os resultados previstos confiáveis.

Assim como na construção do modelo de calibração, para se aplicar um método de transferência, o primeiro passo é a aquisição de dados. Nestes casos necessita-se adquirir espectros a partir do equipamento mestre, equipamento escravo e os devidos parâmetros pelos métodos de referência das mesmas amostras. Após aquisição de um banco de dados representativo tem-se informações suficientes para testar métodos de transferência. No estudo de caso para se avaliar qual o melhor método de transferência para determinado equipamento, deve considerar diversos fatores dentre eles os mais agravantes são: (i) erro, assim como um modelo de calibração, a transferência imbute um erro que deve ser considerado e analisado, (ii) necessidade de retroalimentação, aplicados principalmente em casos onde as amostras podem sofrer evolução, assim como a calibração, a transferência necessita ser periodicamente retroalimentada. Neste ponto deve-se considerar qual a necessidade de retroalimentação implicada pelo método, uma vez que, na maioria dos casos em que se aplica transferência, os equipamentos ficam em laboratórios distantes implicando em logística de transporte de amostras muitas vezes degradativo.

Existem diferentes estudos que implicam em soluções de tratamentos aplicados na construção do modelo de calibração, nos espectros de um equipamento escravo ou ainda nos resultados gerados para que um valor registrado através de um equipamento escravo se torne parecido com o que teria sido medido no equipamento mestre.

Das técnicas existentes, pode-se destacar a padronização direta na qual relaciona-se os espectros provindos do equipamento escravo com os espectros

provindos do equipamento mestre padronizando-os e aproximando o resultando em valores de respostas com erros baixos e aceitáveis. Este tipo de padronização pode ser feito diretamente, ou seja, busca-se correlação espectral em todo o espectro ou ainda padronização direta por partes na qual é feita uma varredura em ambos os espectros limitando a correlação a pequenas regiões [17].

Existem ainda métodos que buscam um modelo global robusto suficiente para prever amostras desconhecidas analisadas por todos os equipamentos representados no banco de dados. Este tipo de modelagem tem pecularidades nas quais um estudo de caso deve-se atentar. Utilizando desta estratégia de transferência, é necessário registrar um conjunto de dados de calibração suficientemente amplo para englobar os efeitos de todas as fontes consideradas. Vale ressaltar que um modelo local tenderia a gerar melhores predições comparadas a um global, porém o modelo global poderia ser mais robusto no sentido de que as suas predições continuariam confiáveis em uma gama mais ampla de situações [17].

Outra linha de aplicação quanto as transferências de calibração pode ser formada utilizando ajustes nos valores previstos através de uma transformação linear e univariada. Assim, as amostras de transferência têm suas propriedades medidas e preditas em ambos os equipamentos utilizando o modelo de calibração desenvolvido no equipamento mestre. Os resultados gerados são usados para ajustar a equação de regressão cujos coeficientes linear e angular irão corrigir os deslocamentos da linha de base e de inclinação. Este tipo de transferência de calibração é melhor aplicado em casos onde os equipamentos mestre e escravos são idênticos, pois suas diferenças não são complexas [17].

MATERIAIS E MÉTODOS

4. – MATERIAIS E MÉTODOS

Toda a parte experimental do presente trabalho foi realizada no laboratório de pesquisa e desenvolvimento da Monsanto do Brasil, na estações experimentais CanaVialis, localizadas nas cidades de Conchal e Araçatuba ambas no estado de São Paulo e Mandaguaçu no estado do Paraná.

4.1. – Amostras

As amostras utilizadas neste trabalho foram materiais provenientes do programa de melhoramento genético de cana-de-açúcar da Monsanto. Todos os dados referentes a identificação de padrões e variedades foram codificados.

Uma amostra representativa corresponde a 10 canas situadas no centro da parcela enviada para o laboratório que irá representar o nível de produção de matéria-prima referente ao material plantado naquela parcela inteira [2]. Na Figura 4.4 pode-se visualizar como é o fluxo de origem de uma amostra de cana-de-açúcar para análises químicas e tecnológicas.

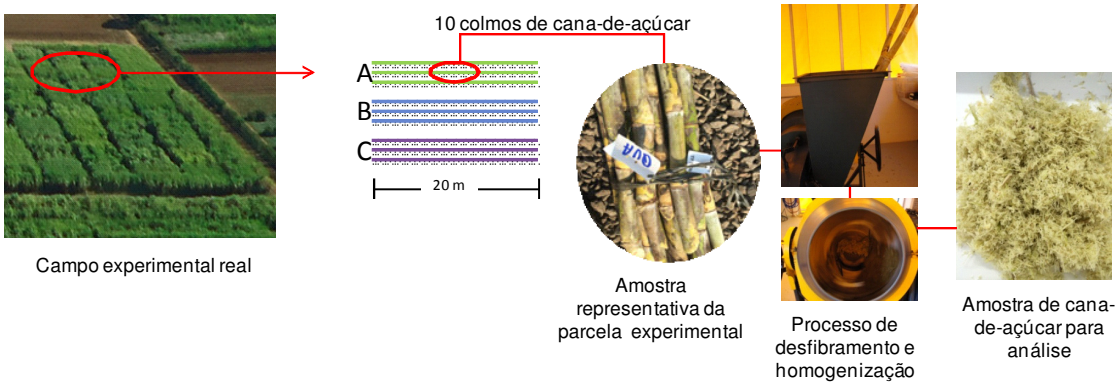

FIGURA 4.4**:** Esquema geral da origem de uma amostra de cana-de-açúcar para análise tecnológica.

No exemplo, cada variedade (A, B e C) é plantada em três linhas (parcelas) com 20 metros lineares cada. A representação da amostra são dez canas localizadas no centro da linha do meio. Estas dez canas são desfibradas e homogenizadas e obtem-se uma amostra para análise. Isto é feito para cada parcela do campo experimental.

Diferentes fases de um programa de melhoramento genético demandam diferentes repetições dos mesmos materiais (variedades experimentais

ou padrões) no mesmo campo. Se determinada fase demanda que um material seja, por exemplo, testado em cinco repetições, cada uma das cinco parcelas terá três ruas e serão formandas cinco amostras.

Amostras de cana-de-açúcar apresentam uma alta complexidade analítica por apresentar uma rápida deteriorização ocasionada pela perda de umidade. A perda de água fará com que os demais componentes sejam porcentualmente concentrados, mascarando os resultados reais das análises quantitativas.

Das demais pecularidades presentes na amostra que podem alterar a qualidade do material prejudicando a análise química, destacam-se a presença de impurezas provenientes do manejo da cana-de-açúcar em campo. A presença de impurezas no material (poeira, palha, terra, folhas, etc) refletem dificuldades no processo analítico e posteriormente nos processos industriais de extração de matéria prima. A Figura 5.4 compreende a divisão dos fatores de um material em uma situação ideal e posteriormente a divisão destes fatores para um mesmo material que sofreu perda de umidade.

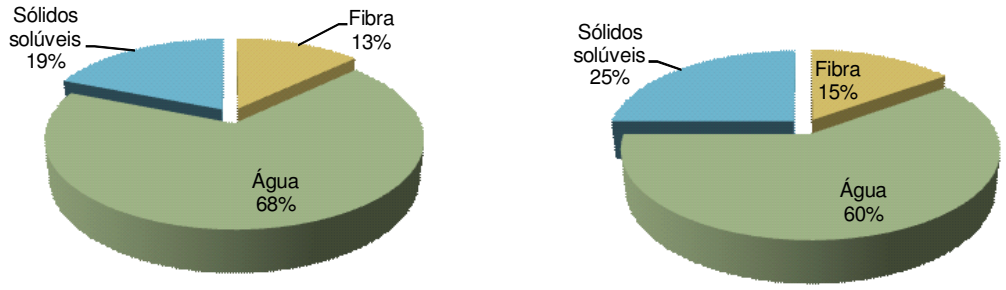

FIGURA 5.4**:** Gráfico de divisão da composição da cana-de-açúcar comparando materiais em condições ideias e materiais deteriorados pela perda de umidade.

As condições ideais de amostragem condizem com análise de no máximo dois dias após coleta em campo e com baixa manifestação de impurezas, ou seja, uma amostra nova e limpa. Como visto na Figura 5.4, esta amostra apresenta um valor real de porcentagem de umidade, fibra e sólidos solúveis. Se esta mesma amostra sofrer perda de umidade e apresentar impurezas do campo, os valores analisados serão mascarados, ou seja, não condizentes com as concentrações reais.

Esta complexibilidade analítica exige que a amostra seja processada e analisada com rapidez gerando confiabilidade nos resultados apresentados [2]. Com a demanda laboratorial alta, necessita-se que estas amostras sejam coletadas em tempo hábil e com qualidade. A problemática para esta presteza analítica é que nem sempre os processos de laboratório atendem esta demanda uma vez que, o fluxo analítico é mais lento que a quantidade de material a ser amostrado.

4.1.1.– Análise pelo método NIR

Para realizar as análises por NIR, foi usado um espectrofotômetro monocromador XDS *online* da marca FOSS. Este equipamento apresenta uma resolução espectral da ordem de 0,5 nm. O software utilizado para aquisição dos espectros foi o Vision da marca FOSS.

Na Figura 6.4, pode-se observar um exemplo de equipamento NIR acoplado com esteira para análise de material desfibrado de cana-de-açúcar.

FIGURA 6.4**:** Espectrofotômetro NIR acoplado a esteira para análise de material desfibrado de cana-de-açúcar.

No procedimento para esta análise, utilizou-se aproximadamente 2000 gramas de cana desfibrada acondicionada na parte inicial da esteira. A esteira corria com uma velocidade controlada e levava o material até a área de análise do espectrofotômetro.

Quando o material era posicionado abaixo do compartimento de luz, acionava-se a leitura no computador e a amostra era analisada 32 vezes pelo

equipamento. Logo após as leituras da amostra, o equipamento realizava a leitura da cerâmica de referência também por 32 vezes. Cerca de um minuto depois, o software apresentava um espectro que era resultado da média das 32 leituras. Na Figura 7.4 pode-se notar o inserção da amostra no começo da esteira e a leitura do espectro no final da mesma.

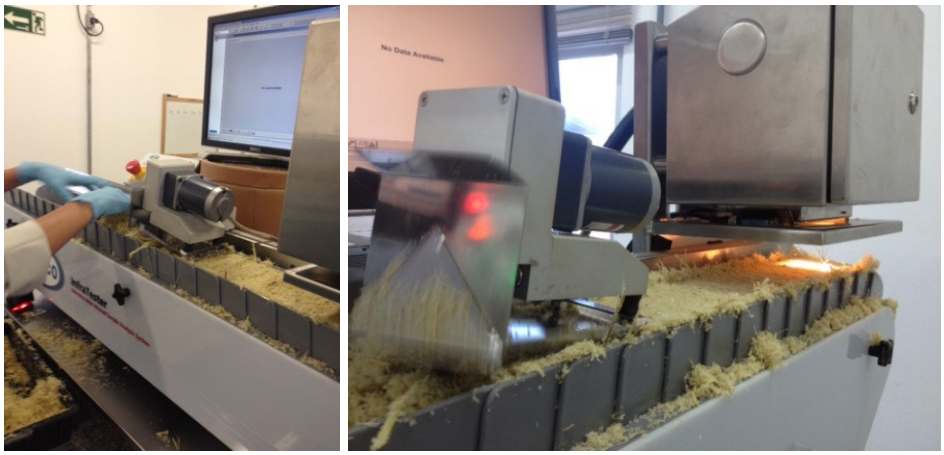

FIGURA 7.4**:** Material desfibrado de cana-de-açúcar inserido na esteira para análise.

O resultado para este procedimento foi um espectro de NIR para o material inserido na esteira. Na Figura 8.4 pode-se visualizar o exemplo de um espectro analisado por NIR para uma amostra de cana-de-açúcar.

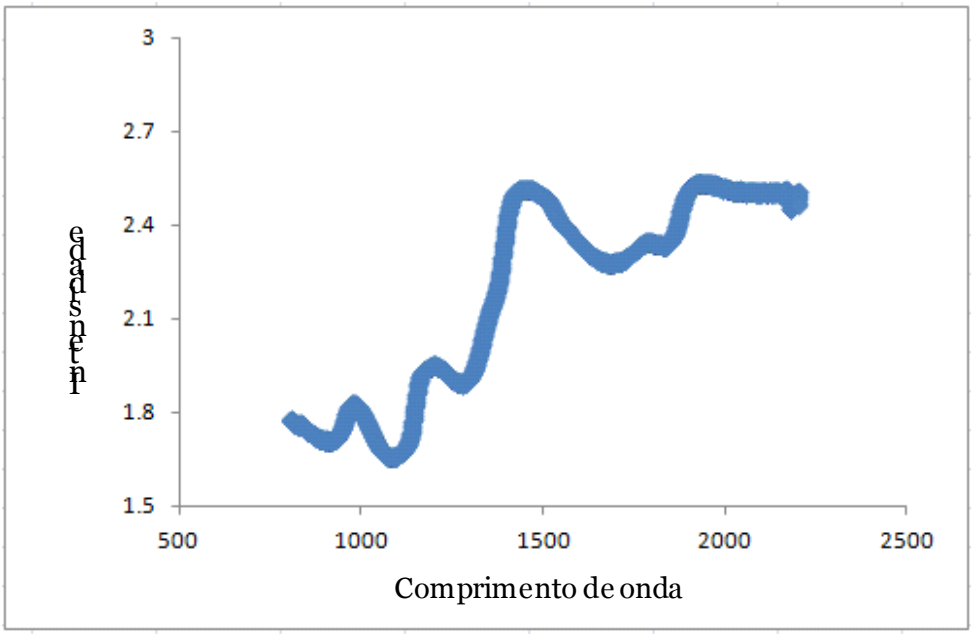

FIGURA 8.4**:** Espectro de cana-de-açúcar

4.1.2.– Análise pelo método *via* úmida

Para aquisição de dados para método de referência foram usados oito diferentes equipamentos homologados pelo Consecana. Para este método analítico chamado de *via* úmida, cada análise foi feita por um diferente equipamento ou em algumas delas, com mais de um. Na Figura 9.4 pode-se ver exemplos dos dois equipamentos analíticos mais importantes para amostragem (Sacarímetro e refratômetro digital). Ainda foram utilizados uma balança semi-analítica, uma estufa de secagem com circulação de ar forçada, uma osmose reversa, uma prensa hidráulica e um agitador magnético.

Neste método, 500 g de material desfibrado deve ser prensando em prensa hidráulica durante 1 minuto a pressão de 500kg/J. Após este procedimento, obtém-se o caldo e o bagaço úmido de fibra. Este bagaço úmido é pesado e passa por um processo de secagem em estufa durante pelo menos 5 horas a 105ºC. O bagaço de fibra seco era novamente pesado e obtinha-se a porcentagem de fibra do material através da equação:

$$F = \frac{(100 * PBS) - (PBU * B)}{5 * (100 - B)}$$

Onde: F= fibra; PBS= peso do bagaço seco; PBU= peso do bagaço úmido e B= valor de brix.

No caldo, a primeira análise feita foi a quantificação de brix. Neste processo, seis gotas do caldo filtrado eram gotejadas sobre o prisma do refratômetro que instantaneamente responde com o valor de brix da amostra.

A segunda análise feita foi leitura de pol, neste procedimento 200 mL de caldo eram homogenizados com sete gramas de mistura clarificante a base de cloreto de alumínio, celite e hidróxido de cálcio. O caldo foi homogenizado utilizando homogenizador magnético durante um minuto e filtrado utilizando papel de filtro. O filtrado desta mistura era analizado por um polarímetro a 20ºC e a resposta do equipamento era o valor de leitura de pol da amostra.

FIGURA 9.4**:** Instrumentação utilizada na *via* úmida (Sacarímetro e Refratômetro).

As outras análises eram feitas baseando-se em resultados das análises de brix, leitura de pol e fibra. Os cálculos usados foram referenciados nas equações preditas pelo CONSECANA [2].

Os valores de pol de caldo foram previsto segundo a equação:

$$S = (1{,}00621 * Z) * (0{,}2605 - 0{,}0009882 * B)$$

Onde: S= pol de caldo; Z = leitura de pol e B = valor de brix.

Os valores de pureza foram previstos segundo a equação:

$$Q = \frac{100 * S}{B}$$

Onde: Q= pureza; S= pol de caldo e B=brix do caldo.

Os valores de pol de cana (PolCana) foram previstos segundo a equação:

$$PolCana = S * (100 - 0.01 * F) * coef$$

Onde: S= pol do caldo; F= fibra da cana e coef= coeficiente utilizado para a transformação da pol do caldo extraído pela prensa em pol de cana.

4.2. – Aquisição do banco de dados

Na intenção de se aplicar o modelo de calibração em amostras de locais que apresentavam diferentes fatores, houve uma necessidade de inserir no banco de dados amostras que representassem ao máximo esta assimilaridade. Para tal foram coletadas amostras da estação experimental de Conchal (75% dos dados), Araçatuba (15% dos dados) e Mandaguaçu (15% dos dados). Estas amostras foram enviadas para Conchal onde está instalado o laboratório mestre contendo os equipamentos de preparação de amostra, equipamento NIR e todos os equipamentos necessários para análise pelo método de *via* úmida (refratômetro digital, sacarímetro, prensa hidráulica, estufa de secagem, agitador magnético, balança semi-analítica e osmose reversa).

Dentro do laboratório, a amostra recebia uma identificação dupla e parte da mesma amostra era separada para aquisição de espectros e parte para quantificação dos 6 parâmetros de interesse. Um esquema global do trabalho pode ser visto na Figura 10.4.

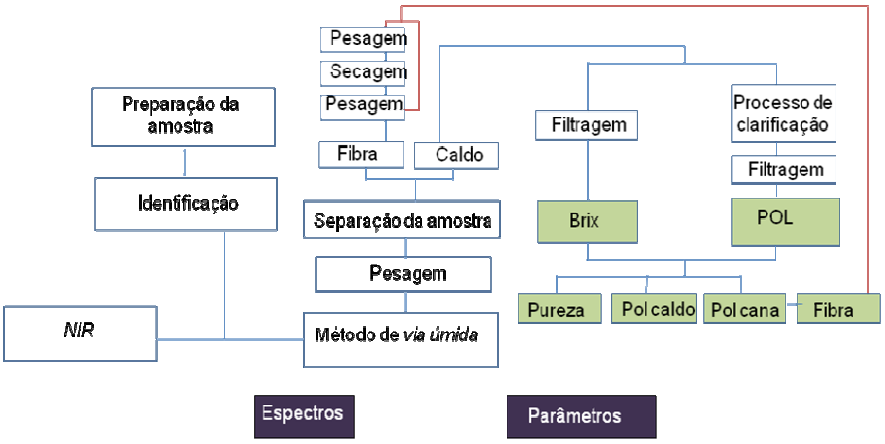

FIGURA 10.4**:** Fluxo de trabalho para aquisição do banco de dados.

Este trabalho foi desenvolvido nos anos de 2012 e 2013 até que houvesse um banco de dados contendo 7000 amostras com o máximo de representatividade possível.

4.3. – Pré processamento dos dados

Os principais objetivos da aplicação de técnicas de pré-processamento são eliminar informações não relevantes do ponto vista químico [11]. Assim, foram utilizados 3 pré-processamentos na matriz de espectros e um pré-processamento na matriz dos parâmetros.

4.3.1.– Normalização

Com o objetivo de reduzir as influências de variações presentes no conjunto de dados deixando cada observação representada de forma otimizada e consistente, a matriz de dados de espectros foi normalizada. A normalização reduz principamente os efeitos da intensidade total de perfis de resposta, devido a variações na concentração da amostra e caminho óptico. No processo de normalização pela norma, um dos mais comumente usados em espectros, realiza-se uma divisão de cada variável pela raiz quadrada da soma dos quadrados de todas as variáveis para uma dada amostra [11]. O conjunto de dados foi normalizado utilizando uma rotina com as seguintes funções no Matlab:

```
>> function Xn=normr(X)
>>% função que recebe o vetor ou matriz e devolve um vetor ou matriz normalizados.
>> [m,n] = size (X) %Este comando permite visualização do número de linhas e colunas da matriz.
>> for i=1:m
>> teste=norm(x(i,: ));
>>% Xn é a matriz normalizada.
>> end
```

4.3.2.– Primeira Derivada

O banco de dados originais possuía um problema de deslocamento e inclinação de linha de base. Este tipo de problemática é corrigido através da

derivação de espectros. Este tipo de pré processamento reduz matematicamente o ruído, aumentando a relação entre sinal e ruído. Nestes métodos uma janela é selecionada e assim o tamanho da janela influencia diretamente o resultado do alisamento [11]. Para calcular a primeira derivada dos dados foi utilizada uma rotina no Matlab que usou o método de Savitzky-Golay com uma janela de 19 e um polinômio de segunda ordem. O conjunto de dados foi derivado utilizando uma rotina (sgolay_diss) com as seguintes funções no Matlab:

```
>> function [var_out,Xout]=sgolay_diss(var_in,X,wid,ordr,nder)
>> % var_in= identificação das variáveis
>>% X= conjunto de dados de espectros
>>% wid = tamanho da janela (necessita ser um número ímpar e maior que 3)
>>% ordr = ordem do polinômio
>>% nder = cálculo (0 para smooth, 1 para primeira derivada e 2 para segunda derivada)
>> %var_out = nova identificação das variáveis
>>% Xder = matriz derivada

>> % Efetivação dos cálculos

>> [m,n]size(X);
>>A=zeros(wid,ordr+1);
>>w=(wid-1)/2;
>>A(:,1)=ones(wid,1);
>>A(:,2)=[-w:w];
>>for i=2:ordr
>>A(:,i+1) = A(:,2).^i;

>> end

>> B= (inv(A'*A))*A;
>> fcoeff=B(nder+1,:);
>> istrt=w+1;
```

```
>> istop=n-w;
>> Xout=zeros(m,n-s*w);
>> for i=1:m
>>for j=istrt:istop
>> Xout(i,j-w)=fcoeff*X(i,j-w:j+w)';

>>end
```

4.3.3.– Centrar na média

Este pré-processamento tem o intuito de mover as coordenadas para a origem do sistema, onde cada variável passará a ter média zero, fazendo com que as diferenças nas intensidades relativas das variáveis sejam percebidas com facilidade. Este processo consiste em calcular a média das intensidades para cada comprimento de onda e subtrair cada uma das intensidades do valor médio [11]. Para executar este pré-processamento, utilizou-se uma rotina (data_pre) do Matlab, esta rotina responde com os valores da matriz centrada na média e a matriz autoescalada. Não havia interesse em realizar o autoescalamento da matriz de espectros então esta matriz foi excluída após os cálculos. A função para execução desta rotina é a seguinte:

```
>> function [Xauto, Xcm, xm, xstd] = data_pre (X);
>> % Xauto= matriz autoescalada
>>% Xcm = matriz centrada na média
>>% xm= vetor contendo a média das variáveis
>>% xstd = vetor contendo o desvio padrão das variáveis
>>% X = matriz na qual se deseja aplicar os tratamentos

>>% Efetivação dos cálculos

>> [m,n]=size(X);
>> xm=mean (X);
>> xstd = std (X);
>> A=ones (m,1)*xm;
```

```
>> B=ones (m,1)*xstd;
>> Xauto= (X-A)./B
>> Xcm=X-A;
```

Ao final destes pré-processamentos, obteve-se uma matriz de espectros normalizada, derivada e centrada na média, tornando os dados melhor condicionados para análise trazendo melhor eficiência na análise exploratória subsequente.

4.3.4.– Autoescalamento

Autoescalar os dados têm como objetivo colocar as variáveis dentro de uma mesma escala quando tais unidades são diferentes ou quando a faixa de variação é muito grande. No processo de autoescalamento, todos os dados passam a ter média zero e desvio padrão igual a 1.

Para executar este pré-processamento, utilizou-se uma rotina (data_pre) do Matlab, esta rotina responde com os valores da matriz centrada na média e a matriz autoescalada. Não havia interesse em centrar na média da matriz de parâmetros (**Y**) então esta resposta foi posteriormente excluída. A função para execução desta rotina foi a data_pre.

4.4. – Análise Exploratória

4.4.1.– PCA – Análise de componentes principais

A PCA constitui, em muitas maneiras, a base para uma análise multivariada dos dados. Como resultados de um cálculo de PCA temos os *scores* que apresenta informações referentes as amostras e os *loadings* que apresenta informações importantes sobre as variáveis. Após o pré processamento dos dados de espectros, calculou-se uma PCA para a matriz **X**.

Para o cálculo da PCA, foi utilizada uma função *svd* já disponível no Matlab. O script da efetivação dos cálculos foi:

```
>> Function: [scores, loads, var_exp] = pca_dis (X)
>> % [scores, loads, var_exp] = pca_dis (X)
>> % scores = scores da matriz de espectros
>> % loadings = loadings da matriz de espectros
>> % var_exp = variância explicada da matriz de espectros
>> % X= matriz de espectros

>>  % Efetivação dos cálculos

>> [m,n] = size (X);
>> [U,S,V]=svd (X, 'econ');
>> scores = U*S;
>> loadings = V;
>> var_exp=(diag(S).^2)*100/(sum(diag(S).^2));
```

4.5. – Construção do modelo de calibração

4.5.1.– PLS – Regressão por mínimos quadrados

Após o pré-processamento dos dados e cálculo da PCA, identificou-se partes importantes do espectro, possíveis outliers, a variância explicada dos dados e o número de componentes principais que explicava a maior parte dos mesmos. Com estas análises conseguiu-se informações importantes para construção do modelo de calibração.

Nesta fase do trabalho, o banco de dados foi dividido em dois grupos: (i) matriz de dados contendo 6300 pares de espectros e respostas dos parâmetros (90% dos dados) e foi utilizada para criação do modelo de calibração e (ii) Matriz contendo 700 pares de resultados que foi separada para a validação externa.

A matriz de calibração com os espectros originais foi importada para o Matlab e com base nos resultados da PCA, selecionou-se apenas as variáveis interessantes. Após esta seleção, a matriz passou a ter apenas 1000 variáveis (900 a 1400nm) que foram novamente preprocessadas com as mesmas técnicas e rotinas anteriormente usadas no banco de dados geral. A matriz de parâmetros de interesse foi separada da mesma forma e autoescalada.

Após aplicação destas rotinas, tinha-se: (i) uma matriz de espectros contendo 6300 amostras com 1000 variáveis normalizadas, derivadas e centradas na média e (ii) uma matriz de respostas contendo 6300 amostras com 6 parâmetros que foram obtidos pelo método de referência. A matriz de espectros recebeu o nome de **X** e a matriz com parâmetros recebeu o nome de **Y**.

Estas duas matrizes foram importadas para o Matlab para que se pudesse construir o modelo através do PLS. Para o calculo do PLS, foi usada uma rotina escrita em ambiente do MatLab (pls_ed) com uma função para o cálculo de PLS usando a validação cruzada pelo método *leave-one-out*. O script da efetivação dos cálculos foi:

```
>> Function [rmsec, relrmsec, rmsecv, coef, Ypred] = pls_ed (Xin, Yin, varID, LV)
>> % Xin = conjunto de dados
>> % Yin = parâmetro a ser modelado
>> % varID = identificação das variáveis
>> % LV = máximo do número de componentes principais
>> % rmsec = root mean square error of calibration
>> % relmsec = rmsec relativo (rmsec/ (ymax-ymin))*100
>> % rmsecv = rmsec para validação cruzada
>> % coef = coeficientes da regressão
>> % Ypred = Y previsto
>> % A matriz de rmsecv é pré estabelecida

>>  % Efetivação dos cálculos ( pls_ed)

>> [m,n]=size(Xin);
>> rmsecv=zeros(m,LV);

>> % Loop para a validação cruzada leave-one-out
>> for i=1:m
    >>Ycal=Yin;
    >>Xcal=Xin;
    >>Ycal (i)=[ ];
```

```
    >>Xcal (i,: )=[ ];
    >>Ycv=Yin(i);
    >>Xcv=Xin(i,: );

>> for j=1:LV
    >> [T,P,U,Q,W,D] = npls (Xcal, Ycal,j);
    >> coef= W=inv(P`*W)*D`;
    >> Ypr_cv=Xcv*coef;
    >> rmsecv (i,j)=sqrt (sum((Ycv-Ypr_cv).^2/lenght(Ycv));

>>end
>>end
```

Como notado anteriormente, utilizou-se uma outra função (npls). O script desta função de efetivação do npls foi:

```
>> function [T,P,U,Q,W,D,Flag]=npls (X,Y,ncom)
>> % X= matriz X
>> % Y= matriz Y
>> % ncom = número de componentes a serem extraídos
>> % T= scores da decomposição da matriz X
>> % P= loadings da matriz X
>> % U= scores da decomposição da matriz Y
>> % Q= loadings da matriz Y
>> % W= pesos da matriz utilizados para calcular T dado X
>> % D= usado para regressão da matriz Y. Y=T*D+error. T é obtido

>>  % Efetivação dos cálculos

>> T=X*W*inv(P`*W)
>> % substituindo T por essa relação, tem-se:
>> Y=X*W*inv(P`*W)*D` + erro ou Y=X*(W*inv(P`*W)*D`)+erro
>> Flag= indica se a convergência foi alcançada
```

4.5.2.– Previsão de amostras desconhecidas

Quando se calcula um modelo PLS, a resposta interessante são os coeficientes de regressão gerados. Cada parâmetro gerou um coeficiente que será utilizado na previsão de parâmetros para amostras desconhecidas.

Uma matriz **X**, contendo amostras desconhecidas, será prevista utilizando a seguinte equação:

$$Xcm=X-xm$$

Onde **X** será a matriz de dados desconhecidos já normalizados e derivados, xm será o x médio calculado através dos dados de calibração e Xcm será a matriz de espectros de amostras desconhecidas que será centrada na média com o x médio da calibração. Depois que a matriz **X** foi centrada na média corretamente, o cálculo do parâmetro de interesse foi previsto com a seguinte equação:

$$Y=(((Xcm*coef(:,i)*ystd(:,1))+ym(:,1))$$

Onde coef será o coeficiente referente ao parâmetro que se deseja prever indicando qual o número de variáveis latentes (i) que melhor se encaixa na análise de determinado parâmetro. Para o cálculo do coeficiente de regressão, os dados da matriz **Y** foram autoescalados, logo, os dados desta equação estarão autoescalados, por este motivo necessita-se fazer as operações matemáticas para reverter este processo (multiplicar pelo desvio padrão e somar com a média).

4.5.3.– Validação externa

Quando separou-se o modelo de calibração, 10% do banco de dados, referente a 700 pares de espectros e parâmetros de interesse foram separados para validação externa. Neste processo, as amostras foram previstas pelo modelo criado e comparadas com os valores respectivos que foram obtidos pelos métodos de referência.

4.6. – Estratégias de transferência de calibração

4.6.1.– Aquisição de dados

O primeiro passo para aplicar uma estratégia de transferência de calibração, assim como na construção de um modelo, é a aquisição de dados. No caso de uma transferência de calibração, os dados de mesmas amostras precisam ser analisados pelo método de referência, pelo equipamento responsável pela calibração, chamado de mestre e pelo equipamento pelo qual se deseja imputar a transferência, chamado de escravo.

Nesta fase de aquisição de dados, tinha-se dois equipamentos escravos que funcionam nos laboratórios de Araçatuba e Mandaguaçu. Para aquisição de dados, os equipamentos passaram por um estágio no laboratório de Conchal onde se encontram todos os equipamentos de análise por método convencional e o equipamento mestre responsável pela aquisição dos dados para construção do modelo de calibração.

Os equipamentos foram instalados em sincronia para que o fluxo de trabalho seguisse da forma que quando uma amostra entrasse no laboratório, fossem coletados os espectros através do equipamento mestre, através dos equipamentos escravos e se quantificasse os seis parâmetros de interesse através da *via* úmida.

Isto foi feito com os dois equipamentos escravos até que se tivesse um banco de dados com 1100 conjunto de dados para o primeiro equipamento escravo e 1100 conjunto de dados para o segundo equipamento escravo como mostra o esquema da Figura 11.4.

Os dados foram previstos utilizando os coeficientes calculados através da calibração mestre.

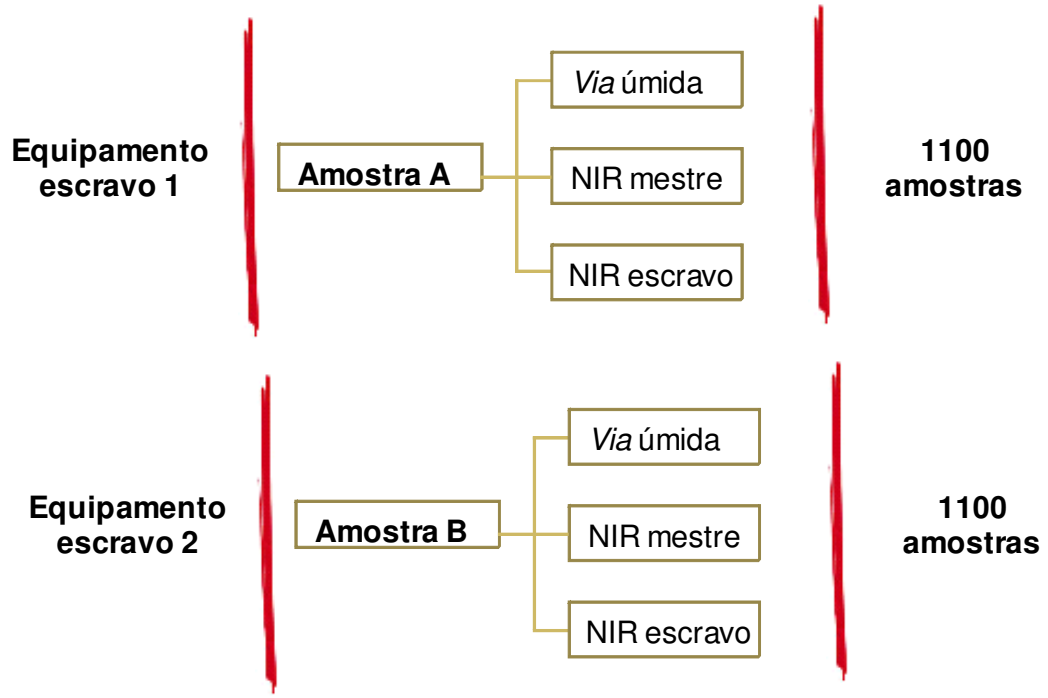

FIGURA 11.4**:** Fluxo de aquisição de dados para aplicação de transferência de calibração.

Nessa dissertação de mestrado foram escolhidas três estratégias de transferência de calibração nesta fase do trabalho e aplicadas no banco de dados.

4.6.2.– Recalibração

Para testar o primeiro método de transferência de calibração, usou-se o método de recalibração onde foi calculado um modelo PLS para cada conjunto de dados.

Para o primeiro equipamento escravo, utilizou-se 1000 pares de dados para construção de um modelo de calibração através de PLS e 100 dados para validação externa. O mesmo foi feito para o segundo equipamento escravo.

4.6.3.– *Model Update*

Para testar a segunda estratégia de transferência de calibração, usou-se o método chamado de *Model Update.* Este método consiste na construção de um

modelo de calibração utilizando um banco de dados global representado pelo mesmo número de amostras provindas de todos os equipamentos dos quais se deseja prever amostras desconhecidas.

Para esta estratégia tinha-se então 1000 amostras do equipamento escravo 1; 1000 amostras do equipamento escravo 2 e foram selecionadas aleatoriamente 1000 amostras do equipamento mestre. Criou-se um banco de dados com 3000 espectros e construiu-se um modelo de PLS com este banco de dados. Teoricamente este modelo geral deveria ser robusto o suficiente para prever amostras provenientes dos 3 equipamentos com erros aceitáveis. Para validação externa, utilizou-se 100 amostras de cada equipamento (total de 300 amostras).

4.6.4.– Correção por *Slope and Intercept*

Para testar a terceira estratégia de transferência de calibração, aplicou-se o método chamado correção por *slope and intercept.* Esta estratégia consiste no ajuste dos dados previstos por um equipamento escravo para que estes resultassem o mais parecidos possíveis com os dados previstos por um equipamento mestre para mesmas amostras.

Para o cálculo dos valores de *slope* e dos valores de *intercept,* foram usados 1000 resultados de amostras analisados pelo equipamento escravo 1 previstas por calibração do equipamento mestre e 1000 dados dos parâmetros de referência para as mesmas amostras. Os valores foram calculados através de fórmulas existentes do Microsoft Excel.

Após os valores calculados, os mesmos foram aplicados em 100 amostras de validação externa segundo a equação a seguir:

$$Z=((Y*slope)+intercept)$$

Onde, **Y** é a matriz de dados do equipamento escravo prevista por calibração de equipamento mestre; **Z** é a matriz de resultados com correção e *slope* e *intercept* são valores criados através de comparações de correção entre os dados [19].

O mesmo foi feito com amostras do segundo equipamento escravo para valores referentes àquele equipamento.

RESULTADOS E DISCUSSÕES

5. – RESULTADOS E DISCUSSÕES

5.1. – Ánalise exploratória usando PCA

A variância explicada para o conjunto de dados de espectros mostra por meio de gráficos qual a melhor escolha de componentes principais do qual melhor se explica o banco de dados. Os resultados apresentados na Figura 12.5 para variância explicada do banco de dados do projeto mostra que o modelo tem por volta de 60% dos dados explicados com 3 componentes principais. A partir disto, o ganho pode ser considerado pequeno.

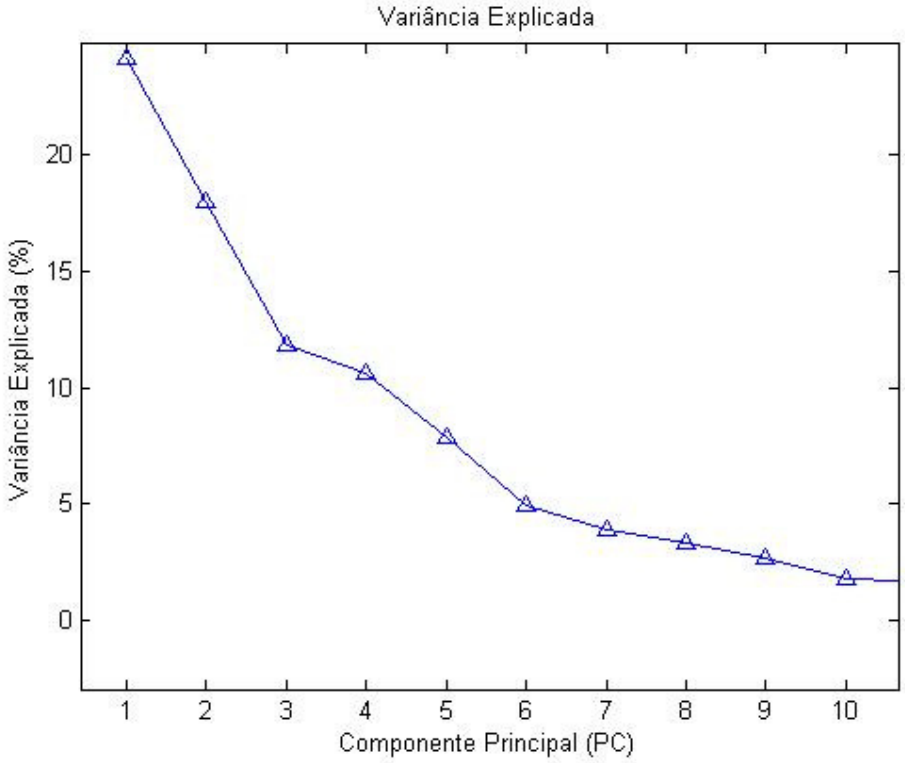

FIGURA 12.5: Variância explicada para o conjunto de dados de espectros.

Os *scores* apresentam os resultados do agrupamento das amostras facilitando a visualização de possíveis outliers e similaridades no conjunto de dados referente a coordenada das amostras. Não havia intenção de retirar nenhuma amostra do banco de dados e a análise dos *scores* vistos na Figura 13.5 mostrou que realmente o conjunto todo de dados poderia ser utilizado.

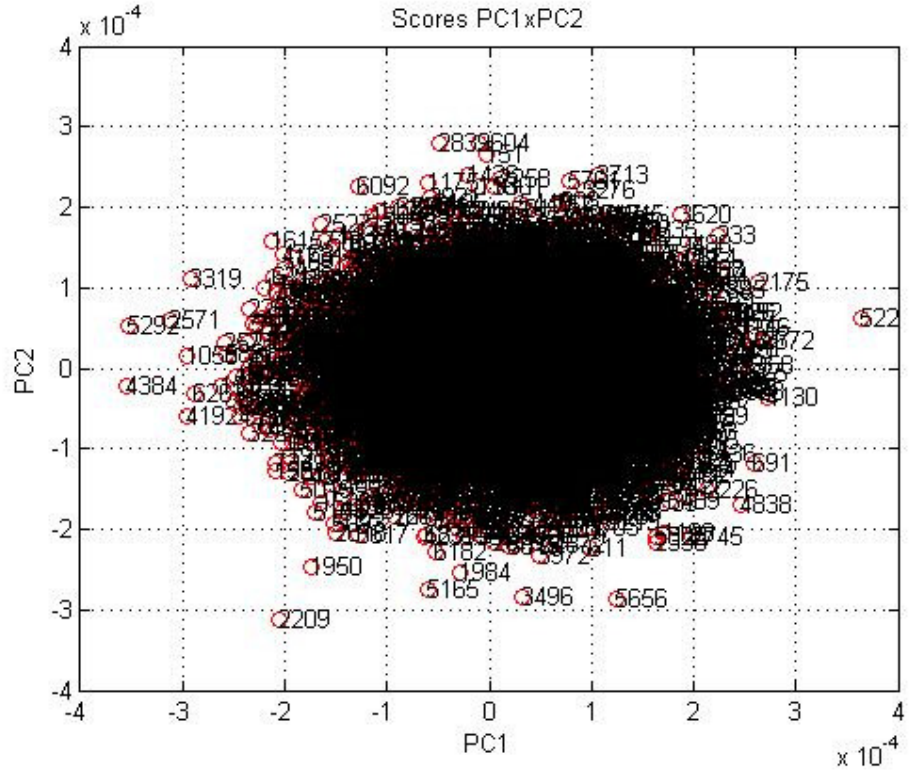

FIGURA 13.5: *Scores* para o conjunto de dados de espectros.

Os *loadings* apresentam os resultados referentes as variáveis dos espectros. A partir da análise de *loadings* percebe-se quais são as áreas onde há maior agrupamento de informação química interessante no espectro analisado. As Figuras 14.5 e 15.5 mostram, respectivamente os *loadings* para a primeira e segunda componentes principais.

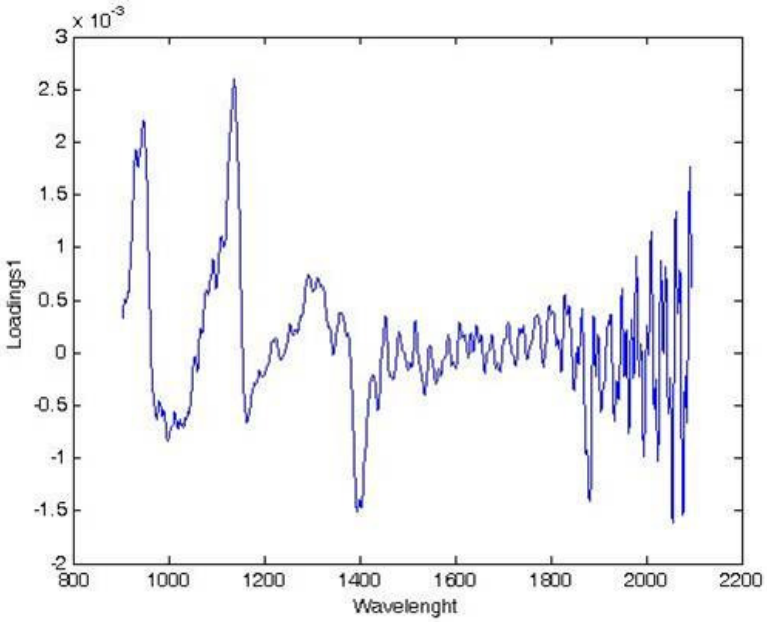

FIGURA 14.5**:** *Loadings* da primeira componente principal.

Verificou-se que entre as regiões de 900 a 1400 nm encontram-se as partes com mais informações químicas importantes. As principais moléculas que apresentam frequência nestas faixas espectrais são: $-CH_{2=}CH_2$, –CH=CH-, $-NO_2$ e $CO_3=$.

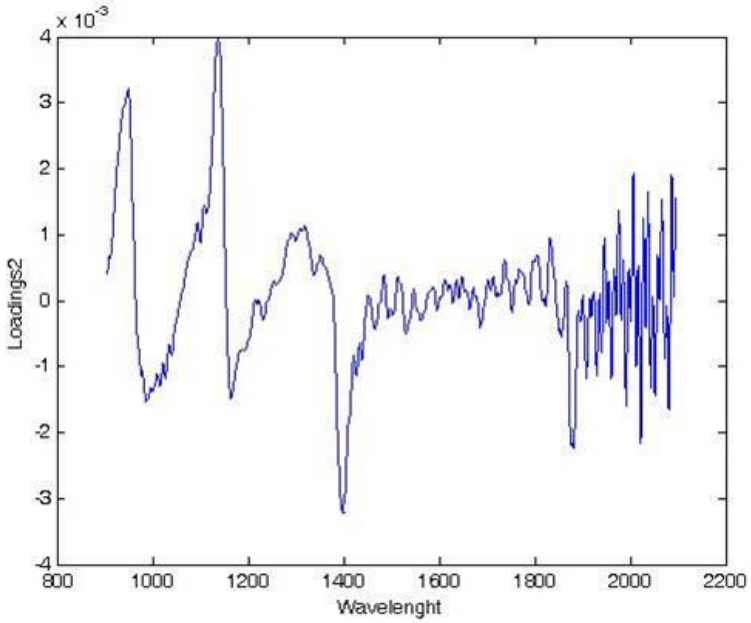

FIGURA 15.5**:** *Loadings* da segunda componente principal.

Comprovou-se que entre as regiões do espectro entre 900 a 1400nm encontram-se as áreas que contém maior informações do ponto de vista químico, as demais apresentavam demasiado ruído espectral.

5.2. – Pré processamento dos dados

De acordo com os resultados apresentados pelos *loadings,* as informações químicas importantes encontravam-se nas faixas entre 900 a 1400 nm. Estas variáveis foram selecionadas para o cálculo de modelo de calibração. A Figura 16.5 mostra o banco de dados apenas com as variáveis interessantes selecionadas.

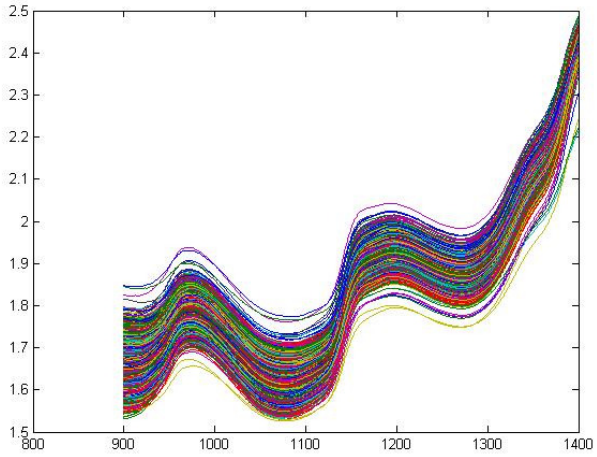

FIGURA 16.5**:** Espectros do banco de dados da calibração (6300 espectros) com as váriaveis interessantes selecionadas.

Ao todo foram utlizados 3 preprocessamentos na matriz de espectros (**X**) e um pré processamento na matriz de parâmetros (**Y**). Nas Figuras 17.5, 18.5 e 19.5 pode-se verificar o conjunto de dados, respectivamente normalizados, derivados e centrados na média.

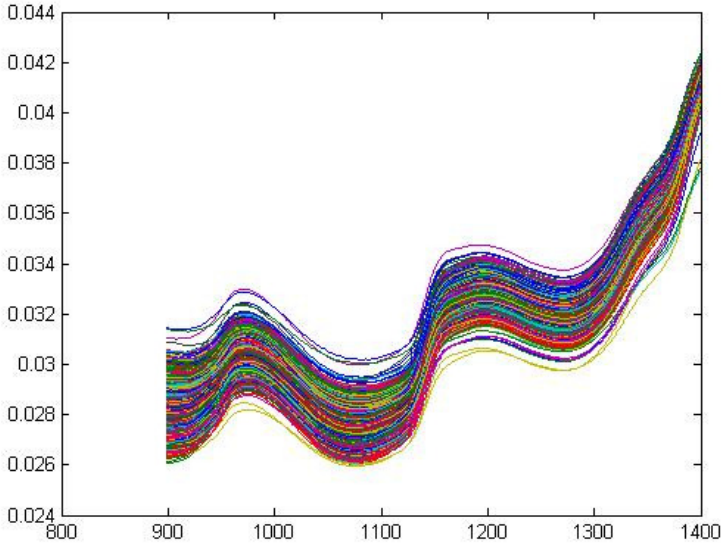

FIGURA 17.5**:** Espectros do banco de dados da calibração (6300 espectros) com as váriaveis interessantes selecionadas normalizados

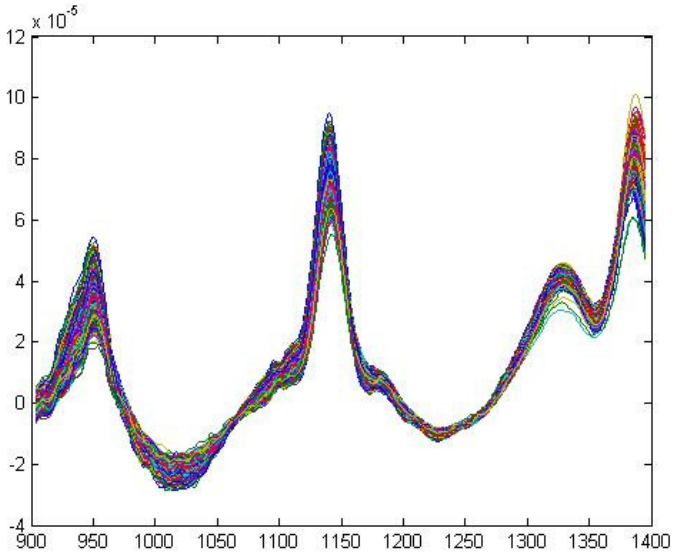

FIGURA 18.5**:** Espectros do banco de dados da calibração (6300 espectros) com as váriaveis interessantes selecionadas normalizados e derivados.

Na tentativa de melhorar os resultados, melhor condicionando os espectros para a construção do modelo, verificou-se a possibilidade de inverter a ordem da primeira derivada e da normalização. Os resultados após este teste não

foram tão satisfatórios comparados aos normalizados e derivados, por isto, foi mantida esta ordem.

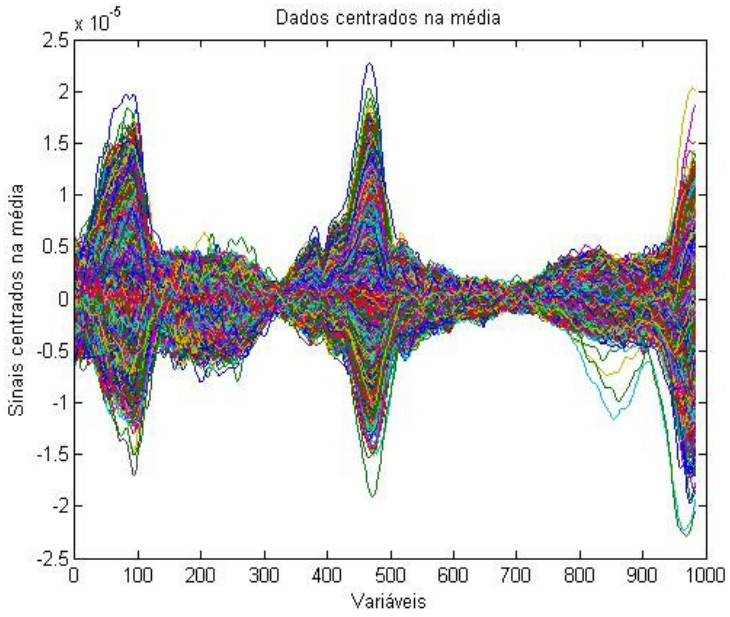

FIGURA 19.5**:** Espectros do banco de dados da calibração (6300 espectros) com as váriaveis interessantes selecionadas normalizados, derivados e centrados na média.

5.3. – Construção do modelo de calibração

Como resposta da construção do modelo de calibração, gerou-se os coeficientes da regressão que consistem em uma matriz de dados que será usada posteriormente para previsão de novas amostras. Estes coeficientes são específicos para cada parâmetro analisado.

Também calculado pela rotina, o sistema apresentou um gráfico com os números de variáveis latentes. Similar as componentes principais na PCA, a escolha do número de variáveis latentes usadas é muito importante e pode causar diferenças na robustez do modelo e problemas na previsão de amostras desconhecidas. A escolha de poucas variáveis latentes pode apresentar um modelo pouco roubusto e com erros inaceitáveis, porém a escolha de muitas variáveis latentes pode gerar um problema chamado de *overfit* ou supercalibração deixando o

modelo otimizado apenas para prever amostras que foram usadas para criar o modelo.

Cada variável analisada apresentou um gráfico de rmsec e rmsecv onde foi possível visualizar melhores condições de variáveis latentes para cada parâmetro. Na Figura 20.9 pode-se ver o gráfico de rmsec e rmsev para o primeiro parâmetro analisado (Brix).

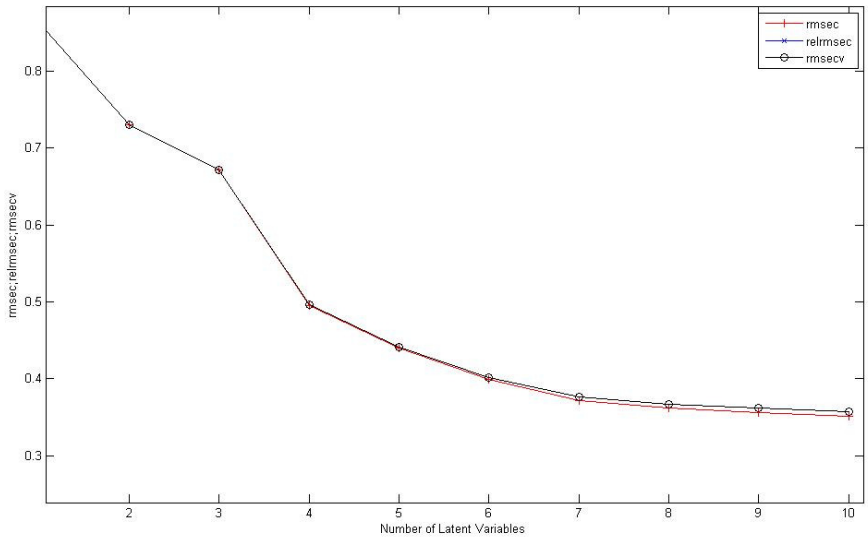

FIGURA 20.5**:** Gráfico de rmsec e rmsev para o parâmetro brix.

Nesta resposta dos resultados de rmsec e rmsecv, pode-se analisar qual a intensidade do erro que a modelagem apresenta para amostras usadas na criação do modelo e também amostras da validação cruzada, ou seja, tem-se uma visão do comportamento do erro para amostras do banco de dados e amostras externas ao banco. Este tipo de análise é de extrema importância pois é a partir dele que se decide qual o número de variáveis latentes melhor condiciona determinado parâmetro no modelo de calibração.

Nesta escolha do número de variáveis latentes, o operador decidirá quais são as melhores condições do modelo. O método proposto para decisão do número de variáveis latentes é chamado de método de cálculo da máxima curvatura modificado [20]. Neste método cada angulação dos pontos é cálculada e o número de variáveis latentes escolhidos é referente ao ponto com maior angulação nos dados. Esta angulação é prescrita conforme as equações:

$$90 - arctg\left[\frac{Yb - Ya}{Xb - Xa}\right] = b1$$

$$90 - arctg\left[\frac{Yc - Yb}{Xc - Xb}\right] = b2$$

$$b3 = 90 + b1 + b2$$

Onde b1, b2 e b3 são as angulações dos pontos.

A Figura 21.5 mostra a esquemática dos pontos b1, b2 e b3. Este processo deve ser feito para todos os pontos do gráfico, principamente nos pontos onde se tem dúvidas de quais tem maiores angulações.

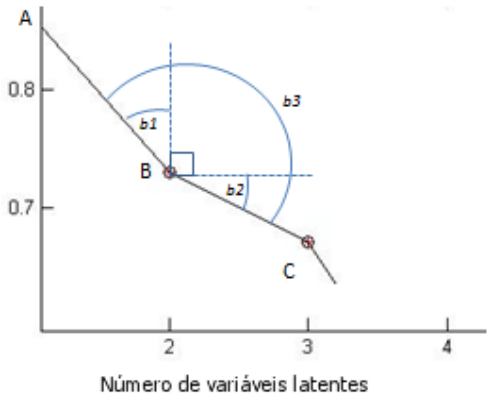

Figura 21.5: Esquemática dos pontos de angulação para cálculo da máxima curvatura modificada.

Para o parâmetro brix, a otimização do modelo é com 4 variáveis latentes. Percebe-se que a partir de 4 variáveis, o modelo não apresenta diferenças de erros significativos para este parâmetro analisado. Para cada outro parâmetro necessita-se deste tipo de análise que irá responder com valores diferentes. Pode-se afirmar que para cada parâmetro há uma correlação distinta, trazendo respostas diferentes fazendo com que o número de variáveis latentes seja específico para cada um deles.

É importante mencionar ainda que as análises pelo método convencional possuem um erro imbutido causada pela complexibilidade da amostra, impurezas e possíveis falhas humanas. Quando comparados os erros apresentados pelo modelo de calibração deve-se levar em consideração o erro que cada amostragem apresenta. Neste passo, para ter base de dados comparativos, foram feitas 30 amostras em triplicata utilizando o método de referência e estimado o erro entre elas.

O brix, responsável pelos sólidos solúveis presentes no caldo da cana é analisado pelo método convencional através de refratômetro digital com o caldo extraído por uma prensa hidráulica. Os valores de brix representados no banco de dados têm um mínimo de 10.9 e máximo de 26,09^{o}brix e apresenta um erro na análise convencional de 0,48. As melhores condições para este parâmetro foram com 4 variáveis latentes.

A pol, responsável pela sacarose aparente presente no caldo da cana é analisada pelo método convencional através de polarímetro digital com o caldo extraído por uma prensa hidráulica e clarificado com celite, hidróxido de calcio e cloreto de alumínio. Os valores de leitura de pol representados no banco de dados tem um mínimo de 25,76 e máximo de 95,94^{o}Z e apresenta um erro na análise convencional de 3,09. As melhores condições para este parâmetro foram com 4 variáveis latentes.

A pureza, responsável pela quantidade de sacarose imbutida nos sólidos solúveis presentes no caldo da cana é analisada pelo método convencional através de refratômetro digital e polarímetro com o caldo extraído por uma prensa hidráulica. Os valores de pureza representados no banco de dados têm um mínimo de 5,03 e máximo de 96,57% e apresenta um erro na análise convencional de 2,03. As melhores condições para este parâmetro foram encontradas com 4 variáveis latentes.

A fibra, responsável pela porcentagem total de fibra contida na amostra é analisada pelo método convencional através de pesagens na fibra após prensagem por prensa hidráulica e pesagem da fibra seca por 5 horas em estufa de secagem. Os valores de fibra representados no banco de dados de calibração têm um mínimo de 7,76 e máximo de 26,34 % e apresenta um erro na análise convencional de 0,78. A correlação para fibra apresentou-se menor comparados aos outros parâmetros, para tanto, as melhores condições de variáveis latentes foram com 3.

A pol de caldo é responsável pela porcentagem de sacarose contida no caldo bruto da amostra de cana-de-açúcar. Esta amostragem é feita baseado nos dados de brix, feitos pelo refratômetro e de leitura de pol, analisadas através do polarímetro. Os valores de pol de caldo representados no banco de dados de calibração têm um mínimo de 6,47 e máximo de 23,95 e apresenta um erro na

análise convencional de 1,12. As melhores condições para este parâmetro foram com 4 variáveis latentes.

Levando em consideração que a quantidade de fibra presente no material interfere diretamente na eficiência da extração de sacarose em uma moenda industrial, o sexto parâmetro analisado, chamado de pol de cana considera a quantidade total de fibra presente para estimar a sacarose que poderá ser extraída do material e isto faz deste parâmetro um dos mais importantes analisados. Os valores de pol de cana representados no banco de dados têm um mínimo de 5,76 e máximo de 19,53 e apresenta um erro na análise convencional de 0,74. As melhores condições para este parâmetro foram com 4 variáveis latentes.

5.4. – Resultados gerais da construção do modelo

Após a construção do modelo, para uma análise geral dos resultados de erros, variáveis latentes e comparação com os erros intrínsecos dos métodos convencionais, pode-se construir uma tabela como a Tabela 1.5. Neste ponto do trabalho, é possível decidir se a modelagem é ou não satisfatória.

	Brix	**L.Pol**	**Pureza**	**Fibra**	**Pol de caldo**	**Pol de cana**
N. variáveis latentes	4	4	4	3	4	4
RMSEV	0,57	3,55	2,71	1,07	0,84	0,66
Erro *via* úmida (n = 3)	0,48	3,09	2,03	0,78	1,12	0,74

TABELA 1.5**:** Visão geral da modelagem.

Considerando resultados gerais apresentados, percebe-se que a fibra tem uma correlação menor quando comparadas aos outros parâmetros. Ainda pode-se perceber que os erros da validação basicamente confundem-se com os erros que do próprio método convencional.

5.5. – Transferência de calibração

Na escolha da melhor estratégia de calibração para determinada gama de análises de certo componente, algumas decisões, baseadas no estudo de caso, devem ser tomadas.

5.5.1. Recalibração

A técnica de recalibração não pode ser considerada uma estratégia de transferência visto que para sua aplicação, cada laboratório deveria possuir a gama de equipamentos dos métodos convencionais necessários para a retroalimentação do modelo local. Isto faz com que ela não seja uma transferência de calibração e sim um processo isolado de construção do modelo. Entretanto, o cálculo da recalibração deve ser feito para se ter dados de erros do que seria a melhor das hipóteses onde cada modelo seria construído e retroalimentado por resultados e espectros colhidos para cada um deles.

5.5.2. Model Update

Neste projeto, um dos principais problemas da aplicação de uma técnica de *model update* é quanto a quantidade de amostras disponíveis para um banco de dados geral. O equipamento mestre se encontrava em trabalho a mais tempo possuíndo um banco de dados de 7000 amostras. Já os dois outros equipamentos escravos disponíveis estavam em trabalho a menos tempo limitados a 1100 amostras cada. Para a aplicação desta técnica a limitação do banco de dados seria de 3000 amostras sendo 1000 de cada equipamento. Isso acabou gerando um problema de robustez do banco de dados comparado a uma calibração realizada com 7000 amostras onde a representação do banco de dados foi muito mais rica. Um outro importante problema apresentado nesta técnica foi a quantidade de amostras necessárias para a retroalimentação da transferência. Esta falta de robustez do modelo necessitaria de uma quantidade maior de transferência de amostras para aumentar a qualidade do modelo criado e, lembrando que os laboratórios ficam em cidades distantes e a amostra tem uma degradação rápida perante aos métodos analíticos, isto dificultaria a aplicabilidade desta técnica.

5.5.3. Correção por ajuste de *slope and intercept*

A terceira técnica de transferência de calibração testada, apresentou melhor beneficio de utilização uma vez que seria possível utilizar os coeficientes calculados através do banco de dados mais robusto e representativo. Isto faz com que os erros apresentados fossem menores. Também como benefício da utilização deste método apresentado, a utilização de um equipamento mestre presente em um laboratório com o método de referência facilitando a retroalimentação e garantindo a qualidade da mesma uma vez que as amostras estivessem em melhores condições analíticas.

A logística de transferência de amostras para retroalimentação da transferência continuava sendo um problema devido a distância e degradação do material, porém nesta técnica a atualização dos valores de *slope* e de *intercept* poderiam ser em menor quantidade melhorando a rapidez com que as amostras chegavam no laboratório mestre garantindo melhor qualidade nos processos analíticos.

5.5.4. Comparação dos erros apresentados

Um dos principais fatores que diferenciam na tomada de decisão no estudo de caso para escolha da técnica da transferência de calibração foram os erros apresentados.

Na Tabela 2.5 podemos comparar quais os erros apresentados para cada uma das técnicas.

Método	Brix	L.Pol	Pureza	Fibra	Pol de caldo	Pol de cana	Média
Sem transferência	2,03	11,66	6,93	3,81	2,75	1,64	4,8
Recalibração	0,51	3,00	2,58	0,98	0,71	0,59	1,39
Model Update	1,00	5,96	4,49	0,99	1,41	1,13	2,49
Slope e Intercept	0,95	4,92	3,22	1,52	1,15	0,68	2,07
Erro via úmida	0,48	3,09	2,03	0,78	1,12	0,74	1,25

TABELA 2.5**:** Comparativo dos erros apresentados pelas estratégias de transferência testados para material desfibrado de cana-de-açúcar.

Na primeira linha tem-se os valores dos dados de um equipamento escravo previstos com uma calibração mestre sem nenhuma correção. Os erros são considerados altos fazendo que os dados não fossem confiáveis.

Os melhores resultados apresentados foram para uma recalibração, isto já era esperado uma vez que os dados utilizados foram referentes ao equipamento de origem, porém esta técnica não foi aplicável devida a falta de estrutura para o projeto.

A técnica de *model update* apresentou resultados aceitáveis, porém maiores comparados as outras técnicas. Considerou-se que devido as problemáticas de transferências de amostras, que esta técnica não iria trazer muitos benefícios.

A última técnica testada, *slope and intercept* apresentou bons resultados com baixos erros e considerou-se a facilidade na aplicação desta técnica uma vez que a retroalimentação necessária poderia ser feita em menor escala e foi possível a utilização do rico banco de dados do equipamento mestre.

CONCLUSÕES

6. – CONCLUSÕES

As análises químicas em um programa de melhoramento genético para cana-de-açúcar são de extrema importância para obtenção de dados e tomada de decisões para a evolução do programa. As quantidades de amostras envolvidas em um programa bem estruturado é um número expressivo e consequentemente a demanda por um processo analítico rápido e confiável é impulsionado por esta necessidade. Os métodos convencionais conseguem atender apenas a base desta demanda. Como alternativa de método analítico, este projeto concluiu que é possível aplicar a tecnologia de infravermelho próximo em conjunto com a quimiometria para utilização de um método analítico rápido e confiável.

Conclui-se ainda que o PLS foi o método quimiométrico que apresenta melhores resultados de predição para análise de material desfibrado de cana-de-açúcar.

Com testes feitos entre laboratórios espalhados pelas áreas de maior concentração de cana no Brasil, constatou-se que o método de transferência de calibração por ajustes de slope e intercept apresenta melhor logística, menores erros e consequentemente melhor aplicabilidade para aumento da capacidade análise/dia.

No ano safra de 2014, foram analisadas 18500 amostras aplicando-se infravermelho próximo e transferência de calibração para análises quantitativas de cana-de-açúcar. Isto resultou em uma economia de recursos financeiros da ordem de R$ 100.000,00. Devido a complexidade e ao tempo de um programa de melhoramento, a curto prazo consegue-se apenas quantificar as economias financeiras, porém este tipo de economia não se compara a evolução que este número de amostras trouxe ao programa de melhoramento, uma vez que por métodos convencionais não se teria tempo hábil para realizar estas análises.

Para a empresa Monsanto do Brasil, foi redigido um padrão de operação confidencial reportando todos os procedimentos, rotinas e funções utilizadas para execução deste trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

7. – REFERÊNCIAS BIBLIOGRÁFICAS

1. Hogarth, D.M., Allsop P.G, MANUAL OF CANEGROWING. Bureau of Sugar Experiments Stations PO Box 86, Indooroopully, 4068 Australia

2. Conselho dos produtores de cana-de-açúcar, açúcar e álcool de São Paulo, CONSECANA - MANUAL DE INSTRUÇÕES. 5ª edição, Piracicaba 2006.

3. Holler, Skoog, Crouch, PRINCÍPIOS DA ANÁLISE INSTRUMENTAL APLICADA, 6ª edição editora bookman 200.

4. A.M.Souza, M.C.Breitkreitz, P.R.Filgueiras, J.J.Rohwedder, R.J.Poppi EXPERIMENTO DIDÁTICO DE QUIMIOMETRIA PARA CALIBRAÇÃO MULTIVARIADA NA DERTERMINAÇÃO DE PARACETAMOL EM COMPRIMIDOS COMERCIAIS UTILIZANDO ESPECTROSCOPIA NO INFRAVERMELHO. Química Nova, 1057-1065 (2013).

5. M.G.O'Shea, S.P.Staunton, M.Burleigh IMPLEMENTATION OF ON-LINE NEAR INFRARED (NIR) TECHNOLOGIES FOR THE ANALYSIS OF CANE, BAGASSE AND RAW SUGAR IN SUGAR CANE FACTORIES TO IMPROVE PERFORMANCE. International Sugar Journal 113 (2009).

6. D.E.Purcell, M.G. O'Shea, S. Kokot ROLE OF CHEMOMETRICS FOR AT FIELD APPLICATION OF NIR SPECTROSCOPY TO PREDICT SUGARCANE CLONAL PERFORMANCE. Chemometrics and intelligent laboratory Systems 113-124 (2007)

7. N. Sorol, E Arancibia, S.A. Bortolato, A.C. Olivieri, VISIBLE/NEAR INFRARED-PARTIAL LEAST-SQUARES ANALYSIS OF BRIX IN SUGAR CANE JUICE A TEST FOR VARIBLE SELECTION METHODS. Chemometrics ans Intelligent Laboratory Systems 100-109 (2010).

8. G.A.Santos, A.B. Pereira, G.H. Korndorfer, USO DO SISTEMA DE ANÁLISES POR INFRAVERMELHO PRÓXIMO (NIR) PARA ANÁLISES DE MATÉRIA ORGÂNICA E FRAÇÃO DE ARGILA EM SOLOS E TEORES FOLIARES DE SILÍCIO E NITROGÊNIO EM CANA-DE-AÇÚCAR. Biosci. J., Uberlândia, 100-108 (2010).

9. W.S. Lyra, E.C. Silva, M.C.Ugulino, W.D, Fragoso, CLASSIFICAÇÃO PERIÓDICA: UM EXEMPLO DIDÁTICO PARA ENSINAR COMPONENTES PRINCIPAIS. Química Nova, 1594-1597 (2010)

10. J.G.Sabin, M.F.Ferrão, J.C.Furtado, ANÁLISE MULTIVARIADA APLICADA NA IDENTIFICAÇÃO DE FARMACOS ANTIDEPRESSIVOS. PARTE II: ANÁLISE POR COMPONENTES PRINCIPAIS (PCA) E O MÉTODO DE CLASSIFICAÇÃO (SIMCA). Revista Brasileira de Ciencias farmacêuticas 40 (2004)

11. A.M. Souza, R.J. Poppi 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, 223-229, (2012).

12. A.M.Souza, M.R.Coelho, P.Figueiras, T.A.F. Cunha, R.O.Dart, J.G.Parés, P.L.Simon, B.G.Cruz, R.J.Poppi, M.M. Santos PROPOSTA DE TUTORIAL DE QUIMIOMETRIA UTILIZANDO TÉCNICAS MODERNAS PARA A ANÁLISE DE SOLOS. Sociedade Brasileira de Ciência do Solo. Sobral, CE (2012)

13. M.A.Morgano, C.G.Faria, M.F.Ferrão, N. Bragagnolo, M.M.Ferreira DETERMINAÇÃO DE PROTEÍNA EM CAFÉ CRU POR ESPECTROSCOPIA NIR E REGRESSÃO PLS. Revista de Ciênica e Tecnologia de Alimentos – Campinas 25 -31 (2005)

14. R.F.Teófilo, MÉTODOS QUIMIOMÉTRICOS: UMA VISÃO GERAL – CONCEITOS BÁSICOS DE QUIMIOMETRIA, Universidade Federal de Viçosa. Viçosa, Vol1 (2013)

15. Parreira, T.F; UTILIZAÇÃO DE MÉTODOS QUIMIOMÉTRICOS EM DADOS DE NATUREZA MULTIVARIADA. Univerisade Estadual de Campinas. Departamento de físico-química. Campinas (2003)

16. Santos, P.M; Wentzell, P.D; Pereira-Filho, E.R; SCANNER DIGITAL IMAGES COMBINED WITH COLOR PARAMETERS: A CASE STUDY TO DETECT ADULTERATIONS IN LIQUID COW'S MILK. Springer Science Business Media, LLC (2011)

17. F.A.Honorato, B.B.Neto, M.N.Martins, R.K.Galvão, M.F.Pimentel, TRANSFERÊNCIA DE CALIBRAÇÃO EM MÉTODOS MULTIVARIADOS. Química. Nova 20-45 (2007)

18. J. Paschoal, F.D.Barboza, R.J.Poppi, ANALYSIS OF CONTAMINANTS IN LUBRICANT OIL BY NEAR INFRARED SPECTROSCOPY AND INTERVAL PARTIAL LEAST-SQUARES, Journal of Near Infrared 2011-2019 (2003)

19. F.D.Barboza, R.J.Poppi, DETERMINATION OF ALCOHOLIC CONTENT IN BEVERAGES USING SHORT WAVE NEAR INFRARED SPECTROSCOPY AND TEMPERATURE CORRECTION BY TRANFER CALIBRATION PROCEDURES. Analitycal and Bioanalytical Chemistry 695-701 (2003)

20. Barros, I.; Tavares, M.; ESTIMATIVA DO TAMANHO ÓTIMO DE PARCELAS EXPERIMENTAIS ATRAVÉS DE CÁLCULOS ALGÉBRICOS. Bragantia, Campinas 54 (1) 209-215 (1995).