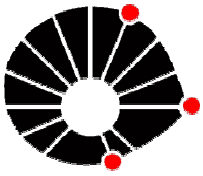
UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA
DEPARTAMENTO DE QUÍMICA ANALÍTICA
LAQQA – LABORATÓRIO DE QUIMIOMETRIA EM QUÍMICA ANALÍTICA



AVALIAÇÃO DE FIGURAS DE MÉRITO EM CALIBRAÇÃO MULTIVARIADA NA DETERMINAÇÃO DE PARÂMETROS DE CONTROLE DE QUALIDADE EM INDÚSTRIA ALCOOLEIRA POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO

DISSERTAÇÃO DE MESTRADO

PATRÍCIA VALDERRAMA

ORIENTADOR: Prof. Dr. RONEI JESUS POPPI

CAMPINAS – SP, JULHO DE 2005

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DO INSTITUTO DE QUÍMICA
DA UNICAMP

Valderrama, Patrícia.
V232a Avaliação de figuras de mérito em calibração multivariada na determinação de parâmetros de controle de qualidade em indústria alcooleira por espectroscopia no infravermelho próximo / Patrícia Valderrama. -- Campinas, SP: [s.n], 2005.

Orientador: Ronei Jesus Poppi.

Dissertação – Universidade Estadual de Campinas, Instituto de Química.

1. Calibração multivariada. 2. Infravermelho próximo. 3. Indústria alcooleira. I. Poppi, Ronei Jesus. II. Instituto de Química. III. Título.

Dedico este trabalho em especial à minha mãe ***ELIZABETT MARTINS VALDERRAMA,*** *ao meu pai* ***OSMAR VALDERRAMA*** *e aos meus irmãos* ***OSMAR ROGÉRIO VALDERRAMA*** *e* ***LEONARDO VALDERRAMA****...*

...Dedico este trabalho ainda, para as pessoas que fazem sorrir meu coração...

...Para aquelas pessoas que fizeram e as que fazem a diferença na minha vida...

...Para as pessoas que quando olho para trás sinto saudades...

...Para aquelas pessoas que me deram uma força quando eu não estava muito animada para o trabalho...

...Para as pessoas que amo e para as pessoas que um dia amei...

...Para as pessoas que encontro apenas em meus sonhos...

...Para as pessoas que se esqueceram o quanto foram importantes para mim, ou que talvez nunca souberam disso...

...Nossa vida é um caminho cheio de surpresas e incertezas, as quais ninguém é capaz de prever, por isso mais importante do que o que se tem na vida é quem temos na vida. Todas as pessoas importantes e que fizeram a diferença na minha vida possuem seu lugar no meu coração!!!

Aprendendo a Viver...

"Depois de algum tempo você aprende a diferença, a sutil diferença entre dar a mão e acorrentar uma alma. E você aprende que amar não significa apoiar-se e que companhia nem sempre significa segurança. E começa a aprender que beijos não são contratos e presentes não são promessas. E começa a aceitar suas derrotas com a cabeça erguida e olhos adiante, com a graça de um adulto e não com a tristeza de uma criança. E aprende a construir todas as suas estradas no hoje, porque o terreno do amanhã é incerto demais para os planos, e o futuro tem o costume de cair em meio ao vão. Depois de um tempo você aprende que o sol queima se ficar exposto por muito tempo. E aprende que não importa o quanto você se importe, algumas pessoas simplesmente não se importam... E aceita que não importa quão boa seja uma pessoa, ela vai feri-lo de vez em quando e você precisa perdoá-la por isso. Aprende que falar pode aliviar dores emocionais. Descobre que se leva anos para se construir confiança e apenas segundos para destruí-la, e que você pode fazer coisas em um instante, das quais se arrependerá pelo resto da vida. Aprende que verdadeiras amizades continuam a crescer mesmo a longas distâncias. E o que importa não é o que você tem na vida, mas quem você tem na vida. E que bons amigos são a família que nos permitiram escolher. Aprende que não temos que mudar de amigos se compreendemos que os amigos mudam, percebe que seu melhor amigo e você podem fazer qualquer coisa, ou nada, e terem bons momentos juntos. Descobre que as pessoas com quem você mais se importa na vida são tomadas de você muito depressa, por isso sempre devemos deixar as pessoas que amamos com palavras amorosas, pode ser a última vez que as vejamos. Aprende que as circunstâncias e os ambientes têm influência sobre nós, mas nós somos responsáveis por nós mesmos. Começa a aprender que não se deve comparar com os outros, mas com o melhor que pode ser. Descobre que se leva muito tempo para se tornar a pessoa que quer ser, e que o tempo é curto. Aprende que não importa onde já chegou, mas onde está indo, mas se você não sabe para onde está indo, qualquer lugar serve. Aprende que, ou você controla seus pensamentos e atos ou eles o controlarão, e que ser flexível não significa ser fraco ou não ter personalidade, pois não importa quão delicada e frágil seja uma situação, sempre existem dois lados. Aprende que heróis são pessoas que fizeram o que era necessário fazer, enfrentando as conseqüências. Aprende que paciência requer muita prática. Descobre que algumas vezes a pessoa que você espera que o chute quando você cai é uma das poucas que o ajudam a levantar-se. Aprende que maturidade tem mais a ver com os tipos de experiência que se teve e o que você aprendeu com elas do que com quantos aniversários você celebrou. Aprende que há mais dos seus pais em você do que você supunha. Aprende que nunca se deve dizer a uma criança que sonhos são bobagens. Poucas coisas são tão humilhantes e seria uma tragédia se ela acreditasse nisso. Aprende que quando está com raiva tem o direito de estar com raiva, mas isso não te dá o direito de ser cruel. Descobre que só porque alguém não o ama do jeito que você quer que ame, não significa que esse alguém não o ama com tudo o que pode, pois existem pessoas que nos amam, mas simplesmente não sabem como demonstrar ou viver isso. Aprende que nem sempre é suficiente ser perdoado por alguém, algumas vezes você tem que aprender a perdoar-se a si mesmo. Aprende que com a mesma severidade com que julga, você será em algum momento condenado. Aprende que não importa em quantos pedaços seu coração foi partido, o mundo não pára para que você o conserte. Aprende que o tempo não é algo que possa voltar para trás. Portanto, plante seu jardim e decore sua alma, ao invés de esperar que alguém lhe traga flores. E você aprende que realmente pode suportar... que realmente é forte, e que pode ir muito mais longe depois de pensar que não se pode mais. Aprende que nossas dúvidas são traidoras e nos fazem perder o bem que poderíamos conquistar, se não fosse o medo de tentar. E que realmente a vida tem valor e que VOCÊ tem valor diante da vida!"

William Shakespeare

AGRADECIMENTOS

- A Deus por ter me permitido alcançar mais esta vitória.
- Ao meu orientador Prof. Dr. Ronei Jesus Poppi pela oportunidade de realização desse trabalho, pela orientação, conhecimentos transmitidos, confiança, paciência, convivência e amizade.
- Ao grupo LAQQA: Jez William Batista Braga, Gilmare Antônia da Silva, Alessandra Borin, Waldomiro Borges Neto, Marcelo Garcia Trevisan, Luiz Carlos Moutinho Pataca, Tiago Pucca Araújo, Luciana Viviani, Joana Guilares de Aguiar, Danilo Althmann, Paulo Henrique Março. Agradeço pelo apoio, convivência, amizade e companheirismo. Agradecimento especial ao colega Jez por todos os conhecimentos transmitidos.
- À COCAMAR – Cooperativa Agroindustrial, por ter cedido os dados gerados por ocasião do meu trabalho como funcionária da empresa para a realização deste trabalho.
- À minha família: minha mãe Elizabett Martins Valderrama por todo esforço realizado para comigo, apoio, incentivo, compreensão, força, confiança e amor; Ao meu pai Osmar Valderrama pelo esforço realizado para comigo, apoio, incentivo, compreensão e amor; Ao meu irmão Osmar Rogério Valderrama pelo esforço realizado para comigo, apoio, incentivo e confiança; Ao meu irmão Leonardo Valderrama que embora muito jovem precisou entender e acostumar-se com minha ausência.
- Aos colegas da COCAMAR: Ageu Kopp dos Santos e Sidnei Leal pela oportunidade; Almir Guido Hawthorne e Aparecido Fadoni pela confiança; Brito, Aldair, Sedival, Davi, Cristiano, Rodrigo, Alécio, Jânio, Lourdes, Maico, Marcos, José Calegari, Gilberto, Juliana, Célio, José Cunha, Altair, Nilson, Luiz Carlos pela colaboração e apoio.
- À Alyadni Janaina Bassi Trento e Helton Cocci que na ocasião da realização da parte experimental deste trabalho foram meus estagiários na COCAMAR, agradeço pela colaboração, apoio e incentivo.
- À CAPES pelo financiamento do projeto.

- À Unicamp por fornecer toda estrutura física e tecnológica para a realização deste trabalho.
- Aos meus amigos (as) Maringaenses que tem convívio direto comigo aqui em Campinas: Juliana, Rafaelle, Adriano, Cris, Rúbia, Aline, Mariana, César, Silvana, Regiane, Emerson, Alessandra, Vamerson, D. Enilda, Leila, Odair, Fernanda pelo companheirismo e amizade.
- Aos meus amigos (as) distantes que, mesmo longe, sempre estiveram muito presente me apoiando e incentivando: Val, Rúbia, Fernanda, Munira, Daniela, Alyadni, Renata, Aline, Lígia, Juliano, Rodrigo, Humberto, Leandro.
- A todos aqueles que, direta ou indiretamente, tiveram sua parcela de participação durante a execução e conclusão deste trabalho.

CURRICULUM VITAE

Dados Pessoais
Nome – Patrícia Valderrama
Nascimento – 09/04/1980
Naturalidade – Japurá-PR
E-mail – patricialaqqa@iqm.unicamp.br

Formação
Graduação – Bacharelado em Química – 1998-2002
UEM – Universidade Estadual de Maringá – Maringá-PR

Experiência Profissional
- COCAMAR – Cooperativa Agroindustrial – 09/2002 a 09/2003
Função – pesquisa, desenvolvimento e implantação de metodologia de espectroscopia no infravermelho próximo para controle de qualidade da cana-de-açúcar.
- Steviafarma Industrial S/A – 01/2001 a 12/2001
Função – estagiária na área de controle de qualidade, desenvolvimento de novos produtos e processo industrial.

Atividades Acadêmicas
Monitoria Acadêmica – Programa de Estágio Docente II – 03/2005 a 07/2005
Disciplina – QG 109 – Química Geral Experimental
Curso – Química e Farmácia
Instituto de Química - UNICAMP

Publicações
- Valderrama, P.; Braga, J. W. B.; Poppi, R. J. "Figures of merit in near infrared spectroscopy and multivariate calibration: An aplication in the determination of quality parameters in alcohol industry" – Journal of Near Infrared Spectroscopy. Artigo submetido para publicação
- Valderrama, P.; Braga, J. W. B.; Poppi, R. J. "Figures of merit for the determination of quality parameters in sugar cane industry by near infrared spectroscopy and multivariate calibration" – Journal Brazilian Chemical Society. Artigo submetido para publicação
- Valderrama, P.; Clemente, E. "Isolation and thermostability of peroxidase isoenzymes from apple cultivars Gala and Fuji" – Food Chemistry, 87:601-606/2004.
- Valderrama, P.; Marangoni, F.; Clemente, E. "Efeito do tratamento térmico sobre a atividade de peroxidase (POD) e polifenoloxidase (PPO) em maçã (*Mallus comunis*)" – Ciência e Tecnologia de Alimentos, 21(3):321-325/2001.

Trabalhos em Eventos Internacionais
- Valderrama, P.; Braga, J.W.B.; Poppi, R.J. "Figures of merit for the determination of quality parameters in sugar cane industry by near infrared spectroscopy and multivariate calibration" – 12th International Conference on Near Infrared Spectroscopy, Auckland-New Zealand, 2005.
- Valderrama, P. Clemente, E. "Peroxidase (POD) e polyphenoloxidase (PPO) in apple (*Mallus comunis*)" – European Conference on Advanced Technology for Safe and High Quality Foods, Berlin-Germany, 2001.
- Valderrama, P.; Coqueiro, A.; Clemente, E. "Quality of uvaia fruit (*Pseudo myrcianthes pyriformis* (Camb.) *Klaus*) and pulp during freezing storage" -

European Conference on Advanced Technology for Safe and High Quality Foods, Berlin-Germany, 2001.
- Valderrama, P.; Clemente, E. "Análise da termoestabilidade de peroxidase (POD) e polifenoloxidase (PPO) em maçã (*Mallus comunis*) – 4º Simpósio Latino Americano de Ciência de Alimentos, Campinas-SP, 2001.
- Valderrama, P.; Goto, A. "Processo de purificação do extrato de estévia utilizando colunas de troca iônica" - 4º Simpósio Latino Americano de Ciência de Alimentos, Campinas-SP, 2001.
- Valderrama, P.; Clemente, E. "Characterisation of peroxidase (PPO) and poliphenoloxidase (POD) in apple (*Mallus comunis*)" – 8º Simpósio Internacional de Iniciação Científica da USP, São Carlos-SP, 2000.

Trabalhos em Eventos Nacionais
- Valderrama, P.; Braga, J.W.B.; Poppi, R.J. "Figuras de mérito em calibração multivariada na determinação de açúcares polarizáveis para indústria alcooleira utilizando espectroscopia no infravermelho próximo" – 13º Encontro Nacional de Química Analítica, Niterói-RJ, 2005.
- Valderrama, P.; Braga, J.W.B.; Poppi, R.J. "Identificação de amostras anômalas em calibração multivariada. Aplicação na determinação do Brix em caldo de cana por NIR" – 28º Reunião Anual da Sociedade Brasileira de Química, Poços de Caldas-MG, 2005.
- Valderrama, P.; Poppi, R.J. "Seleção de variáveis através de iPLS em calibração multivariada utilizando NIR para parâmetro de qualidade da indústria alcooleira" - 28º Reunião Anual da Sociedade Brasileira de Química, Poços de Caldas-MG, 2005.
- Valderrama, P.; Braga, J.W.B.; Poppi, R.J. "Figuras de mérito em calibração multivariada na determinação de sólidos solúveis em indústria alcooleira utilizando NIR" - 28º Reunião Anual da Sociedade Brasileira de Química, Poços de Caldas-MG, 2005.
- Valderrama, P.; Clemente, E. "Estudo da termoestabilidade de peroxidase (POD) e polifenoloxidase (PPO) em maçã(*Mallus comunis*)" – 4º Congresso Brasileiro de Engenharia Química em Iniciação Científica, Maringá-PR, 2001.
- Valderrama, P.; Clemente, E. "Isolamento e termo estabilidade de polifenoloxidase (PPO) e peroxidase (POD) em maçã (*Mallus comunis)*" – X Encontro Anual de Iniciação Científica, Ponta Grossa-PR, 2001.
- Valderrama, P.; Marangoni, F.; Clemente, E. "Caracterização de polifenoloxidase (PPO) e peroxidase (POD) em maçã (*Mallus comunis)*" – IX Encontro Anual de Iniciação Científica, Londrina-PR, 2000.

Participação em Eventos Nacionais
- Curso: "Aplicações analíticas da espectroscopia no infravermelho próximo - NIR" – São Paulo-SP, 2002.
- "VII Encontro regional sul de ciência e tecnologia de alimentos" – Curitiba-PR, 2001.
- "XVI Semana de química" – Maringá-PR, 2000.
- "1º seminário sobre gás natural e suas aplicações" – Maringá-PR, 2000.
- "VIII Encontro anual de iniciação científica" – Cascavel-PR, 1999.
- "XV Semana de química" – Maringá-PR, 1999.
- "VI Encontro de química da região sul" – Maringá-PR, 1998.
- "XIV Semana de química" – Maringá-PR, 1998.

RESUMO

AVALIAÇÃO DE FIGURAS DE MÉRITO EM CALIBRAÇÃO MULTIVARIADA NA DETERMINAÇÃO DE PARÂMETROS DE CONTROLE DE QUALIDADE EM INDÚSTRIA ALCOOLEIRA POR ESPECTROSCOPIA NO INFRAVERMELHO PRÓXIMO

Autora: Patrícia Valderrama
Orientador: Ronei Jesus Poppi

Sólidos solúveis (Brix), sacarose (Pol) e açúcares redutores (AR) são parâmetros importantes no controle de qualidade de indústrias alcooleiras visto que o pagamento dos produtores de cana-de-açúcar é feito a partir destes parâmetros. Assim, foi realizada a validação através da determinação de figuras de mérito para os modelos de calibração multivariada desenvolvidos a partir da espectroscopia de infravermelho próximo (NIR) na região de 1100-2500 nm por regressão de mínimos quadrados parciais (PLS) e na região de 1600-1850 nm correspondente à seleção de variáveis por regressão de mínimos quadrados parciais por intervalo (iPLS) para determinação destes parâmetros. Um total de 1003 e 378 amostras compôs os conjuntos de calibração e validação, respectivamente, sendo a divisão realizada pelo algoritmo de Kennard-Stone. A calibração foi otimizada pela eliminação dos *outliers*, com base nas amostras com *leverage* extremo, resíduos não modelados nos dados espectrais e resíduos não modelados na variável dependente. Para a validação, foram avaliados o *leverage*, os resíduos não modelados nos dados espectrais e os resíduos com base na repetibilidade espectral. Foram calculadas as figuras de mérito: exatidão, precisão, sensibilidade, sensibilidade analítica, seletividade, ajuste, razão sinal/ruído, limites de detecção e quantificação e intervalo de confiança. Os resultados obtidos, indicam que os modelos desenvolvidos podem ser utilizados na indústria alcooleira como uma alternativa à refratometria e medidas de polarização (metodologias padrão para determinação de Brix e Pol, respectivamente). Para o AR é necessária uma avaliação, por parte da indústria e produtores, da performance do modelo NIR em relação ao método padrão de titulação e a estimativa feita pela indústria até o momento.

ABSTRACT

AVALIATION OF FIGURES OF MERIT IN MULTIVARIATE CALIBRATION IN THE DETERMINATION OF QUALITY CONTROL PARAMETERS IN ALCOHOL INDUSTRY BY NEAR INFRARED SPECTROSCOPY

Author: Patrícia Valderrama
Adviser: Ronei Jesus Poppi

Soluble solids (Brix), sucrose (Pol) and Reducing Sugar (RS) are important properties in the quality control of alcohol industry to determine grower payment. Thus, the validation was achieved through determination of figures of merit for multivariate calibration models using near infrared spectroscopy (NIR) in region of 1100 - 2500 nm by partial least square regression (PLS). The region of 1600 – 1850 nm corresponds to variables selection by interval partial least square regression (iPLS). A total of 1003 and 378 samples constitute the calibration and validation sets, respectively, divided by Kennard-Stone algorithm. The calibration set was optimized by outliers elimination based on data with extreme leverage, unmodelled residuals in spectral data and unmodelled residuals in the dependent variable. For validation , besides the leverage and unmodelled residuals in spectral data was also evaluated residuals based on in spectral repeatability. The figures of merit such as accuracy, precision, sensitivity, analytical sensitivity, selectivity, adjust, signal-to-noise ratio, limits of detection and of quantification and confidence limit were calculated. The results obtained ondicate that the models developed can be used in the alcohol industry as an alternative to refractometry and lead clarification to polarization measurements (standard methods for Brix and Pol, respectively). For RS it is necessary to have evaluation by the industry and growers, to model NIR performed in relation to the standard method of titration and the estimate currently made by the industry.

LISTA DE SIGLAS

AOTF – Filtro Óptico Acústico Sintonizável (do inglês, Acousto-Optic Tunable Filter)

AR – Açúcar Redutor

ASTM – American Society for Testing and Materials

ATR – Açúcar Total Recuperável

Bias – Erro Sistemático

CONSECANA – Conselho dos Produtores de Cana-de-Açúcar, Açúcar e Álcool

CV – Validação Cruzada (do inglês, Cross validation)

EIV – Erros nas Variáveis (do inglês, Error in variables)

FAR – Infravermelho Distante (do inglês, Far Infrared)

IPLS – Mínimos Quadrados Parciais por Intervalo (do inglês, Interval Partial Least Square)

LD – Limite de Detecção

LQ – Limite de Quantificação

LS – Leitura Sacarimétrica

MSEC – Erro Médio Quadrático da Calibração (do inglês, Mean Square Error of Calibration)

$MSEC_p$ – Pseudo Erro Médio Quadrático da Calibração (do inglês, Pseudo Mean Square Error of Calibration)

MSECV – Erro Médio Quadrático da Calibração estimado por Validação Cruzada (do inglês, Mean Square Error of Cross validation)

MID – Infravermelho Médio (do inglês, Middle Infrared)

MLR – Regressão Linear Múltipla (do inglês, Multiple Linear Regression)

NAS – Sinal Analítico Líquido (do inglês, Net Analyte Signal)

NIPALS – Nonlinear Interative Partial Least Squares

NIR – Infravermelho Próximo (do inglês, Near Infrared)

PC – Componente Principal (do inglês, Principal Component)

PCs – Componentes Principais (do inglês, Principal Components)

PCA – Análise de Componentes Principais (do inglês, Principal Components Regression)

PCR – Regressão por Componentes Principais (do inglês, Principal Component Regression)

PCTS – Pagamento de Cana pelo Teor de Sacarose

PDF – Pseudograus de Liberdade (do inglês, Pseudo-Degress of Freedom)

PDS – Método para Selecionar Amostras (do inglês, Piecewise Direct Standardisation)

PLS – Mínimos Quadrados Parciais (do inglês, Partial Least Squares)

RHM – Método para Detecção de Amostras Anômalas (do inglês, Resampling by the Half-Means)

RMSEC – Raiz Quadrada do Erro Médio Quadrático da Calibração (do inglês, Root Mean Square Error of Calibration)

RMSECV – Raiz Quadrada do Erro Médio Quadrático de Validação Cruzada (do inglês, Root Mean Square Error of Cross Validation)

RMSEP – Raiz Quadrada do Erro Médio Quadrático de Previsão (do inglês, Root Mean Squares Error of Prediction)

RMSSR – Raiz Quadrada Média dos Resíduos Espectrais

SDV – Desvio Padrão dos Erros de Validação (do inglês, Standard Deviation of Validation)

SHV - Método para Detecção de Amostras Anômalas (do inglês, Smallest Half-Volume)

SPA – Método para Selecionar Amostras (do inglês, Successive Projections Algorithm)

UV – Ultra Violeta

UVE-PLS – Mínimos Quadrados Parciais com Eliminação de Variáveis não Informativas (do inglês, Elimination of Uniformative Variables in Partial Least Square)

VIS – Visível

VL – Variável Latente

VLs – Variáveis Latentes

V(PE) – Variância dos Erros de Previsão

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

PREFÁCIO

Prefácio

Na química analítica, determinações quantitativas normalmente fazem uso de alguma técnica instrumental em que a quantificação não é realizada de forma direta mas sim de uma forma indireta através de medidas físicas. Assim, para este tipo de determinação, faz-se necessário encontrar uma função que relacione o resultado da medida instrumental com a propriedade de interesse.

Essa relação entre o resultado da medida instrumental com a propriedade de interesse é conhecida como calibração. Quando se tem apenas um valor escalar registrado para cada amostra a calibração é dita de ordem zero. Por outro lado, quando os dados referentes a uma amostra podem ser arranjados na forma de um vetor a calibração é classificada como de primeira ordem e, finalmente, quando para uma amostra, é obtida uma matriz de dados instrumentais a calibração é dita de segunda ordem.

Dentre os métodos citados, a calibração de ordem zero, conhecida como calibração univariada, é o mais difundido sendo de aplicação relativamente fácil, porém restrita, tendo em vista que a amostra deve ser livre de interferentes. Para a calibração de primeira ordem, conhecida como calibração multivariada, a medida instrumental pode ser realizada mesmo na presença de interferentes, com a restrição de que estes estejam presentes no conjunto de amostras da calibração. Já para a calibração de segunda ordem, ou multi-modos a calibração pode ser realizada na presença de interferentes desconhecidos, fato conhecido como vantagem de segunda ordem. Estas calibrações ainda apresentam a possibilidade de determinações simultâneas e análises mesmo sem resolução, o que as tornam uma alternativa quando os métodos univariados não encontram aplicação.

Sempre que um procedimento analítico é proposto ou desenvolvido, existe a necessidade de se averiguar se o método apresenta a performance adequada para as condições nas quais ele será aplicado. A validação de um método estabelece, por estudos sistemáticos realizados em laboratório, que este atende ao seu propósito e às normas impostas por órgãos de fiscalização nacionais e internacionais, como por exemplo, American Society for Testing and Materials

(ASTM), Farmacopéias e o Conselho dos Produtores de Cana-de-Açúcar, Açúcar e Álcool (CONSECANA) o qual faz a regulamentação de acordo com cada estado. Essa validação pode ser atestada através da determinação de diversos parâmetros que são conhecidos como figuras de mérito. Dependendo de onde o método será aplicado, de seu propósito, ou a que órgão de fiscalização estará sujeito, a quantidade de figuras de mérito que devem ser determinadas ou o nível que deve ser atingido em cada uma delas pode variar.

Tendo em vista a dificuldade da validação em calibração multivariada, esse procedimento ainda é um fator limitante para suas aplicações. Entretanto, a atenção de diversos pesquisadores, órgãos de fiscalização e normatização vem sendo voltada nos últimos anos para o desenvolvimento de procedimentos e normas para a validação da calibração multivariada.

Motivando-se com os problemas apresentados esta Dissertação teve como objetivo validar modelos de calibração multivariada desenvolvidos pelo método de mínimos quadrados parciais (PLS – do inglês, Partial Least Squares) através da determinação das figuras de mérito. A aplicação foi focada a dados espectroscópicos na região do infravermelho próximo, gerados na indústria alcooleira. A escolha dessa aplicação deu-se devido à busca do setor por métodos alternativos de análise do caldo de cana tendo em vista o pagamento aos fornecedores da indústria. Isso tem a finalidade de aumentar a confiabilidade, uniformidade do método e a precisão das medidas, além do que determinações espectroscópicas na região do infravermelho próximo atualmente são regulamentadas pelo órgão de fiscalização do controle de qualidade da cana-de-açúcar.

A presente Dissertação foi dividida em 5 capítulos mais conclusões e referências bibliográficas. O primeiro capítulo, intitulado Cana-de-açúcar e indústria alcooleira, apresenta a principal matéria-prima empregada na fabricação industrial do álcool no Brasil, a metodologia empregada atualmente no controle de qualidade da cana-de-açúcar, a busca por metodologias alternativas e as principais tendências tecnológicas do setor.

O segundo capítulo, intitulado Espectroscopia no infravermelho, apresenta um histórico, as aplicações, os fundamentos teóricos e os aspectos instrumentais da espectroscopia no infravermelho próximo.

Métodos de análise multivariada consiste o assunto do terceiro capítulo, onde estão apresentados a quimiometria, os métodos de calibração e os métodos de regressão mais empregados para calibração de primeira ordem.

A validação dos modelos de calibração multivariada é proposta com base na determinação de figuras de mérito, que são descritas no quarto capítulo intitulado Validação e figuras de mérito, em que, para cada figura de mérito, é apresentada uma definição e o procedimento para sua determinação em modelos construídos empregando o método PLS.

A determinação quantitativa dos parâmetros de controle de qualidade da cana-de-açúcar está descrita no quinto capítulo intitulado Aplicação, onde o principal objetivo consiste na construção e validação de modelos de calibração PLS construídos com todas as variáveis do espectro de infravermelho próximo, assim como a construção e validação de modelos PLS com as variáveis selecionadas pelo método iPLS. Os modelos de calibração multivariada são propostos como um método alternativo à refratometria, medidas de polarização e titulação de oxidação-redução, metodologias padrão para determinação de Brix, Pol e AR, respectivamente. Os modelos multivariados são validados e comparados com base nos parâmetros apresentados no quarto capítulo, de forma a atestar sua performance.

Esta Dissertação encerra-se com as Conclusões do trabalho, avaliando os resultados para as estimativas das figuras de mérito bem como sua aplicação em substituição às metodologias padrão empregadas atualmente pelo setor industrial. Por fim, segue uma lista de Referências Bibliográficas em que são apresentados os trabalhos que contribuíram para a elaboração desta Dissertação.

CAPÍTULO 1
Cana-de-Açúcar e Indústria Alcooleira

1. Cana-de-açúcar e indústria alcooleira

1.1. Cana-de-açúcar

A cana-de-açúcar é uma planta da família das gramíneas composta de folhas, colmos, raízes e, eventualmente, flores. Cana é o termo genericamente aceito para designar os colmos industrializáveis da cana-de-açúcar, os quais são cortados na base, rente ao solo, despontados no último entrenó maduro e livres de impurezas oriundas da própria cana como plantas daninhas, terra, folhas, ponteiros, entre outras[1].

A matéria-prima entregue na indústria para a fabricação do álcool é composta por cana mais as impurezas carreadas com os colmos durante o carregamento mecanizado ou o corte, seguido pelo carregamento, realizado por colheitadeiras[1].

A composição da cana é extremamente variável em função de diversos fatores, como a idade cronológica e fisiológica da cultura, época de amostragem, variedade, estágio de corte, sanidade das plantas, condições climáticas durante o desenvolvimento e maturação, adubação e fertilização, tipo de solo, entre outros. Essa composição também é variável no sentido longitudinal e transversal da cana. Do ponto de vista tecnológico, a cana é constituída de caldo mais os sólidos insolúveis em água, os quais são representados pelas fibras. O caldo contém água mais os sólidos solúveis totais, que correspondem aos açúcares e não-açúcares sendo denominado de Brix. O principal componente da cana é a água, que pode chegar a 78% do seu peso no início do desenvolvimento vegetativo, decrescendo para 68% quando a cana atinge seu ponto máximo de maturação[1].

Os açúcares presentes na cana são representados principalmente pela sacarose, glicose e frutose e, na indústria, esses açúcares são denominados como Pol e açúcares redutores, respectivamente[2]. Em geral, para a cana madura o teor de açúcares redutores é baixo, menos que 0,5% comparado com o teor de sacarose que pode atingir altos teores, como por exemplo, até 17,5% em determinadas variedades, com análise de cana inteira, limpa e bem

despontada. Algumas novas variedades estão sendo produzidas em programas de melhoramento, sendo que nelas esse valor de Pol pode ser ultrapassado quando a cana se encontra no seu pico de maturação[1].

Durante a época do ano em que prevalecem temperaturas altas e a máxima atividade pluvial, a cana atinge um grande crescimento vegetativo. Ao terminarem as chuvas e com a diminuição da temperatura a síntese de sacarose na cana atinge níveis máximos que é denominado de maturidade tecnológica da cana. Este ciclo de crescimento e maturação se repete anualmente, num curso de 12 a 14 meses[3].

A cana-de-açúcar é cultivada em mais de cem países, sendo considerada a planta que possui mecanismos fisiológicos mais aperfeiçoados para a produção de sacarose, pois suas vias fotossintéticas para produzi-la, a partir dos açúcares simples, são mecanismos altamente eficientes, que o homem, através de um processo longo e continuado de melhoramento, vem aperfeiçoando e desenvolvendo até criar variedades comerciais com alto teor de sacarose e resistentes a doenças. Em alguns países, a cana é cortada ao término de um ciclo anual e em outros se faz depois de dois ciclos, com o objetivo de obter maior massa de cana por hectare. Por sua eficiência de assimilação da fotossíntese e capacidade de produzir massa verde, composta por açúcares, amidos, proteína e compostos lignocelulósicos, todas matérias-primas para um amplo campo de produtos de importância econômica, esta planta é uma das que possuem maiores qualidades entre as culturas comerciais. Para dar uma idéia, um hectare de cana por ano, com rendimento médio, é capaz de contribuir com 100 toneladas de matéria verde. Em termos de energia total, equivale a mais de 1.000 toneladas de petróleo, e considerando-a em termos de energia metabolizável, a 75.000 Mcal[3].

Os exemplos mencionados levam em consideração que os valores utilizados respondem às variedades comerciais em exploração. Estes índices podem ser superiores em se tratando de variedades desenvolvidas com propósitos dirigidos o que forneceria possibilidades de conseguir maior conteúdo de matéria verde, fibra ou açúcares, em menores períodos de tempo.

1.2. Controle de qualidade e pagamento de fornecedores

A qualidade da cana-de-açúcar como matéria-prima industrial pode ser definida como uma série de características intrínsecas da própria planta, alterada pelo manejo agrícola e industrial que definirão seu potencial para a produção do álcool[4].

O Brix é o parâmetro mais utilizado na indústria do álcool. Estritamente, expressa a porcentagem peso/peso dos sólidos solúveis contidos em uma solução pura de sacarose, ou seja, mede o teor de sacarose em soluções puras. No entanto, para o caldo de cana, representa a estimativa do teor de sacarose tendo em vista as demais impurezas que se encontram presentes. Este parâmetro não fornece qualquer informação qualitativa acerca dos açúcares presentes. A determinação quantitativa deste parâmetro pode ser realizada através de densímetros ou refratômetos, onde no caso do Brix areométrico o densímetro é calibrado com uma solução aquosa pura de sacarose a 20ºC e no caso do Brix refratométrico é medido o índice de refração de soluções de açúcar que fornecerão o próprio índice e/ou a porcentagem de sólidos solúveis da solução[4].

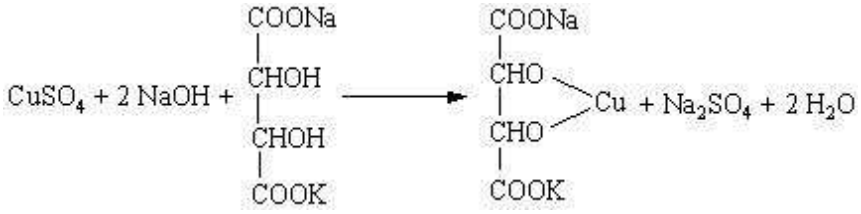
Um outro parâmetro que caracteriza a qualidade da cana-de-açúcar é o Pol ou a percentagem de sacarose no caldo da cana. Para a quantificação deste parâmetro o caldo da cana passa por um processo de clarificação com auxílio de uma mistura de acetato de chumbo e hidróxido de chumbo, também conhecido pela indústria como subacetato de chumbo ($Pb(CH_3COO)_2.Pb(OH)_2$). Após filtração através de papel de filtro pregueado, o grau de polarização das amostras é determinado através da leitura sacarimétrica (LS) utilizando um sacarímetro automático digital e a equação[4]:

$$\text{Pol}_{\%\ \text{caldo}} = \text{LS} \times (0{,}2605 - 0{,}0009882 \times \text{Brix}_{\%\ \text{caldo}}) \qquad (1)$$

Por fim, o último parâmetro utilizado pela indústria alcooleira para atribuir qualidade à cana-de-açúcar são os açúcares redutores, representados pela glicose e frutose. A determinação dos açúcares redutores é realizada através da metodologia padrão proposta por Eynon & Lane[4] que consiste na titulação de

oxidação-redução do licor de Fehling pelo caldo de cana filtrado através de algodão para eliminação de partículas suspensas. As substâncias redutoras do caldo de cana (glicose e frutose) reduzem o cobre de Cu^{2+} para CuO_2 do licor de Fehling, tendo como indicador do ponto final da titulação o azul de metileno a 1%[4].

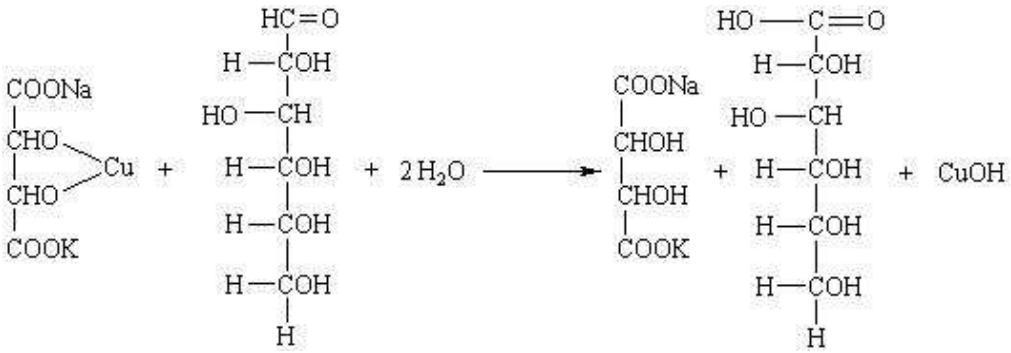
Normalmente, em meio alcalino o cobre sofre precipitação, no entanto, em presença de compostos orgânicos ricos em oxigênio como ácido cítrico, ácido tartárico, ou seus sais, a precipitação não ocorre devido à formação de um complexo com o cobre. Portanto, para evitar a precipitação do cobre no licor de Fehling, este é complexado com tartarato duplo de sódio e potássio conforme a reação abaixo[5]:



Tartarato de Sódio e Potássio

Complexo Tartarato-Cobre

O complexo formado pelos reagentes do licor de Fehling oxida os açúcares redutores do caldo da cana a ácido carboxílico durante a titulação conforme a reação[5]:



O teor de açúcares redutores presente em cada amostra é obtido através da equação (2)[4]:

$$AR_{\%caldo} = \frac{\left[\frac{5,2096 - \left(1,74993 x LS x \frac{Vg}{Vp}\right)}{500}\right]}{25,64 x \frac{Vg}{Vp} x (0,00398 x Brix + 0,99692)} \quad (2)$$

em que: Brix = Brix percento caldo

Vg = Volume de caldo de cana gasto na titulação

Vp = Volume padrão de solução de açúcar invertido a 1% gasto na titulação

LS = Leitura sacarimétrica

Estas equações foram propostas a partir de médias quinzenais de análise em várias safras para desenvolvimento do sistema de pagamento de cana pelo teor de sacarose (PCTS), até então, realizado por tonelada de cana-de-açúcar entregue. Na equação de quantificação do Pol, são utilizados fatores que "descontam" a quantidade de sólidos solúveis que não representam açúcares. Para a equação de quantificação dos açúcares redutores a parte superior da razão calculada consiste em um fator que considera a LS da sacarose na análise, enquanto que a parte inferior considera a massa específica do caldo, ou seja, a relação da massa com o volume e a densidade como sendo a relação da massa específica com a massa da água na temperatura padrão de 4ºC[1].

Tendo em vista que a produção industrial do álcool no Brasil utiliza a cana-de-açúcar como matéria-prima básica[6] e, ainda, com o desenvolvimento do sistema PCTS, atualmente, na formação do custo da tonelada de cana-de-açúcar pela indústria, a qualidade da cana consiste no principal parâmetro. A qualidade da cana-de-açúcar entregue pelo fornecedor na unidade industrial é apurada conforme a sua concentração em açúcar total recuperável em quilogramas por

tonelada (ATR), o qual é função do Pol, do Brix e do AR e mostrada na equação (3)[4]:

$$ATR = (10x0,88x1,0526xPol) + (10x0,88xAR) \quad (3)$$

em que: Pol = Pol % caldo, conforme a equação (1)

AR = AR % caldo, conforme a equação (2)

Para o pagamento dos fornecedores de cana-de-açúcar, os açúcares redutores não são determinados via análise, sendo apenas estimados através da seguinte equação que leva em consideração o Brix e o Pol % caldo[4]:

$$AR_{estimado} = \left(9,9408 - 0,1049x\left(\left(\frac{Pol}{Brix}\right)x100\right)\right) \quad (4)$$

A determinação deste parâmetro torna-se inviável para a indústria devido ao tamanho da amostragem necessária para cada fornecedor de acordo com a área colhida[4].

1.3. Métodos alternativos

Métodos alternativos de análise do caldo da cana para pagamento dos fornecedores vem sendo investigados e testados com a finalidade de aumentar a confiabilidade, uniformidade do método e também a precisão das medidas[4,7].

Atualmente, é regulamentado pelo conselho dos produtores de cana-de-açúcar, açúcar e álcool (CONSECANA) que o Brix, o Pol e o AR do caldo extraído poderão, também, ser determinados utilizando espectroscopia no infravermelho próximo (NIR), após a definição de modelos de calibração, construídos com os resultados da metodologia padrão. Entretanto, a implantação do NIR para pagamento dos fornecedores deverá ser aprovada pelo referido conselho, após a avaliação de um conjunto de pares de dados, superior a trezentos com valores do NIR e da metodologia convencional[4].

A busca por esses métodos alternativos tem incentivado a pesquisa e alguns trabalhos científicos apresentam propostas de novas técnicas e metodologias para determinações em indústrias do setor sucro-alcooleiro. Algumas determinações fazem uso da espectroscopia no infravermelho médio (MID) para determinações de açúcares como sacarose, glicose e frutose[8,9], outro faz uso da fluorescência de excitação e emissão para análise de cor na produção industrial do açúcar[10]. Foi investigada a previsão e determinação de adulteração de mel na indústria do açúcar através da espectroscopia no infravermelho próximo[11,12], análise de caldo de cana clarificado por subacetato de chumbo[13] e avaliação de parâmetros de qualidade no caldo de cana[7,14,15]. Entretanto, não se tomou conhecimento a respeito da validação de metodologias, utilizando a espectroscopia no infravermelho próximo e calibração multivariada, empregadas em indústrias do setor, que fazem uso de determinações de figuras de mérito. A validação é importante e necessária sempre que um procedimento analítico é proposto ou desenvolvido para averiguar se o método apresenta a performance adequada para as condições nas quais será aplicado.

1.4. Tendências tecnológicas

No mundo estão-se produzindo mudanças descontínuas no desenvolvimento da produção industrial, resultantes principalmente de inovações tecnológicas sob a influência da microeletrônica, da informática, das comunicações, dos novos materiais e dos novos conceitos em prática, que colocam como principal recurso o conhecimento[3]. Uma parte importante dos produtos fabricados atualmente tem cada vez menor valor, quanto a materiais e matérias-primas, enquanto crescem na sua composição os custos de *design*, apresentação ou inovação. Os produtos vão tendo um maior componente abstrato em seus custos, de tal forma que, conjuntamente como valor da produção, estejam os valores criados pela sua forma de comercialização, publicidade, serviços, pós-venda e outros[3].

Os sinais que caracterizam a indústria deste século são a sustentabilidade, uma alta flexibilidade para a mudança, a permanente inovação tecnológica e o conhecimento como o elemento central e mais importante. Estas são as

verdadeiras vantagens comparativas que os países possuem atualmente e futuramente[3]. No início dos anos 70, devido à crise energética, a produção do álcool atingiu o auge. O uso fundamental tem sido como substituto da gasolina, pois sua mistura aumenta a octanagem de forma adicional e permite reduzir o emprego de chumbo tetraetila com ação cancerígena. Além disso, a substituição total da gasolina pelo álcool permite reduzir nos gases de escape o monóxido de carbono e o óxido de nitrogênio, que são muito nocivos[3]. Desde então, tendências tecnológicas vêm abalando a indústria do setor sucro-alcooleiro, a qual vem aproveitando as vantagens de ser uma indústria auto-energética, capaz de não precisar de nenhum combustível externo para o seu processo e ainda ser capaz de gerar excedentes de eletricidade. Melhorias genéticas, dirigidas a propósitos específicos, vem sendo conseguidas em variedades da cana-de-açúcar, obtendo maiores teores de açúcar ou crescimentos em menores prazos, de tal forma que as suas características satisfaçam os requisitos das diferentes produções a que estarão destinadas. Inovações tecnológicas para garantir um melhor controle de qualidade da matéria-prima também consistem em uma destas tendências tecnológicas contribuindo para a nova indústria diversificada deste século[3].

CAPÍTULO 2
Espectroscopia no Infravermelho

2. Espectroscopia no Infravermelho

2.1. Histórico

A origem da espectroscopia no infravermelho data do início do século dezenove com o trabalho pioneiro do músico e astrônomo alemão Frederick William Herschel. A astronomia, de início um passatempo, passou a motivá-lo à realização de estudos sérios que consistiam essencialmente no mapeamento dos corpos celestes. Destes estudos resultaram a descoberta de várias estrelas e nebulosas. No entanto, sua grande descoberta no campo da astronomia foi o planeta Urano em 1781. O interesse pela astronomia despertou a curiosidade de Herschel com relação às propriedades físicas da radiação eletromagnética na região do visível, acreditando que a compreensão destas propriedades poderia ajudá-lo em seus estudos a respeito dos corpos celestes. Em 1800, durante a execução de um experimento que consistia na utilização de um prisma para separação das faixas espectrais associadas à região do visível, Herschel observou que uma das cores decompostas pelo prisma apresentava uma quantidade de energia distinta e para monitorar a quantidade de energia associada a cada cor utilizou um termômetro, verificando que abaixo do vermelho, onde não havia mais luz visível, era a região que apresentava maior temperatura[16].

O experimento realizado por Herschel foi importante, não somente pela descoberta da radiação infravermelha mas, também, por demonstrar que existem formas de luz que não podem ser observadas pelo olho humano. A partir de seus estudos, outros trabalhos foram desenvolvidos utilizando faixas espectrais como o infravermelho próximo, médio e distante e também a região do ultravioleta[16].

A espectroscopia na região do infravermelho alcançou grande desenvolvimento devido à potencialidade que a técnica apresentou na caracterização e quantificação de diferentes espécies químicas. Inicialmente, devido às limitações instrumentais, os trabalhos envolvendo espectroscopia em química restringiram-se basicamente à identificação e quantificação de algumas poucas espécies químicas em casos bem específicos[17]. Posteriormente, com o desenvolvimento de equipamentos mecânicos e ópticos mais precisos as aplicações foram ampliadas.

Em uma primeira fase, as aplicações qualitativas concentraram-se principalmente na faixa espectral do infravermelho médio, uma vez que nessa região é possível a observação de bandas de absorção de grupos orgânicos específicos como N-H, C-H, O-H, C-C, entre outros. Assim, esta técnica foi largamente empregada pelos químicos orgânicos para auxiliar a caracterização de diversas substâncias químicas[16].

A partir dos anos setenta uma nova fase de estudos espectroscópicos foi iniciada, agora também no campo das análises quantitativas, promovido pelo desenvolvimento dos espectrofotômetros com Transformada de Fourier, da informática, do interfaceamento de instrumentos eletrônicos com computadores e a introdução de recursos matemáticos mais sofisticados. Com essas inovações tecnológicas, os estudos quantitativos expandiram suas fronteiras para as regiões espectrais no infravermelho próximo e médio. Isso desencadeou um surpreendente interesse de vários grupos acadêmicos de pesquisas, governamentais e industriais, na tentativa de desenvolver metodologias de análise, trazendo como conseqüência o impacto direto no crescimento da produção e produtividade industrial, redução de gastos e da quantidade de resíduos industriais[16].

A primeira aplicação da espectroscopia no infravermelho foi para monitorar a qualidade e controlar a produção na indústria petroquímica, por ocasião da segunda guerra mundial. Os principais compostos monitorados foram combustíveis, lubrificantes e polímeros e os equipamentos utilizados, projetados pelas maiores companhias químicas da época como, Dow, Shell e Cyanamid, eram configurados somente para medidas de absorção na região do infravermelho médio[18].

A espectroscopia no infravermelho próximo não foi considerada inicialmente como uma técnica analítica com algum valor prático, sendo originalmente, uma extensão da região do visível que não foi explorada até por volta de 1970[18]. Os primeiros trabalhos que proporcionaram interesse pelo estudo da espectroscopia no infravermelho próximo como ferramenta de análise industrial foram desenvolvidos na década de setenta, pelo grupo de pesquisa do professor Karl

Norris, quando este era responsável por um grupo de pesquisa do Departamento de Agricultura dos Estados Unidos. Entretanto, as limitações tecnológicas da época não permitiram o desenvolvimento vertiginoso do NIR como atualmente é observado em diversos setores industriais como agrícola, petroquímico, alimentício e farmacêutico[19].

Atualmente, o valor, a funcionalidade e os benefícios oferecidos pela espectroscopia no infravermelho, principalmente a espectroscopia no infravermelho próximo, são incontestáveis[18].

2.2. Aplicações da espectroscopia no infravermelho

A região espectral que corresponde ao infravermelho compreende a radiação com números de onda no intervalo de aproximadamente 12800 a 10 cm^{-1}. Do ponto de vista da aplicação como dos instrumentos empregados, o espectro infravermelho é dividido em infravermelho próximo (NIR – do inglês, Near Infrared), médio (MID – do inglês, Middle Infrared) e distante (FAR – do inglês, Far Infrared). A Tabela 1 apresenta os limites aproximados para cada região[20].

Tabela 1. Regiões espectrais do infravermelho

Região	Intervalo de número de onda ($\hat{\nu}$) – (cm^{-1})	Região em comprimento de onda (λ) – (nm)	Região de freqüência (ν) – (Hz)
Próximo (NIR)	***12800 a 4000***	***780 a 2500***	***$3,8 \times 10^{14}$ a $1,2 \times 10^{14}$***
Médio (MID)	***4000 a 200***	***2500 a 5000***	***$1,2 \times 10^{14}$ a $6,0 \times 10^{12}$***
Distante (FAR)	***200 a 10***	***5000 a 100000***	***$6,0 \times 10^{12}$ a $3,0 \times 10^{11}$***

Fonte: Skoog, D. A.; Holler, F. J.; Nieman, T. A. *Princípios de análise instrumental*[20].

Na região do infravermelho próximo as principais aplicações encontram-se na análise quantitativa de materiais industriais e agrícolas e no controle de processos, destacando as aplicações farmacêuticas e petroquímicas, sendo também uma ferramenta valiosa para a identificação e determinação de aminas primárias e secundárias na presença de aminas terciárias em misturas. A princípio, as medidas eram somente realizadas em fotômetros e

espectrofotômetros dispersivos baseados em filtros e redes de difração, respectivamente. A configuração destes equipamentos era semelhante à de equipamentos que operavam na região do ultravioleta/visível (UV/VIS), sendo que, em muitos casos tratavam-se de equipamentos que compreendiam a região UV/VIS/NIR[20]. Atualmente, devido ao reconhecimento do potencial da aplicação do NIR principalmente nas análises quantitativas, equipamentos modernos, em sua maioria interferométricos com Transformada de Fourier, vêm sendo desenvolvidos especificamente para análises nesta região e já se encontram disponíveis acessórios para análises de amostras sólidas, líquidas e gasosas[16].

A espectroscopia no infravermelho próximo, além de fornecer os resultados de maneira mais rápida, é um método não destrutivo, assim como não gera subprodutos tóxicos e apresenta simplicidade na preparação de amostras, sendo que a maior desvantagem da técnica, é provavelmente, a baixa sensibilidade a constituintes em menores concentrações[16,20].

A região do infravermelho médio é provavelmente onde se encontra a maioria das pesquisas desenvolvidas e o maior número de aplicações. Esta região começou a ser utilizada no final dos anos 50 para a análise qualitativa de compostos orgânicos devido à grande quantidade de informação que pode ser utilizada para a caracterização funcional de compostos orgânicos. Para esta região, até o início dos anos 80, a maioria dos instrumentos era do tipo dispersivo baseados em redes de difração. A partir de então, com o surgimento dos equipamentos interferométricos, a maior parte dos instrumentos atuais é baseada na Transformada de Fourier. Essa mudança aumentou significativamente o número de aplicações do MID, tanto na área qualitativa como na quantitativa[20]. Entretanto, ainda hoje, a maioria das aplicações do MID consiste na identificação de compostos orgânicos pois nessa região ocorrem essencialmente transições fundamentais e existe uma faixa espectral conhecida como região de impressão digital (1200 a 700 cm^{-1}). Nessa região pequenas alterações na estrutura e na constituição de uma molécula resultam em mudanças significativas na distribuição dos picos de absorção do espectro que são relacionados com a estrutura da molécula. De posse destas informações, a identificação de compostos pode ser

realizada pela comparação do seu espectro MID com bancos de dados existentes[20].

A utilização da região do infravermelho distante teve seu uso limitado em tempos passados devido às limitações instrumentais, pois são poucas as fontes para este tipo de radiação e, ainda, para essa região, é necessária a utilização de filtros de interferência para evitar que radiações de ordens superiores atinjam o detector. O desenvolvimento dos espectrofotômetros com Transformada de Fourier resolve grande parte do problema encontrado nessa região e a tornou muito mais acessível para o desenvolvimento de aplicações e pesquisas. O FAR é útil principalmente para estudos de compostos inorgânicos, onde as absorções devido à vibrações de estiramento e deformação angular de átomos metálicos e ligantes, tanto inorgânicos como orgânicos, podem ser observados abaixo de 650 cm^{-1}. Moléculas compostas apenas por átomos leves também absorvem no FAR, desde que estas possuam modos de deformação angular da estrutura que envolva mais de dois átomos que não sejam o hidrogênio. Outras aplicações da região consistem ainda no estudo de gases que apresentam momentos de dipolo permanentes como por exemplo H_2O, O_3, HCl e AsH_3[20].

Para as regiões do infravermelho, em geral, é possível realizar medidas de amostras em todos os estados e formas como, gases, líquidos, sólidos, sistemas binários e terciários como as amostras semi-sólidas, pastas, géis e outras[18].

2.3. Princípios da espectroscopia no infravermelho

A radiação infravermelha não é suficientemente energética para causar transições eletrônicas e a absorção desta radiação está muito restrita a espécies moleculares que possuem diferenças de energia pequenas entre vários estados vibracionais e rotacionais. Para absorver radiação infravermelha a molécula precisa sofrer uma variação no momento de dipolo como conseqüência do movimento vibracional ou rotacional. Apenas nessas circunstâncias o campo elétrico alternado da radiação pode interagir com a molécula e causar variações na amplitude de um de seus movimentos. O momento dipolar é determinado pela magnitude da diferença de carga e a distância entre os dois centros de carga.

Quando uma molécula que possui essa variação do momento dipolar vibra, uma variação regular do momento dipolar ocorre e surge um campo que pode interagir com o campo elétrico associado à radiação. Se a freqüência da radiação coincidir exatamente com a freqüência vibracional natural da molécula, ocorre uma transferência de energia efetiva e resulta em uma variação da amplitude da vibração molecular e a conseqüência é a absorção de radiação. Do mesmo modo, a rotação de moléculas assimétricas em torno dos seus centros de massa resulta em uma variação periódica do dipolo que pode interagir com a radiação. Nenhuma variação efetiva no momento de dipolo ocorre durante a vibração ou rotação de uma molécula homonuclear, como O_2, N_2 ou Cl_2 e, conseqüentemente, essas substâncias não podem absorver no infravermelho[18,20].

A energia necessária para causar uma mudança de nível rotacional é pequena, da ordem de 100 cm^{-1} ou menor que 100000 nm. O espectro infravermelho de um gás consiste normalmente de uma série de linhas aproximadamente espaçadas, isso porque há vários estados rotacionais de energia para cada estado vibracional. Por outro lado, a rotação está rigorosamente restrita em líquidos e sólidos e nestas condições as linhas discretas vibracionais/rotacionais desaparecem deixando picos vibracionais alargados[20].

Em uma molécula, as posições relativas dos átomos não estão fixas variando continuamente em conseqüência dos tipos de vibrações e rotações em torno das ligações da molécula. Para uma molécula diatômica ou triatômica simples é possível definir com certa facilidade o número e a natureza de tais vibrações e relacioná-las às energias de absorção. Entretanto, para moléculas constituídas de muitos átomos possuindo um grande número de centros de vibração, como também para moléculas apresentando interações entre vários centros, é muito difícil definir o número e a natureza das vibrações envolvidas e relacioná-las à suas respectivas energias de absorção[20].

As vibrações são divididas em duas categorias: *estiramentos* e *deformações angulares.* Uma vibração de estiramento envolve uma variação contínua na distância interatômica ao longo do eixo da ligação entre dois átomos podendo acontecer de forma simétrica ou assimétrica, enquanto que as deformações

angulares são caracterizadas pela variação do ângulo entre duas ligações e podem acontecer no plano ou fora do plano da molécula. Além desses tipos de vibração, interações ou *acoplamentos* de vibrações podem ocorrer se as vibrações envolverem ligações de um mesmo átomo central e o resultado disso é uma variação nas características das vibrações envolvidas[20].

Considerando que as vibrações acontecem de forma isolada em uma molécula, estas podem ser representadas por um modelo mecânico simples e conhecido como oscilador harmônico. As características da vibração de estiramento pode se aproximar às de um modelo mecânico consistindo de duas massas ligadas por uma mola e as freqüências fundamentais de quaisquer dois átomos ligados podem ser calculadas assumindo que a energia segue o comportamento de um oscilador harmônico que obedece a Lei de Hooke. Assim as massas representam os átomos e a mola representa a ligação química entre eles. Uma perturbação de uma das massas ao longo do eixo da mola resulta em uma vibração denominada de movimento harmônico simples e a força restauradora é proporcional ao deslocamento (lei de Hooke) e tende a restaurar as massas para sua posição original. A freqüência da vibração é dada por[16]:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \tag{5}$$

em que: ν é a freqüência da vibração, k é a constante de força da ligação e μ é a massa reduzida dos dois átomos de massas m_1 e m_2, definida como:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \tag{6}$$

Este modelo funciona bem para o cálculo de freqüências fundamentais de moléculas diatômicas simples apresentando resultados que não ficam muito distantes dos valores médios encontrados para o estiramento de uma ligação entre dois átomos em uma molécula poliatômica. Contudo, essa aproximação

fornece apenas a média ou a freqüência central de transições de estados vibracionais e rotacionais de ligações diatômicas. Em moléculas poliatômicas os elétrons sofrem influencia de átomos ou grupos vizinhos e isso influencia o estiramento, o comprimento da ligação, o ângulo da ligação e conseqüentemente a freqüência da vibração das ligações químicas. Estas diferenças específicas que ocorrem devido a essas interações são o que proporcionam que cada substância apresente um espectro característico. Os valores da constante de força da ligação (k) variam muito e proporcionam diferenças de energia que podem ser utilizadas para a interpretação dos espectros[16].

O modelo descrito pela mecânica clássica apresentado até o momento, considera que as vibrações moleculares podem ter qualquer energia potencial prevendo com isso níveis de energia contínuos para as vibrações moleculares. No oscilador harmônico a energia potencial pode ser arbitrariamente considerada como sendo zero quando se trata da condição de repouso ou equilíbrio. À medida que a mola é comprimida ou esticada, a energia potencial do sistema aumenta de uma quantidade igual ao trabalho necessário para deslocar a massa e, assim, a variação da energia potencial será igual à força multiplicada pela variação da distância y, que depois de integrada entre a posição de equilíbrio e a distância esticada ou comprimida é dada por[20]:

$$E = \frac{1}{2} k y^2 \qquad (7)$$

A curva de energia potencial é uma parábola como ilustrada na Figura 1. (1), de onde se observa que a energia potencial é máxima quando a mola está esticada ou comprimida na amplitude A e decresce para zero na posição de equilíbrio[20].

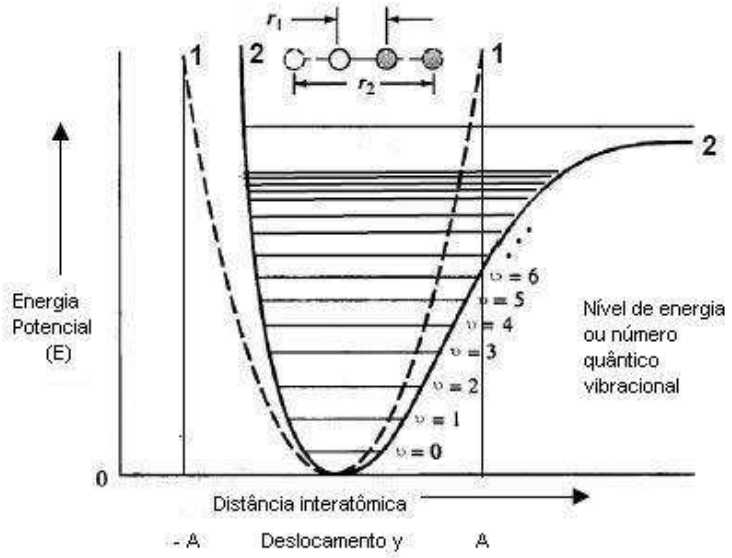


Figura 1. Diagrama de energia potencial. (1) oscilador harmônico, (2) oscilador anarmônico.

Para a mecânica quântica as vibrações moleculares podem ter apenas determinadas energias discretas e a partir do conceito do oscilador harmônico descrito pela mecânica clássica pode-se desenvolver as equações de onda da mecânica quântica. As soluções dessas equações para energia potencial têm a forma[20]:

$$E = \left(\nu + \frac{1}{2}\right)\frac{h}{2\pi}\sqrt{\frac{k}{\mu}} \qquad (8)$$

em que, h é a constante de Planck e ν é o número quântico vibracional que pode tomar valores positivos e inteiros incluindo o zero.

Para moléculas poliatômicas, os níveis de energia se tornam numerosos e uma aproximação pode tratar essas moléculas como uma série de osciladores harmônicos diatômicos e independentes. Nesse caso, a equação para a energia potencial pode ser generalizada como[16]:

$$E_{(\nu_1,\nu_{2...})} = \sum_{i}^{i=3N-6}\left(\nu_i + \frac{1}{2}\right)\frac{h}{2\pi}\sqrt{\frac{k}{\mu}} = \sum_{i}^{i=3N-6}\left(\nu_i + \frac{1}{2}\right)h\nu_m \qquad (9)$$

em que: 3N-6 é o número de vibrações possíveis em uma molécula com N átomos (para moléculas lineares existem 3N-5 vibrações possíveis), ν_m é a freqüência vibracional do modelo clássico e $\nu_{1,}\, \nu_{2,}$ ...= 0, 1, 2, ...[16].

A energia envolvida na transição do nível 1 para o nível 2 ou do nível 2 para o 3 deveria ser idêntica à da transição de 0 para 1 e mais, a teoria quântica indica que as únicas transições que poderiam ocorrer seriam aquelas em que o número quântico vibracional muda de uma unidade. Segundo a teoria quântica, portanto, somente transições fundamentais poderiam existir sendo esta restrição denominada de regra de seleção ($\Delta\nu = \pm 1$). Uma vez que os níveis vibracionais são igualmente espaçados, apenas um único pico de absorção deveria ser observado para uma certa vibração molecular[16,20].

A descrição da vibração molecular, considerando os tratamentos clássico e mecânico-quântico do oscilador harmônico, é imperfeita quando se levam em consideração aproximações qualitativas. Por exemplo, à medida que dois átomos se aproximam, a repulsão colombiana entre os dois núcleos produz uma força que age na mesma direção da força de restauração da ligação, assim, espera-se que a energia potencial cresça mais rapidamente do que é previsto pelo modelo do oscilador harmônico. Por outro lado, quando a distância entre os átomos aumenta, um decréscimo na força de restauração e, portanto, da energia potencial, ocorre quando a distância interatômica se aproxima daquela em que ocorre a dissociação dos átomos[20].

O comportamento anarmônico conduz a desvios de duas espécies. O primeiro deles é que os níveis de energia não são igualmente espaçados, como pode ser visto na representação da Figura 1.(2). A diferença entre os níveis de energia diminui à medida que a energia aumenta. A segunda é que em números quânticos altos, ΔE se torna menor e a regra de seleção não é seguida rigorosamente e como resultado, harmônicos que ocorrem em freqüências com

aproximadamente duas ou três vezes a de uma transição fundamental são observadas, isto é $\Delta v = \pm 2$ ou ± 3, sendo esse tipo de transição conhecida como sobretom (do inglês, "overtone"). A maior parte dos sinais do tipo sobretom ocorre na região do infravermelho próximo, e a intensidade dessas transições é cerca de 10 a 1000 vezes menor que as observadas para transição fundamental. Este fato é inconsistente com a teoria quântica, uma vez que esta prevê que tais transições seriam proibidas[16,20].

Além dos dois desvios descritos, às vezes são encontradas bandas de combinação quando um fóton excita simultaneamente dois modos vibracionais ocorrendo principalmente entre 5500 e 4000 cm^{-1}. A freqüência da banda de combinação é aproximadamente a soma ou a diferença das duas freqüências fundamentais. Esse fenômeno ocorre quando um quantum de energia é absorvido por duas ligações em vez de uma[17,20].

Na teoria, as equações de onda da mecânica quântica permitem a obtenção de curvas de energia potencial mais corretas para as vibrações moleculares, entretanto, a complexidade matemática dessas equações impede a aplicação quantitativa. Qualitativamente, as curvas devem tomar a forma anarmônica (Figura 1.(2)) e essas curvas se diferenciam do comportamento harmônico em alguns aspectos, dependendo da natureza da ligação e dos átomos envolvidos. No entanto, as curvas harmônicas e anarmônicas são muito parecidas para energias potenciais pequenas e isso explica o sucesso dos métodos de aproximação descritos[20,21].

Para a região da espectroscopia vibracional correspondente ao infravermelho próximo, os sinais observados são essencialmente devido a sobretom e bandas de combinação de estiramentos e deformações angulares de transições fundamentais de ligações X-H, em que X representa átomos de oxigênio, nitrogênio, carbono, grupos aromáticos e também grupos funcionais importantes como os C-O, grupos carbonila, C-N, C-C, entre outros que sofrem estiramento[17].

2.4. Instrumentos para espectroscopia no infravermelho

A Figura 2 mostra um esquema básico dos componentes principais de um equipamento de infravermelho:

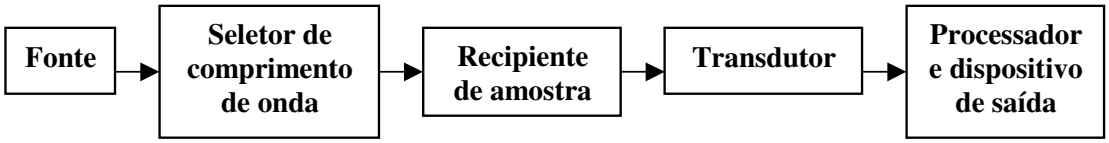


Figura 2. Componentes básicos de um equipamento que opera na região do infravermelho.

As fontes de infravermelho consistem em um sólido inerte como a fonte de Nernst que é composta de óxido de terras raras, a fonte Globar que consiste em uma barra de carbeto de silício, a fonte de filamento incandescente que é representada através de um espiral de níquel-cromo, o arco de mercúrio, a lâmpada de filamento de tungstênio, o laser de dióxido de carbono, são exemplos de fontes para a região do infravermelho que encontram aplicações de acordo com o interesse específico ou a região NIR/MID/FAR[18,20].

A princípio, os equipamentos de infravermelho não diferem muito dos equipamentos utilizados nas regiões do ultravioleta e do visível no que diz respeito aos componentes básicos. Contudo, diferenças notórias são encontradas principalmente na fase de seleção de comprimento de onda, que originam espectrofotômetros dispersivos ou interferométricos ou, ainda, fotômetros que são equipados com filtros[18,20].

Os recipientes para a amostra variam de acordo com a região NIR/MID/FAR e com o estado físico da amostra[20].

Os transdutores para o infravermelho são de três tipos gerais:

1. Transdutores térmicos: a resposta depende do efeito de aquecimento da radiação sendo comumente encontrados em fotômetros e espectrofotômetros dispersivos.
2. Transdutores piroelétricos: consiste em um transdutor térmico muito especializado com propriedades térmicas e elétricas especiais e são encontrados em fotômetros e espectrofotômetros dispersivos.

3. Transdutor fotocondutor: consiste na absorção da radiação pelos materiais componentes do transdutor, essa absorção promove elétrons não-condutores de valência a um estado condutor, de energia mais elevada, decrescendo a resistência elétrica do semicondutor. A composição dos materiais componente do transdutor varia de acordo com a região NIR/MID/FAR. Esse tipo de transdutor é encontrado em instrumentos multiplexados com transformada de Fourier[20].

2.4.1. Espectrofotômetros dispersivos

Os espectrofotômetros dispersivos no infravermelho são instrumentos registradores de feixe duplo, que usam redes de difração para dispersar a radiação. Essa é uma característica importante devido à baixa intensidade das fontes de infravermelho, à baixa sensibilidade dos transdutores para infravermelho e à conseqüente necessidade de grandes amplificações de sinal[18,20].

Geralmente, este tipo de equipamento incorpora um modulador de baixa freqüência que permitirá ao detector distinguir entre o sinal da fonte e sinais de radiação espúria, tal como emissão infravermelha ao redor do transdutor. Baixas velocidades de modulação são requeridas devido ao tempo de resposta lento dos transdutores para infravermelho usados na maioria dos instrumentos dispersivos. Em geral, a óptica dos instrumentos dispersivos é muito parecida com a de equipamentos no UV/VIS, exceto que os compartimentos de amostra e referência estão sempre localizados entre a fonte e o monocromador nos instrumentos no infravermelho. Essa disposição é possível porque a radiação infravermelha, em contraste com a UV/VIS, não é energética o suficiente para causar decomposição fotoquímica da amostra. Colocando a amostra e a referência antes do monocromador, tem-se a vantagem de que a maior parte da radiação espalhada, gerada no compartimento da célula, é efetivamente removida pelo monocromador e não atinge o transdutor[20].

2.4.2. Espectrofotômetro com Transformada de Fourier

Esse tipo de equipamento utiliza um interferômetro ao invés de redes de difração. Sua popularização ocorreu com o surgimento de microcomputadores, com o seu interfaceamento e com a utilização da transformada de Fourier para o tratamento de dados. Espectrofotômetros com Transformada de Fourier apresentam algumas vantagens, como a grande eficiência no transporte da radiação até o detector, o que melhora a relação sinal/ruído, apresenta um alto poder de resolução e reprodutibilidade do comprimento de onda, todos os elementos de resolução para um espectro são medidos simultaneamente o que possibilita a aquisição de dados de um espectro inteiro em cerca de um segundo ou menos[18,20].

A maioria dos equipamentos utilizados atualmente, utilizam o princípio interferométrico, sendo que diante de tantas vantagens, a existência e utilização de equipamentos dispersivos para a região NIR é justificada pelo fator custo e por se tratar de uma região onde os sinais observados são em sua maioria bandas relativamente largas, em que uma grande resolução não é requerida na maioria das vezes, uma vez que as aplicações geralmente são quantitativas[16,18].

2.4.3. Instrumentos não-dispersivos

Estes instrumentos são simples e robustos e foram projetados para análise quantitativa. Alguns são fotômetros de filtro único ou não-dispersivo, outros usam filtros no lugar de um elemento dispersivo para a obtenção de um espectro completo, outros ainda, não fazem uso de qualquer dispositivo para seleção de comprimento de onda. Em geral, este tipo de instrumento é menos complexo, mais robusto, fácil de manter e menos caro que os instrumentos descritos anteriormente.

Os fotômetros de filtro podem utilizar filtros de interferência e são normalmente designados para análises quantitativas. Uma variedade de filtros é disponível atualmente, cada um indicado para uma aplicação específica[20]. Um exemplo que faz uso deste tipo de equipamento é a análise quantitativa de substâncias orgânicas na atmosfera[22]. Um outro tipo de filtro que pode ser

utilizado pelos fotômetros são os filtros ópticos acústicos sintonizáveis (AOTF – do inglês, Acousto-Optic Tunable Filter) que podem efetuar uma varredura espectral como a que é feita através de redes de difração e, quando comparados a equipamentos dispersivos, apresentam alto rendimento, resolução e velocidade de varredura[18].

Fotômetros que não utilizam qualquer dispositivo de seleção de comprimento de onda são amplamente usados para monitorar fluxos de gases para um único componente[23]. Este tipo de instrumento é altamente seletivo pois possui um gás sensor que é aquecido apenas com a estreita porção do espectro que é absorvida pelo monóxido de carbono na amostra, podendo ser adaptado para a determinação de qualquer gás que absorva no infravermelho[20].

2.4.4. Instrumentos para espectroscopia no infravermelho próximo

Os instrumentos para a região do infravermelho próximo são semelhantes aos utilizados para a espectroscopia de absorção no UV/VIS. Lâmpadas de tungstênio/halogênio com janelas de quartzo servem como fontes, as células para medidas de absorção são normalmente de quartzo ou sílica fundida, transparentes até 3.000 nm. O caminho óptico varia de 0,1 a 10 cm. Os detectores são, em geral, fotocondutores de sulfeto de chumbo[20].

Estão disponíveis uma série de fotômetros e espectrofotômetros projetados especificamente para a região do infravermelho próximo. A variedade de instrumentos é muito grande, indo dos mais sofisticados com Transformada de Fourier, aos de feixe duplo com redes de difração ou arranjo de diodos, até os instrumentos mais simples baseados em filtros[16,24,25].

CAPÍTULO 3
Métodos de Análise Multivariada

3. Métodos de análise multivariada

3.1. Quimiometria

A quimiometria pode ser definida como a pesquisa e utilização de métodos matemáticos e estatísticos para o tratamento de dados químicos de forma a extrair uma maior quantidade de informações e melhores resultados analíticos. Os métodos utilizados na quimiometria, a princípio, foram desenvolvidos em outras disciplinas que com a aplicação e pesquisas voltadas para o tratamento de dados químicos acabou dando origem a uma nova área dentro da química analítica. Isso se deu após a segunda metade dos anos 60, com o surgimento de métodos instrumentais computadorizados para a análise química que promoveu a geração de uma grande quantidade de dados. Até este período, os químicos baseavam suas decisões em uma pequena quantidade de dados que, na maioria das vezes, eram obtidos de forma lenta e dispendiosa. A partir dos anos 60, com a grande quantidade de dados de obtenção rápida e com menor esforço, foi preciso analisar todos esses dados e extrair maior quantidade de informações relevantes. Foi então, que teve início a pesquisa e utilização dos métodos matemáticos e estatísticos que acabaram resultando nessa nova área conhecida como quimiometria[26].

A quimiometria pode ser considerada uma das áreas mais recentes da química analítica. Desde o seu surgimento no final dos anos 60 até os dias de hoje, foram desenvolvidos muitos métodos que tem tornado possível o processamento e interpretação de dados que antes seriam impossíveis de serem analisados. Um exemplo importante do sucesso da utilização da quimiometria são as análises realizadas na região do infravermelho próximo, as quais, sem a utilização de modelos de calibração multivariada não apresentam possibilidades para determinações quantitativas[26].

Com o crescimento da quimiometria foram desenvolvidas novas ferramentas para tratamento de dados encontrando aplicações distintas conforme o objetivo do estudo como, por exemplo, a otimização de processos, a classificação de dados,

as determinações quantitativas, entre outros. Assim, a quimiometria foi dividida em diversas frentes de pesquisa e aplicação:

- Processamento de sinais analíticos
- Planejamento e otimização de experimentos
- Reconhecimento de padrões e classificação de dados
- Calibração multivariada
- Métodos de inteligência artificial[26].

Dentro dessa divisão, a principal linha de pesquisa da quimiometria aplicada à química analítica tem sido a construção de modelos de regressão a partir de dados de primeira ordem, ou seja, dados que podem ser representados através de um vetor para cada amostra, sendo a construção desses modelos denominada de calibração multivariada[27].

3.2. Calibração

A calibração pode ser definida como uma série de operações que estabelecem, sob condições específicas, uma relação entre medidas instrumentais e valores para uma propriedade de interesse correspondente[28].

Um modelo de calibração, na verdade, é uma função matemática que relaciona dois grupos de variáveis, uma delas denominada dependente (Y) e a outra denominada independente (X):

$$Y=f(X) = Xb \qquad (10)$$

Esta etapa representa a calibração e, por isso, o conjunto de dados empregado para essa finalidade é chamado de conjunto de calibração. Os parâmetros do modelo são denominados de coeficientes de regressão (b) determinados matematicamente a partir dos dados experimentais[29,30].

O passo seguinte à calibração é a validação. Nesta etapa, as variáveis independentes obtidas, para um outro conjunto de amostras, são utilizadas em conjunto com o coeficiente de regressão, para calcular os valores previstos para a variável dependente. No conjunto de validação utilizam-se amostras cujas

variáveis dependentes sejam conhecidas para que seja possível estabelecer uma comparação entre os valores previstos pelo modelo e os valores conhecidos previamente através de uma metodologia padrão, o que permitirá a avaliação sobre o desempenho do modelo de calibração proposto[30].

Existem diversos métodos para a construção de modelos de calibração, sendo que a função que ajusta as variáveis dependentes e independentes pode ser linear ou não, dependendo da complexidade do sistema em estudo[31].

Dentre os métodos de calibração existentes, sem dúvida, os mais difundidos são ainda os métodos de calibração univariada que também são conhecidos como calibração de ordem zero, ou seja, tem-se apenas uma medida instrumental para cada uma das amostras de calibração, isto é, para cada amostra tem-se apenas um escalar. Esses métodos são descritos na literatura em vários trabalhos[32,33,34,35] e sua aplicação e validação são relativamente fáceis. No entanto, a aplicação da calibração univariada é restrita, visto que, quando a amostra não é livre de interferentes e a medida é realizada diretamente na metodologia instrumental, isso provavelmente provocará desvios na determinação da propriedade de interesse e a aplicação deste método de calibração torna-se inviável.

Em calibração multivariada, mais de uma resposta instrumental é relacionada com a propriedade de interesse. Esses métodos de calibração possibilitam a análise mesmo na presença de interferentes, desde que esses interferentes estejam presentes nas amostras utilizadas para a construção do modelo de calibração. Outras possibilidades apresentadas por este tipo de calibração são determinações simultâneas e análises mesmo sem resolução. Isso faz com que os modelos de calibração multivariada sejam uma alternativa quando os métodos univariados não encontram aplicação[30]. Neste tipo de calibração a resposta instrumental é representada na forma de matriz, enquanto a propriedade de interesse, determinada por uma metodologia padrão, é representada por um vetor. A Figura 3 ilustra como uma matriz de dados pode ser construída a partir de um vetor de respostas instrumental.

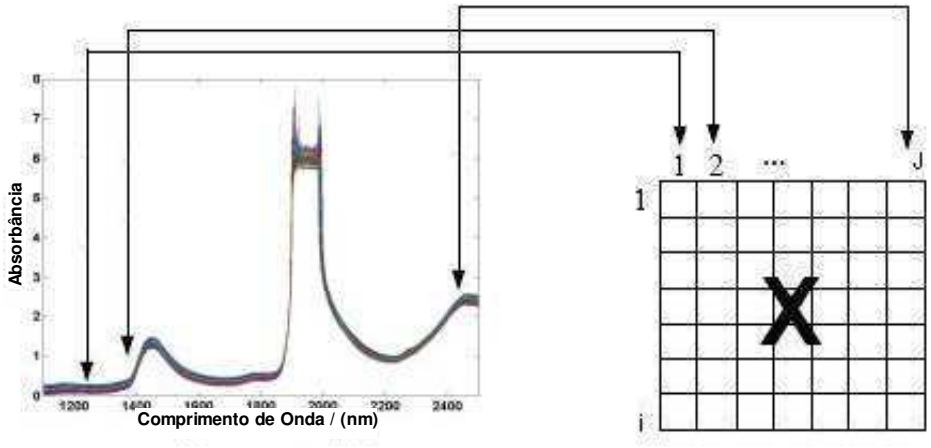


Figura 3. Construção da matriz **X** para calibração multivariada.

Uma diversidade de métodos de regressão vem sendo utilizado em química analítica para a construção de modelos de calibração multivariada, dentre esses os mais empregados tem sido a regressão linear múltipla (MLR), regressão por componentes principais (PCR) e regressão por mínimos quadrados parciais (PLS), que são métodos para ajuste linear entre as variáveis. Tem-se verificado que a maioria dos métodos de calibração multivariada empregados em espectroscopia utiliza ajuste linear entre as variáveis, uma vez que este representa o modelo de mais fácil elaboração e interpretação.

3.3. Métodos de Regressão

3.3.1. Regressão Linear Múltipla - MLR

O modelo mais simples em calibração multivariada consiste na resolução de um sistema de equações lineares em uma regressão linear múltipla (MLR – do inglês, Multiple Linear Regression)[26,36,37,38]. Este modelo de calibração inversa, para a determinação de um analito, pode ser obtido a partir de uma matriz **X** de respostas instrumentais com dimensão (i x j), arranjada de acordo com a Figura 3, onde i representa o número de amostras a ser utilizada na construção do modelo (conjunto de calibração) e j representa o número de variáveis, e um vetor **y** de

dimensão (i x 1), que contém as concentrações de referência das amostras de calibração. A partir daí cada variável dependente de **y** é expressa como uma combinação linear das variáveis independentes da matriz **X** e um vetor que contém os coeficientes de regressão, $\mathbf{b}_{MLR}$, relaciona **X** e **y** por meio da expressão[27]:

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{MLR} \tag{11}$$

em que $\mathbf{b}_{MLR}$ é um vetor de dimensão (j x 1).

Para a obtenção do vetor dos coeficientes de regressão, esta equação pode ser resolvida por mínimos quadrados[27]:

$$\hat{\mathbf{b}}_{MLR} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{12}$$

em que, os índices sobrescritos –1 e T representam a inversão e transposição de uma matriz ou vetor, respectivamente.

Uma estimativa para a concentração da espécie de interesse ($\hat{\mathbf{y}}$) pode ser obtida por[27]:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}_{MLR} \tag{13}$$

A regressão linear múltipla apresenta dois problemas que limitam sua aplicação. O primeiro deles é que o número de amostras deve ser igual ou superior ao número de variáveis, uma vez que o modelo consiste na resolução de um sistema de equações lineares simultâneas quando o número de variáveis é superior ao número de amostras, ou vice-versa, o sistema de equações a ser resolvido torna-se indeterminado. O segundo problema constatado para a MLR é

que a matriz **(X^TX)** pode não apresentar inversa devido a alta correlação entre as variáveis[27].

3.3.2. Regressão por Componentes Principais – PCR

A regressão por componentes principais (PCR- do inglês, Principal Components Regression) surge como uma alternativa no intuito de contornar os problemas apresentados pela MLR. Neste método de regressão utiliza-se a análise de componentes principais (PCA – do inglês, Principal Component Analysis) como a técnica de ortogonalização baseada em mudança de base vetorial. Este procedimento resolve os dois principais problemas da MLR, uma vez que a PCA pode ser utilizada para a redução do número original de variáveis sem acarretar na perda significativa de informação resolvendo, assim, o problema de existência de alta colinearidade entre as colunas de **X** e a necessidade de um número excessivo de amostras para a construção do modelo por MLR[26,27,36,37,38].

O primeiro passo para a análise de componentes principais é a formação de uma matriz de variância/covariância dos dados (**Z**) que irá isolar a fonte de variação dos dados:

$$\mathbf{Z}=\mathbf{X}^T\mathbf{X} \tag{14}$$

A matriz de covariância é, então, diagonalizada por uma transformação unitária:

$$\mathbf{\Lambda} = \mathbf{P}^{-1}\mathbf{Z}\mathbf{P} \tag{15}$$

em que $\mathbf{\Lambda}$ é uma matriz diagonal cujos elementos são autovalores de **Z**, **P** é a matriz de autovetores, denominada *loadings* (pesos). Basicamente, os *loadings* formam uma nova base ortonormal que explica a variância dos dados de **X** e a projeção dos dados nessa base é denominada *scores* (escores), (**T**). Desse modo, os dados são decompostos por um conjunto de vetores *loadings* e *scores*:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{T} \quad (16)$$

O conjunto *loadings* e *scores* é denominado componente principal (PC). A Figura 4 ilustra a decomposição da matriz **X** de dimensão (n x m) pela análise de componentes principais até A componentes principais.

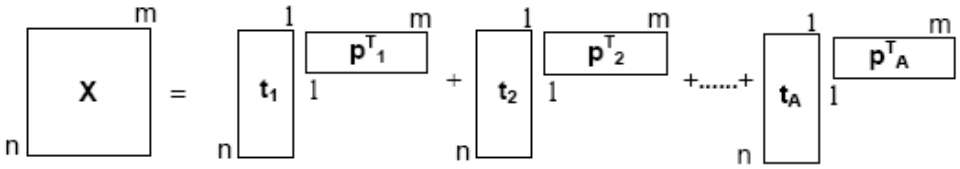


Figura 4. Decomposição em componentes principais por PCA.

O número máximo de componentes principais obtidos (PCs) é igual ao número de vetores de dados utilizados (posto da matriz **X** de dados independentes), sendo que, nem todas as PCs possuem informações úteis. Normalmente, as últimas PCs modelam ruído inerente aos dados. Sendo assim, a eliminação das PCs freqüentemente aumenta a relação sinal/ruído[26,36,37,38]. Para a determinação do número correto de PC o método mais utilizado consiste no método de Validação Cruzada (CV – do inglês, Cross Validation), o qual se baseia na habilidade de previsão de um modelo construído por parte de um conjunto de dados seguido pela previsão do restante do conjunto de dados, que é realizada pelo modelo construído. A validação cruzada pode ser realizada em blocos, ou seja, um número determinado de amostras é deixado de fora no processo de construção do modelo e a seguir essas amostras são previstas pelo modelo construído, ou ainda por um caso conhecido como “leave one out” (deixe um fora), onde uma amostra é deixada de fora no processo de construção do modelo e a seguir essa amostra é prevista pelo modelo construído. Em ambos os casos, o processo é repetido até que todas as amostras tenham sido previstas e a raiz quadrada da soma do quadrado dos erros da validação cruzada (RMSECV – do inglês, Root Mean Square Error of Cross validation) é calculada[26,27]:

$$RMSECV = \frac{\sum_{i=1}^{k} RMSEPi}{k} \quad (17)$$

em que, k é o número de blocos e o RMSEP (do inglês, Root Mean Square Error of Prediction) é calculado como:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad (18)$$

em que $\hat{y}_i$ e y_i são os valores previstos e de referência para a propriedade de interesse, respectivamente e n é o número de amostras de calibração.

O cálculo é realizado para o número de componentes de 1 até A, e os resultados de RMSECV são plotados em um gráfico em função do número de PCs. O comportamento típico para esses gráficos é a observação de um mínimo ou um patamar, que indica a melhor dimensionalidade do modelo de regressão, ou seja, o melhor número de PCs que produziu o menor erro de previsão sem perda significativa da variância dos dados[26,27] .

Idealmente, o número de PC deveria ser igual ao número de espécies químicas presentes na amostra. Isso permite que técnicas quimiométricas, que empregam PCA, possam ser utilizadas em circunstâncias onde se deseja determinar apenas algumas espécies de interesse em um meio complexo. Essa propriedade também é referida como vantagem de primeira ordem, que faz com que interferentes na amostra possam ser modelados, desde de que estejam presentes no desenvolvimento do modelo. Assim, a seletividade do sinal analítico deixa de ser essencial, como é nos modelos univariados[27].

PCR procede a uma regressão para converter os scores **T** nas concentrações **y**, que pode ser representada por:

$$\mathbf{y}=\mathbf{T}\boldsymbol{\beta}_{PCR} \quad (19)$$

em que, $\boldsymbol{\beta}_{PCR}$ possui dimensão (A x 1), A é o número de PCs escolhido.

O vetor $\boldsymbol{\beta}_{PCR}$ pode ser obtido então, por mínimos quadrados[27]:

$$\hat{\boldsymbol{\beta}}_{PCR} = \left(\mathbf{T}^{T}\mathbf{T}\right)^{-1}\mathbf{T}^{T}\mathbf{y} \quad (20)$$

Na prática, a obtenção das concentrações estimadas pelo modelo é feita pela multiplicação da matriz de dados instrumentais por um vetor de coeficientes de regressão, $\mathbf{b}_{PCR}$**.** O vetor $\mathbf{b}_{PCR}$ que estabelece essa relação direta é obtido por[27]:

$$\hat{\mathbf{b}}_{PCR} = \mathbf{P}\boldsymbol{\beta}_{PCR} \quad (21)$$

Através do vetor $\mathbf{b}_{PCR}$ a concentração da espécie de interesse $\hat{y}$ pode ser estimada diretamente de[27]:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}_{PCR} \quad (22)$$

3.3.3.Regressão por Mínimos Quadrados Parciais – PLS

A regressão por mínimos quadrados parciais (PLS – do inglês, Partial Least Squares) é considerada o método de regressão mais utilizado para a construção de modelos de calibração multivariada a partir de dados de primeira ordem. Este método, assim como o PCR, não requer um conhecimento exato de todos os componentes presentes nas amostras podendo realizar a previsão de amostras mesmo na presença de interferentes, desde de que estes também estejam presentes por ocasião da construção do modelo (vantagem de primeira ordem)[27].

Para o método de regressão PCR, a decomposição da matriz **X** realizada pelo PCA é feita de forma independente do vetor **y**, enquanto que para o método de regressão PLS a informação de **y** é incorporada, de forma que cada PC do modelo

sofre uma pequena modificação para buscar a máxima covariância entre **X** e **y** e passa a receber a terminologia de Variável Latente (VL)[26].

O modelo PLS é obtido através de um processo iterativo, no qual se otimiza ao mesmo tempo a projeção das amostras sobre os *loadings* para a determinação dos *scores* e o ajuste por uma função linear dos *scores* da matriz **X** aos *scores* da matriz **Y** de modo a minimizar os desvios. Essa otimização simultânea ocasiona pequenas distorções nas direções dos *loadings*, de modo que, rigorosamente eles perdem a ortogonalidade, levando à pequenas redundâncias de informação. No entanto, são essas pequenas redundâncias que otimizam a relação linear entre os scores e estas distorções da ortogonalidade entre os PCs no PLS fazem com que os mesmos não sejam mais componentes principais (que são ortogonais) e sim variáveis latentes[26,36-38].

A regressão por mínimos quadrados parciais estende o conceito do modelo inverso (propriedade como função da resposta instrumental) trocando as variáveis originais por um sub-conjunto truncado das variáveis latentes dos dados originais[26,36-38]. Considerando um caso geral para a determinação de mais de uma espécie de interesse, logo **Y** é uma matriz de dimensão (n x z), onde z é o número de colunas de **Y**, tem-se a decomposição de ambas as matrizes **X** de dimensão (n x m) e **Y** em suas matrizes de scores e loadings[26]:

$$\mathbf{X} = \mathbf{TP}^{T} + \mathbf{E}_x = \sum \mathbf{t}_A \mathbf{p}_A^T + \mathbf{E}_x \quad (23)$$

$$\mathbf{Y} = \mathbf{UQ}^{T} + \mathbf{E}_y = \sum \mathbf{u}_A \mathbf{q}_A^T + \mathbf{E}_y \quad (24)$$

em que, **X** é a matriz de respostas instrumentais, **Y** é a matriz de respostas da propriedade de interesse obtida por metodologia padrão, **T** e **U** são os scores de **X** e **Y**, respectivamente, **P** e **Q** são os loadings de **X** e **Y**, respectivamente, $\mathbf{E}_x$ e $\mathbf{E}_y$ corresponde a matriz de resíduos composta pelas variáveis latentes descartadas, ou seja, as matrizes que contem a parte não modelada[26].

Entre os *scores* de **X** e os *scores* de **Y**, uma relação linear é, então, estabelecida[26]:

$$\hat{\mathbf{u}}_A = \mathbf{b}_A \hat{\mathbf{t}}_A \tag{25}$$

em que, $\mathbf{b}_A$ é o vetor de coeficientes de regressão do modelo linear para cada variável latente, obtido através de:

$$\mathbf{b}_A = \frac{\mathbf{u}_A^T \mathbf{t}_A}{\mathbf{t}_A^T \mathbf{t}_A} \tag{26}$$

A Figura 5 ilustra a decomposição das matrizes **X** e **Y** no produto das matrizes de scores e loadings. A decomposição pode ser realizada através de diversos algoritmos que procedem a referida decomposição por passos diferentes chegando ao final em resultados praticamente iguais. Um exemplo desses algoritmos, que foi empregado neste trabalho, é o NIPALS (do inglês, Nonlinear Interative Partial Least Squares)[39].

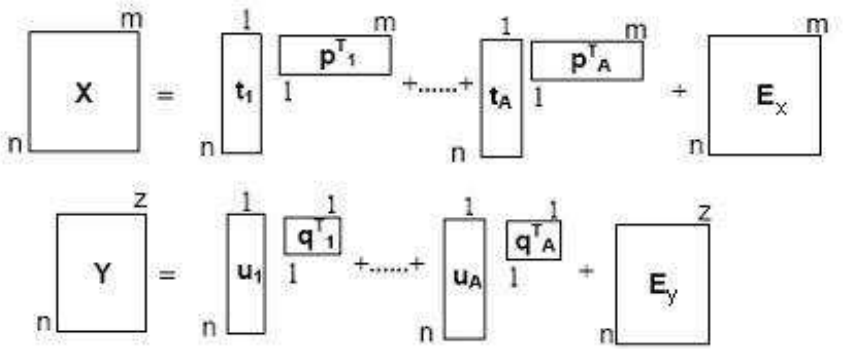


Figura 5. Decomposição em variáveis latentes das matrizes **X** e **Y** para modelos PLS.

O algoritmo NIPALS decompõe iterativamente a matriz de dados em uma soma do produto de scores e loadings até o número de variáveis latentes A[39]:

$$\mathbf{X} = \sum_{i=1}^{A} \mathbf{t}_i \mathbf{p}_i = \mathbf{t}_1 \mathbf{p}_1 + \mathbf{t}_2 \mathbf{p}_2 + \ldots + \mathbf{t}_A \mathbf{p}_A \quad (27)$$

Para a primeira VL (A=1) os seguintes passos são realizados para a matriz **X**:

1. Um vetor **x** qualquer de **X** é denominado $\mathbf{t}_A$
2. O seguinte cálculo é realizado para a primeira estimativa do conjunto dos loadings:

$$\mathbf{p}_A^T = \frac{\mathbf{t}_A^T \mathbf{X}}{\mathbf{t}_A^T \mathbf{t}_A} \left(= \frac{\mathbf{u}_A^T \mathbf{X}}{\mathbf{u}_A^T \mathbf{u}_A} \right) \quad (28)$$

3. A estimativa do conjunto dos loadings é normalizada para comprimento 1, ou seja, a estimativa do autovetor é autoescalada:

$$\mathbf{p}_{A,norm}^T = \frac{\mathbf{p}_A^T}{\left\| \mathbf{p}_A^T \right\|} \quad (29)$$

4. Primeira estimativa dos scores baseada nos loadings estimados no passo 3.

$$\mathbf{t}_A = \frac{\mathbf{X}\mathbf{p}_A}{\mathbf{p}_A^T \mathbf{p}_A} \quad (30)$$

5. Comparação de $\mathbf{t}_A$ obtido no passo 2 e no passo 4. Se forem iguais o algoritmo para e o procedimento inicia-se para a segunda variável latente, caso contrário, se forem diferentes, retorna-se ao passo 2 até convergir.

6. Os efeitos da primeira VL são removidos pela subtração do produto dos scores e loadings da matriz original:

$$\mathbf{X}_A = \mathbf{X}_{A-1} - \mathbf{t}_A \mathbf{p}_A^T \tag{31}$$

Para a primeira VL (A=1) os seguintes passos são realizados para a matriz **Y**:

1. Um vetor **y** qualquer de **Y** é denominado $\mathbf{u}_A$
2. O seguinte cálculo é realizado para a primeira estimativa do conjunto dos loadings:

$$\mathbf{q}_A^T = \frac{\mathbf{u}_A^T \mathbf{Y}}{\mathbf{u}_A^T \mathbf{u}_A} \left(= \frac{\mathbf{t}_A^T \mathbf{Y}}{\mathbf{t}_A^T \mathbf{t}_A} \right) \tag{32}$$

3. A estimativa do conjunto dos loadings é normalizada para comprimento 1, ou seja, a estimativa do autovetor é autoescalada:

$$\mathbf{q}_{A,norm}^T = \frac{\mathbf{q}_A^T}{\left\| \mathbf{q}_A^T \right\|} \tag{33}$$

4. Primeira estimativa dos scores baseada nos loadings estimados no passo 3.

$$\mathbf{u}_A = \frac{\mathbf{Y}\mathbf{q}_A}{\mathbf{q}_A^T \mathbf{q}_A} \tag{34}$$

5. Comparação de $\mathbf{u}_A$ obtido no passo 2 e no passo 4. Se forem iguais o algoritmo pára e o procedimento se inicia para a segunda variável latente, caso contrário, se forem diferentes, retorna-se ao passo 2 até convergir.

6. Os efeitos da primeira VL são removidos pela subtração do produto dos scores e loadings da matriz original:

$$\mathbf{Y}_A = \mathbf{Y}_{A-1} - \mathbf{u}_A \mathbf{q}_A \tag{35}$$

Os passos descritos acima para o algoritmo são escritos como relações completamente separadas entre as matrizes **X** e **Y**. O caminho no qual um único algoritmo pode implementar informação com relação às duas matrizes simultaneamente é dado no NIPALS no passo 2, para ambos os casos. Assim, esse algoritmo pode ser escrito como[39]:

Para a primeira VL (A=1) os seguintes passos são realizados para as matrizes **X** e **Y** simultaneamente :

1. Um vetor **y** qualquer de **Y** é denominado $\mathbf{u}_A$
2. O seguinte cálculo é realizado para a primeira estimativa do conjunto dos loadings de **X**:

$$\mathbf{w}_A^T = \frac{\mathbf{u}_A^T \mathbf{X}}{\mathbf{u}_A^T \mathbf{u}_A} \tag{36}$$

3. A estimativa do conjunto dos loadings de **X** é normalizada para comprimento 1, ou seja, a estimativa do autovetor é autoescalada:

$$\mathbf{w}_{A,norm}^T = \frac{\mathbf{w}_A^T}{\left\| \mathbf{w}_A^T \right\|} \tag{37}$$

4. Primeira estimativa dos scores de **X** baseada nos loadings estimados no passo 3.

$$\mathbf{t}_A = \frac{\mathbf{X}\mathbf{w}_A}{\mathbf{w}_A^T\mathbf{w}_A} \tag{38}$$

5. O seguinte cálculo é realizado para a primeira estimativa do conjunto de loadings de **Y**:

$$\mathbf{q}_A^T = \frac{\mathbf{t}_A^T\mathbf{Y}}{\mathbf{t}_A^T\mathbf{t}_A} \tag{39}$$

6. A estimativa do conjunto dos loadings de **Y** é normalizada para comprimento 1, ou seja, a estimativa do autovetor é autoescalada:

$$\mathbf{q}_{A,norm}^T = \frac{\mathbf{q}_A^T}{\left\|\mathbf{q}_A^T\right\|} \tag{40}$$

7. Primeira estimativa dos scores de **Y** baseada nos loadings do passo 6.

$$\mathbf{u}_A = \frac{\mathbf{Y}\mathbf{q}_A}{\mathbf{q}_A^T\mathbf{q}_A} \tag{41}$$

8. Como o algoritmo não fornece os valores de **t** ortogonais, os $\mathbf{p}^T$ são substituídos por $\mathbf{w}^T$ e um passo extra é incluído depois da convergência tornando os valores de **t** ortogonais:

$$\mathbf{p} = \frac{\mathbf{t}^T\mathbf{X}}{\mathbf{t}^T\mathbf{t}} \tag{42}$$

9. Comparação de $\mathbf{t_A}$ obtido no passo 4 com o do passo da iteração anterior. Se forem iguais o algoritmo pára e o procedimento inicia-se para a segunda

variável latente, caso contrário, se forem diferentes, retorna-se ao passo 2 até convergir.

10. Os efeitos da primeira VL são removidos pela subtração do produto dos scores e loadings da matriz original:

$$\mathbf{X}_A = \mathbf{X}_{A-1} - \mathbf{t}_A\mathbf{p}_A \tag{43}$$

$$\mathbf{Y}_A = \mathbf{Y}_{A-1} - \mathbf{u}_A\mathbf{q}_A \tag{44}$$

No caso de se ter um vetor **y** ao invés de uma matriz, os passos de 5 a 9 podem ser omitidos pois **q**=1[39]. Quando um modelo de calibração por PLS é construído a partir de uma matriz de dados **X** e um vetor de variáveis dependentes, o método é conhecido como PLS1[30].

Esse processo é repetido até o número de variáveis latentes desejado ou definido. No final do processo a variância explicada pela primeira VL será maior que a variância explicada pela segunda VL e a terceira VL explicará uma variância menor que a segunda VL, e assim sucessivamente até o número de VL definido e o algoritmo, geralmente, converge rapidamente[39].

O coeficiente de regressão, $\hat{\mathbf{B}}_{\mathbf{PLS}}$ é encontrado através de:

$$\hat{\mathbf{B}}_{\mathrm{PLS}} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{q} \tag{45}$$

E o modelo pode ser representado como:

$$\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}_{\mathrm{PLS}} \tag{46}$$

3.3. Seleção de variáveis

Evidências teóricas e experimentais, indicam que a escolha adequada de regiões espectrais pode melhorar significativamente a eficiência do modelo de calibração multivariada. A seleção de variáveis envolve a escolha de determinadas regiões do espectro (um comprimento de onda ou um conjunto de comprimentos de onda) independentes e mais restrito, que permitem ao modelo de calibração minimizar os erros de previsão. Como conseqüência da seleção de variáveis, é possível produzir um modelo mais robusto, simples de interpretar e com melhor precisão nas previsões[40]. Assim, em análises espectroscópicas no infravermelho próximo, os comprimentos de onda que apenas induzem a ruídos, informações irrelevantes ou não-linearidades, podem ser eliminados[41].

Existem vários métodos para a escolha da região espectral, sendo que esses métodos diferem com relação ao procedimento realizado para a escolha[41]. Alguns exemplos desses métodos são o algoritmo genético[42], o método de mínimos quadrados parciais por intervalo (iPLS – do inglês, Interval Partial Least Square)[41,43] e o método de mínimos quadrados parciais com eliminação de variáveis não-informativas (UVE-PLS – do inglês, Elimination of Uniformative Variables in Partial Least Square)[44].

Neste trabalho foi utilizado o método iPLS para a seleção de variáveis, tendo em vista que este método seleciona faixas do espectro e não variáveis isoladas, e também, por se tratar de um método mais simples em relação aos demais métodos citados acima.

O método iPLS é uma extensão interativa desenvolvida para o PLS, onde é feita uma regressão por mínimos quadrados parciais em cada sub-intervalo eqüidistante ao longo de toda a extensão do espectro. Desta forma é avaliada a relevância da informação nas diferentes sub-divisões espectrais, de onde é possível identificar e selecionar apenas as variáveis que apresentam informações mais relevantes. As regiões espectrais, cujas variáveis se apresentam como supostamente de menor importância e detentoras de possíveis ruídos, são removidas, e um novo modelo é construído a partir das variáveis selecionadas[43].

A seletividade do algoritmo PLS para as variáveis que apresentam ruído é realçada através da informação do erro de validação cruzada em função de cada intervalo no qual o espectro é sub-dividido. Portanto, através dos valores de RMSECV, para cada intervalo de sub-divisão do espectro, em relação ao valor de RMSECV para o modelo global, com todas as variáveis, avalia-se a performance da previsão de cada intervalo. O melhor intervalo, ou seja, a melhor seleção de variáveis, será aquela que apresentar valor de RMSECV menor em relação ao modelo global. Normalmente, a construção de modelos PLS, a partir dessas variáveis, necessita de um número de VL diferente daquele necessário ao modelo global para alcançar uma variância relevante de **y**. Isso ocorre, principalmente, devido à largura dos sub-intervalos, número de substâncias que absorvem/interferem e ruído[43].

Como a seleção de variáveis busca-se encontrar um conjunto de variáveis independentes mais restrito que produza um menor erro de previsão, o algoritmo monitora, concomitantemente à seleção das variáveis, um parâmetro regulador, E(b), que descreve o desvio das previsões em relação aos valores esperados, conforme a equação[41]:

$$E(b) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \qquad (47)$$

em que, y_i e $\hat{y}_i$ consistem nos valores de referência e previstos, respectivamente, N é o número de amostras e b são os coeficientes de regressão calculados pelo PLS.

A eliminação de qualquer variável fica, portanto, vinculada à minimização da função descrita pela equação (47), ou seja, é necessário atingir um mínimo da função simultaneamente à eliminação do valor de um determinado coeficiente de regressão (b) do modelo global referente à variável eliminada[41].

3.4. Algoritmo de Kennard-Stone

A divisão de um conjunto entre amostras de calibração e validação deve ser realizada, de tal maneira, que as amostras de validação sejam bem representadas pelas amostras de calibração. Existem vários métodos para realizar esta separação como por exemplo: Algoritmo de projeções sucessivas (SPA – do inglês, Successive Projections Algorithm)[45], PDS (do inglês, Piecewise Direct Standardisation)[46,47] e o algoritmo de Kennard-Stone[48], entre outros[49,50,51,52].

Para este trabalho, a separação do conjunto de dados entre os conjuntos de calibração e validação foi realizada através do algoritmo de Kennard-Stone por se tratar de um algoritmo que seleciona as amostras com base em suas distâncias, é de fácil aplicação, além do que consiste em uma alternativa viável para cumprir o objetivo almejado.

No algoritmo de Kennard-Stone, a primeira amostra selecionada pelo algoritmo é a que apresenta a maior distância em relação à amostra média. A segunda amostra a ser selecionada será a que apresentar maior distância em relação à primeira amostra selecionada. A próxima amostra a ser selecionada apresentará maior distância em relação à ultima amostra selecionada, e assim sucessivamente até atingir o número de amostras desejadas[48].

Normalmente, esse algoritmo é aplicado para realizar a seleção das amostras que irão compor o conjunto de calibração, uma vez que este procede a seleção das amostras de maior variabilidade, ou seja, as amostras mais “externas” do conjunto total.

3.6. Detecção de amostras anômalas - Outliers

Outliers é o termo utilizado para designar amostras anômalas que podem estar presentes nos conjuntos de calibração e de validação, que serão empregados na construção e validação de um modelo de calibração multivariada, respectivamente. Normalmente, essas amostras anômalas possuem um comportamento diferente das demais amostras do conjunto de dados[37].

A presença desse tipo de amostra no conjunto de calibração pode conduzir a modelos com baixa capacidade de previsão, ou seja, que produzem altos

valores de erro[37]. Quando presentes no conjunto de validação, podem influenciar os resultados, geralmente, levando a resultados que indicam que o modelo não é adequado ou que a sua capacidade é inferior à que poderia ser apresentada na ausência destas amostras anômalas. Assim, a identificação de anomalias é um passo importante para a otimização dos conjuntos de calibração e validação, sendo que, a exclusão destes permite a construção de modelos mais eficientes e precisos e com melhor capacidade de previsão[30].

Existem diversas técnicas para a identificação de amostras anômalas, como por exemplo: a distância de Mahalanobis[53], método da incerteza[54,55,56], uma metodologia conhecida como "convex hull"[55,56] , funções potenciais[55,,57], RHM (do inglês, Resampling by the Half-Means)[55,58,59], SHV (do inglês, Smallest Half-Volume)[55,58], entre outros[60], sendo que esta área constitui uma linha de pesquisa recente e em andamento que, no entanto, é de grande importância tendo em vista a variedade de tipos de equipamentos e a grande quantidade de amostras que geram dados de primeira ordem.

No presente trabalho, a identificação dos outliers presentes nos conjuntos de calibração e validação foi realizada seguindo as recomendações da norma E1655-00 da ASTM (American Society for Testing and Materials)[61] e da referência 30.

3.6.1. Amostras anômalas na calibração

Os *outliers* na calibração são, usualmente, avaliados com base no *leverage* extremo, resíduos não modelados nos dados espectrais e resíduos não modelados na variável dependente[30,61].

O *leverage* representa o grau que uma amostra está distante da média do conjunto de dados, ou seja, o "peso relativo" de uma amostra em relação às demais presentes em um mesmo conjunto. Qualitativamente, tomando como exemplo dados espectrais, o leverage mede o quanto o espectro de uma amostra difere dos espectros das demais amostras presentes no conjunto de dados. O *leverage* pode ser representado por[30,61]:

$$h_i = \hat{\mathbf{t}}_i^T \left(\hat{\mathbf{T}}^T \hat{\mathbf{T}}\right)^{-1} \hat{\mathbf{t}}_i \quad (48)$$

em que: **T** são os *scores* de todas as amostras de calibração, $\mathbf{t}_i$ é o vetor de scores de uma amostra em particular.

Amostras com altos valores de h podem ser consideradas como amostras anômalas. De acordo com a ASTM E1955-00[61], amostras com $h_i > 3\frac{A+1}{n}$, onde n corresponde ao número de amostras da calibração, devem ser removidas do conjunto de calibração, e o modelo deve, então ser reconstruído. No entanto, é comum um novo espectro apresentar $h_i > 3\frac{A+1}{n}$ após a construção do novo modelo. Quando aplicações repetidas do teste continuam a apresentar $h_i > 3\frac{A+1}{n}$ identificando novas amostras como sendo *outliers*, um fenômeno conhecido como "*snowball*" (do inglês, efeito bola de neve) pode estar acontecendo. Se este fenômeno ocorrer, é indicativo de que algum problema com a estrutura dos dados espectrais pode estar acontecendo. O espaço das variáveis do modelo podem ser examinados por uma distribuição de "*clusterings*" (agrupamentos) não usual. Entretanto, se for observado a ocorrência de "*snowball*" na seqüência de desenvolvimento dos modelos, o teste para identificação de *outliers* com base no *leverage* pode ser relaxado: (1) um primeiro modelo é construído a partir do conjunto de calibração inicial, (2) espectros de calibração com $h_i > 3\frac{A+1}{n}$ são eliminados do conjunto de calibração, (3) um segundo modelo usando o mesmo número, A, de variáveis latentes é construído com o subconjunto de espectros de calibração, e (4) espectros de calibração com $h_i > 3\frac{A+1}{n}$ são identificados no segundo modelo. O segundo modelo pode ser utilizado como critério de parada do teste para identificação de *outliers* baseado no *leverage*, desde que as amostras de calibração apresentem valor de h menor do

que 0,5. Se for o caso, os *outliers* para o segundo modelo são removidos e um terceiro modelo é, ainda, construído[61].

Identificação de anomalias em relação aos resíduos não modelados nos dados espectrais são obtidas por comparação do desvio padrão residual total (s(ê)), definido como:

$$s(\hat{e})^2 = \frac{1}{nJ - J - A\max(n,J)} \sum_{i=1}^{n} \left(\sum_{j=1}^{J} (x_{i,j} - \hat{x}_{i,j})^2 \right) \quad (49)$$

em que, J é o número de variáveis espectrais e n é o número de amostras da calibração.

O desvio padrão residual de uma amostra i ($s(\hat{e}_i)$) é calculado por:

$$s(\hat{e}_i)^2 = \frac{n}{nJ - J - A\max(n,J)} \sum_{j=1}^{J} (x_{i,j} - \hat{x}_{i,j})^2 \quad (50)$$

Se uma amostra apresentar $s(\hat{e}_i) > 2s(\hat{e})$ esta é removida do conjunto de calibração[30].

Com relação aos resíduos não modelados na variável dependente, os *outliers* são identificados através da comparação da raiz quadrada do erro médio da calibração (RMSEC) com o erro absoluto daquela amostra. Se a amostra apresentar erro absoluto $(y_i - \hat{y}_i)$ maior do que (3xRMSEC), esta é definida como sendo um *outlier*[30]. O RMSEC é determinado como:

$$RMSEC = \frac{1}{n - A - 1} \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (51)$$

em que, $\mathbf{y_i}$ é o valor de referência e $\hat{\mathbf{y}}_\mathbf{i}$ é o valor estimado.

3.6.2. Amostras anômalas na validação

Para a identificação dos *outliers* no conjunto de validação, além dos testes baseados no *leverage* e nos resíduos espectrais, que foram descritos na seção anterior, pode ser realizado um teste que considera os resíduos não modelados da repetibilidade espectral de amostras em três níveis de concentração com sete replicatas para cada nível, como descrito na ASTM E1655-00[61]. Este teste é baseado na raiz quadrada média dos resíduos espectrais (RMSSR), que, para uma amostra, é definida como:

$$\mathrm{RMSSR} = \frac{1}{\mathrm{J}}\sqrt{\sum_{\mathrm{j}=1}^{\mathrm{J}}\left(\mathrm{x_j} - \hat{\mathrm{x}}_\mathrm{j}\right)^2} \tag{52}$$

em que, J é o número de variáveis.

Um RMSSR limite pode ser determinado através da equação:

$$\mathrm{RMSSR_{limit}} = \left[\sum \frac{\mathrm{RMSSR_{reply(i)}}}{\mathrm{RMSSR_{cal(i)}}}\right]\mathrm{RMSSR_{max}} \tag{53}$$

em que, $\mathrm{RMSSR_{max}}$ é o valor máximo observado para as amostras de calibração, $\mathrm{RMSSR_{cal(i)}}$ é a raiz quadrada da média dos resíduos espectrais para uma replicata de cada nível que são incluídas no conjunto de calibração e $\mathrm{RMSSR_{reply(i)}}$ é a raiz quadrada da média dos resíduos espectrais para a média de seis replicatas de cada nível que não estarão incluídas no conjunto de calibração.

Se o valor da RMSSR de uma amostra de validação apresentar um valor maior do que $\mathrm{RMSSR_{limit}}$, esta amostra deverá ser removida do conjunto de validação.

CAPÍTULO 4
Validação e Figuras de Mérito

4. Validação e figuras de mérito

4.1. Validação

A validação é um processo de averiguação da performance de um método, com o intuito de avaliar se este apresenta uma performance adequada para as condições nas quais será aplicado. O processo de validação deve ser realizado sempre que um procedimento analítico é proposto ou desenvolvido[62]. A validação de um método estabelece, por estudos sistemáticos realizados em laboratório, que o método atende ao seu propósito e às normas impostas por órgãos de fiscalização nacionais e internacionais. Para o caso da implantação do método por espectroscopia no infravermelho próximo na indústria alcooleira, o órgão de fiscalização responsável é o conselho dos produtores de cana-de-açúcar, açúcar e álcool, o qual faz a regulamentação de acordo com cada estado[4]. A validação de métodos baseados em espectroscopia no infravermelho próximo, é também fiscalizada pela American Society for Testing and Materials (ASTM)[61].

A validação pode ser atestada através da determinação de parâmetros conhecidos como figuras de mérito, que, dependendo de onde o método será aplicado, do seu propósito e ou do órgão de fiscalização a que estará sujeito, o número de figuras de mérito ou nível que deve ser atingido em cada uma delas, pode variar[4,61,62].

As principais figuras de mérito são:

- Exatidão
- Precisão
- Sensibilidade
- Seletividade
- Linearidade
- Razão sinal/ruído
- Limite de detecção
- Limite de quantificação
- Robustez
- Intervalos de confiança

- Teste para erros sistemáticos
- Extensão da faixa de trabalho

A maneira pela qual essas figuras de mérito devem ser determinadas é estabelecida pelos órgãos de fiscalização e encontra-se descrita em normas específicas[61], guias de validação[62,63,64] e trabalhos científicos[65,66,67,68]. Entretanto, a maioria dos guias, normas e trabalhos científicos, ainda são referentes à calibração univariada e são poucos os trabalhos científicos que realizam a determinação de figuras de mérito para validação de modelos de calibração multivariada.

Determinações quantitativas utilizando espectroscopia no infravermelho próximo tem crescido a cada ano, e estas determinações fazem uso de modelos de calibração multivariada[16]. A ausência de validação desses modelos, devido à carência de normas oficiais que descrevem como essa validação deve ser realizada tem restringido sua implementação. Nos últimos anos, vem sendo direcionada uma considerável atenção para a elaboração de guias, normas e trabalhos científicos que enfocam a necessidade da validação de modelos de calibração multivariada, de modo que, atualmente, já se encontram disponíveis trabalhos que descrevem procedimentos e propostas de como esta validação deve ser realizada. Alguns documentos apresentam caráter geral[62,64], enquanto que outros, como por exemplo a norma E1655-00 da ASTM[61] descrevem especificamente o desenvolvimento e validação de modelos de calibração multivariada a partir da espectroscopia no infravermelho. Muitos trabalhos científicos tratam de problemas específicos, e, portanto, abordam a determinação de apenas algumas figuras de mérito específicas relacionadas ao trabalho realizado. São exemplos destes trabalhos: a determinação de limite de detecção[69], estimativa de intervalos de confiança[70], determinação da sensibilidade e sensibilidade analítica[71], determinação da seletividade[72], precisão no nível de repetibilidade e precisão intermediária[73], sendo que estes trabalhos mostram que figuras de mérito, raramente abordadas durante a validação de modelos multivariadas, podem ser determinadas. São exemplos, ainda, trabalhos que propõem melhora de

definições[74] e relações entre diferentes definições para uma mesma figura de mérito[75].

Para este trabalho, as figuras de mérito foram estimadas seguindo a norma E1655-00 da ASTM[61], com o intuito de validar os modelos PLS construídos a partir de espectros obtidos na região do NIR, para determinações de parâmetros de controle de qualidade da indústria alcooleira.

4.2. Sinal analítico líquido – NAS

O conceito de sinal analítico líquido (NAS – do inglês, Net Analyte Signal) exerce uma importante função na determinação de figuras de mérito para calibrações multivariadas.

O método para o cálculo do NAS para modelos multivariados de calibração inversa foi proposto por Lorber[76,77].

O NAS é definido, para uma propriedade de interesse k, como sendo a parte do sinal analítico que é ortogonal às contribuições de possíveis interferentes presentes na amostra. Sua propriedade de ortogonalidade pode ser observada pela representação geométrica da Figura 6[78]:

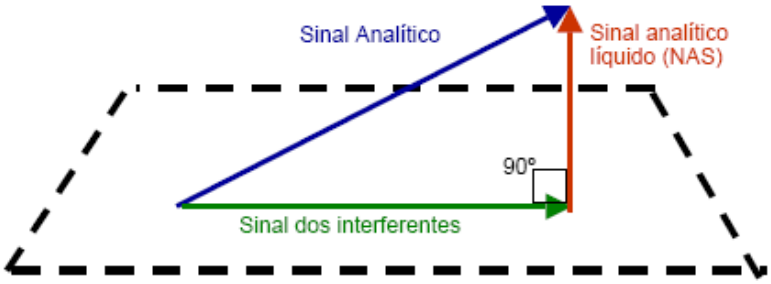


Figura 6. Representação geométrica da propriedade de ortogonalidade do NAS.

O método proposto por Lorber[76,77] foi corrigido por Ferre, Brown e Rius[79] para possibilitar o cálculo exato do NAS para modelos de calibração construídos a partir dos métodos de regressão baseados no PLS e PCR.

Para o cálculo do NAS, primeiramente, **X** e **y** são reconstruídos com A variáveis latentes gerando $\hat{\mathbf{X}}_A$ e $\hat{\mathbf{y}}_A$, segundo as equações[76]:

$$\hat{\mathbf{X}}_{\mathrm{A}} = \mathbf{T}_{\mathrm{A}}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}} + \mathbf{E} \tag{54}$$

$$\hat{\mathbf{y}}_{\mathrm{A}} = \mathbf{U}_{\mathrm{A}}\mathbf{q}_{\mathrm{A}} + \mathbf{f} \tag{55}$$

em que, **T** e **U** são os *scores*, **P** e **q** são os *loadings* e **E** e **f** representa o erro de decomposição da matriz **X** e do vetor **y**, respectivamente.

O próximo passo é a determinação da matriz $\hat{\mathbf{X}}_{\mathrm{A},-\mathrm{k}}$, que contém a informação de todas as espécies presentes na amostra exceto a informação referente à espécie k. Essa determinação é realizada através de uma projeção ortogonal baseada na operação matricial que estabelece que para uma matriz **X** qualquer, a matriz $\mathbf{XX}^{+}$ (onde "+" indica a Moore-Penrose pseudo inversa da matriz) é uma matriz de projeção com as seguintes propriedades[76].

$$\mathbf{X} = \left(\mathbf{XX}^{+}\right)\mathbf{X} \tag{56}$$

$$\mathbf{X}^{+} = \mathbf{X}^{+}\left(\mathbf{XX}^{+}\right) \tag{57}$$

A partir dessas propriedades, se qualquer vetor **z** for uma combinação linear da matriz **X**, a multiplicação de **z** pela matriz $\mathbf{XX}^{+}$ fornecerá como resultado o próprio vetor **z**. No entanto, a multiplicação de **z** por $\mathbf{(I-XX}^{+}\mathbf{)}$, (onde **I** representa a matriz identidade de dimensões adequadas) resultará em um vetor de zeros. Assim, a multiplicação de um vetor pela matriz $\mathbf{(I-XX}^{+}\mathbf{)}$ fornecerá como resultado um vetor que será ortogonal à matriz **X**[76]. Dessa forma, a matriz $\hat{\mathbf{X}}_{\mathrm{A},-\mathrm{k}}$ é obtida por[80]:

$$\hat{\mathbf{X}}_{\mathrm{A},-\mathrm{k}} = \left(\mathbf{I} - \hat{\mathbf{y}}_{\mathrm{A,k}}\hat{\mathbf{y}}_{\mathrm{A,k}}^{+}\right)\hat{\mathbf{X}}_{\mathrm{A}} \tag{58}$$

em que, $\hat{\mathbf{y}}_{A,k}$ é o vetor de concentrações da espécie de interesse k estimado com A variáveis latentes segundo a equação[80]:

$$\hat{\mathbf{y}}_{A,k} = \hat{\mathbf{X}}_A \hat{\mathbf{X}}_A^{+} \mathbf{y}_k \tag{59}$$

Assim, a matriz $\hat{\mathbf{X}}_{A,-k}$ fica livre de qualquer contribuição da espécie k uma vez que a projeção realizada na equação (58) indica a parte da matriz **X** que é ortogonal ao vetor **y** que contém os valores do método de referência e, portanto, a matriz $\hat{\mathbf{X}}_{A,-k}$ contém a parte de **X** que não possui relação com **y**. O vetor NAS pode, então, ser obtido com uma nova projeção que indicará a parte da matriz **X** que é ortogonal à matriz de interferentes $\hat{\mathbf{X}}_{A,-k}$, resultando assim, na parte de **X** que não possui relação com os interferentes:

$$\hat{\mathbf{x}}_{A,k}^{nas} = \left(\mathbf{I} - \hat{\mathbf{X}}_{A,-k}^{T} \left(\hat{\mathbf{X}}_{A,-k}^{T} \right)^{+} \right) \hat{\mathbf{x}}_A \tag{60}$$

em que, $\hat{\mathbf{x}}_A$ é o vetor de respostas instrumentais de uma amostra estimado com A variáveis latentes.

Uma vez que $\hat{\mathbf{x}}_{A,k}^{nas}$ é livre de interferentes, é possível substituí-lo por uma representação escalar, sem perda de informação. Assim:

$$\mathbf{nâs}_i = \left\| \hat{\mathbf{x}}_{A,k}^{nas} \right\| \tag{61}$$

em que, $\| \ \|$ representa a norma Euclideana do vetor $\hat{\mathbf{x}}_{A,k}^{nas}$.

Com a possibilidade de calcular um valor escalar livre de interferentes, a partir de um vetor contendo contribuições de constituintes desconhecidos, é

possível construir uma nova forma de calibração multivariada, em que o modelo pode ser representado em uma forma pseudo-univariada.

A representação univariada de um modelo de calibração multivariada possibilita avaliar a porção do sinal que eficientemente participa do modelo[81].

De posse do cálculo do NAS para as i amostras de calibração, o coeficiente de regressão pode ser determinado, por mínimos quadrados, entre o vetor **nas** e o vetor de concentrações **y**:

$$\hat{b}_{nas} = \left(\mathbf{n\hat{a}s}^T \mathbf{n\hat{a}s}\right)^{-1} \mathbf{n\hat{a}s}^T \mathbf{y} \tag{62}$$

O modelo de regressão pode ser, então, representado por:

$$\hat{\mathbf{y}} = \hat{b}_{nas} \mathbf{n\hat{a}s} + \in \tag{63}$$

em que os resultados obtidos por meio das equações (63) e (46) são equivalentes.

Quando os dados são centrados na média para a construção do modelo de calibração multivariada, o vetor **nâs** precisa ser corrigido para evitar um erro de sinal que é introduzido pelo uso da norma Euclideana, antes da determinação do coeficiente de regressão $\hat{b}_{nas}$. Esta correção pode ser feita multiplicando cada elemento do vetor **nâs** pelo seu sinal correspondente no vetor $(\mathbf{y} - \bar{y})$, onde $\bar{y}$ é a média do vetor **y** que contém os valores de referência[82].

4.3. Figuras de mérito

As figuras de mérito descritas a seguir estão de acordo com normas específicas e trabalhos científicos divulgados por meio de periódicos.

4.3.1. Exatidão

Expressa o grau de concordância entre o valor estimado pelo modelo multivariado e o valor tido como verdadeiro ou de referência. Em calibração

multivariada, normalmente, é expressa através da raiz quadrada do erro médio quadrático de previsão (RMSEP – do inglês, Root Mean Squares Error of Prediction)[30]:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{I_v} (y_i - \hat{y}_i)^2}{I_v}} \quad (64)$$

em que, I_v é o número de amostras do conjunto de validação, y_i e $\hat{y}_i$ correspondem aos valores de referência e aos previstos pelo modelo, respectivamente.

4.3.2. Precisão

Expressa o grau de concordância entre os resultados de uma série de medidas realizadas para uma mesma amostra homogênea em condições determinadas. Em geral a precisão pode obtida em diversos níveis, tais como: repetibilidade, precisão intermediária, reprodutibilidade e precisão média[61].

- Repetibilidade: É a precisão do método em um curto intervalo de tempo. Segundo a norma E1655-00 da ASTM[61], são necessárias um mínimo de três amostras em concentrações diferentes cobrindo a faixa útil do modelo de calibração e seis replicatas para nível de concentração.
- Precisão intermediária: A extensão em que a precisão intermediária deve ser determinada depende das circunstâncias em que o método será aplicado. Este nível de precisão é obtido a partir da variação de uma determinada condição experimental. Variações típicas estudadas incluem, por exemplo, dias, analistas, equipamentos, entre outras.
- Reprodutibilidade: É acessada por meio de ensaios interlaboratoriais. Este nível de precisão normalmente é requerido em casos de

padronização de procedimentos analíticos, como por exemplo, a inclusão de procedimentos em farmacopéias.

Neste trabalho a precisão foi calculada no nível de repetibilidade, seguindo as recomendações da ASTM[61]:

$$\text{precisão} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\hat{y}_{ij} - \hat{\bar{y}}_i\right)^2}{n(m-1)}} \tag{65}$$

em que, n representa o número de amostras e m o número de replicatas.

4.3.3. Sensibilidade

É definida como a fração do sinal responsável pelo acréscimo de uma unidade de concentração à propriedade de interesse. Para modelos de calibração multivariada, como PLS, pode ser determinada como[67,77,79,83]:

$$\hat{\text{SEN}} = \frac{1}{\|\mathbf{b}_k\|} \tag{66}$$

em que, $\mathbf{b}_k$ é o vetor dos coeficientes de regressão estimados pelo PLS.

Quando o NAS é determinado, o vetor de sensibilidade líquida $\mathbf{s}_k^{nas}$ para cada amostra do conjunto de calibração, pode ser determinado a partir do vetor $\hat{\mathbf{x}}_{A,k}^{nas}$ como:

$$\mathbf{s}_k^{nas} = \left\| \frac{\hat{\mathbf{x}}_{A,k}^{nas}}{\mathbf{y}} \right\| \tag{67}$$

em que, o vetor de sensibilidades $\mathbf{s}_k^{nas}$ deve ser igual para todas as amostras de calibração, $\hat{\mathbf{x}}_{A,k}^{nas}$ é o vetor de sinal analítico líquido para a espécie k, **y** é o vetor que contém os valores de referência.

O escalar SÊN, pode ser determinado por:

$$S\hat{E}N = \left\| s_k^{nas} \right\| \tag{68}$$

4.3.4. Sensibilidade Analítica

A sensibilidade analítica, γ, não é abordada em normas ou guias de validação. No entanto, esse parâmetro apresenta a sensibilidade do método em termos da unidade de concentração que é utilizada, sendo definida como a razão entre a sensibilidade e o desvio padrão do sinal de referência (δx)[83,84].

$$\gamma = \frac{S\hat{E}N}{\|\delta x\|} \tag{69}$$

em que, SÊN é obtido através das equações (66) ou (68) e δx é o desvio padrão do sinal de referência estimado através do desvio padrão do valor de NAS para 15 espectros do sinal de referência.

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

4.3.5. Seletividade

É a medida do grau de sobreposição entre o sinal da espécie de interesse e os interferentes presentes na amostra, indicando também, a parte do sinal que é perdida por essa sobreposição[85]. Para modelos de calibração multivariada, a seletividade, $S\hat{E}L_{k,i}$, é definida como[77,79]:

$$\hat{SEL}_{k,i} = \frac{\hat{nas}_{k,i}}{\left\|\mathbf{x}_{k,i}\right\|} \quad (70)$$

em que, $\hat{nas}_{k,i}$ é o valor escalar do sinal analítico líquido para a amostra i e $\mathbf{x}_{k,i}$ representa o vetor de respostas instrumental para a amostra i.

4.3.6. Linearidade

A avaliação desta figura de mérito é problemática em calibração multivariada utilizando PLS ou PCR, uma vez que as variáveis são decompostas pelos componentes principais. Assim, uma medida quantitativa para a linearidade não corresponderia a uma tarefa simples, ou mesmo possível. Qualitativamente, o gráfico dos resíduos para as amostras de calibração e validação pode indicar se os dados seguem um comportamento linear se a distribuição destes resíduos for aleatória[30].

4.3.7. Ajuste

O ajuste para um modelo não consiste em uma figura de mérito abordada com freqüência em guias e normas oficiais. Entretanto, devido à dificuldade em realizar estimativas para linearidade, esse parâmetro é com freqüência determinado em trabalhos que envolvem modelos de calibração multivariada. Este parâmetro pode ser estimado a partir da correlação entre os valores tidos como referência e os valores estimados pelo modelo de calibração multivariada, para a propriedade de interesse. Isso é feito determinando por mínimos quadrados, a reta que melhor se ajusta aos valores de referência e os valores estimados pelo modelo, para as amostras de calibração. Quando o escalar “nas” é determinado, é possível também determinar o ajuste do modelo através da melhor reta que se ajusta à curva do “nas” contra a concentração, para as amostras de calibração[30,65,77].

4.3.8. Razão sinal/ruído

Em calibração univariada, a razão sinal/ruído, $S/N_{k,i}$, representa o quanto do sinal do analito é maior do que o ruído instrumental. No caso da calibração multivariada, esta razão indica o quanto da intensidade do NAS da espécie de interesse está acima do desvio padrão do sinal de referência[76,77,79]:

$$S/N_{k,i} = \frac{n\hat{a}s_{k,i}}{\delta x} \tag{71}$$

em que, $n\hat{a}s_{k,i}$ é o valor escalar do sinal analítico líquido para a amostra i e δx é o desvio padrão do sinal de referência.

4.3.9. Robustez

Em processos industriais, ou mesmo em análises de bancada, existem diversas variáveis, instrumentais ou ambientais, que não são possíveis de se controlar. Alguns exemplos são a umidade, temperatura, pequenas variações na quantidade dos componentes para a formação de um produto na indústria, entre outros. Para tanto, um método analítico robusto não deve ser sensível a esses tipos de variações, pois isto acarretaria na introdução de erros que podem ser significativos ao resultado. A robustez em calibração multivariada, consiste em testar a performance do modelo de calibração multivariada frente a alguns tipos de variações e averiguar se estas são ou não significativas[62].

4.3.10. Extensão da faixa de trabalho

É estabelecida determinando a espécie de interesse em diferentes concentrações. Através dos resultados obtidos, determina-se a faixa de concentração na qual os resultados apresentam um nível aceitável de incerteza para o método empregado[62].

4.3.11. Limite de Detecção e Quantificação

O limite de detecção (LD) e o limite de quantificação (LQ) de um procedimento analítico, expressam as menores quantidades da espécie de interesse que podem ser detectadas e determinadas quantitativamente, respectivamente. Para um conjunto de dados que apresenta comportamento homoscedástico (variância constante ao longo da faixa de trabalho, erros com previsão não correlacionados e que seguem uma distribuição normal), os LD e LQ na calibração multivariada podem ser calculados por[86]:

$$LD = 3\delta x \left\| b_k \right\| = 3\delta x \frac{1}{S\hat{E}N} \tag{72}$$

$$LQ = 10\delta x \left\| b_k \right\| = 10\delta x \frac{1}{S\hat{E}N} \tag{73}$$

em que, δx é o desvio padrão do sinal de referência, $\mathbf{b_k}$ é o vetor dos coeficientes de regressão do modelo PLS para a espécie k, SÊN corresponde ao valor de sensibilidade obtido através das equações 66 ou 68.

5.3.12. Intervalos de Confiança

O intervalo de confiança para o valor estimado de y para uma amostra i, pode ser definido como o intervalo no qual se pode afirmar com certo grau de confiança, ou probabilidade, que inclui o valor verdadeiro da propriedade de interesse. O cálculo desses intervalos depende de uma estimativa razoável da variância dos erros de previsão para amostras desconhecidas. Em modelos de calibração multivariada uma determinação razoável de uma estimativa dessas variâncias não constitui uma tarefa simples, pois, ao contrário de modelos univariados (que utilizam uma estimativa de variância que representa a distribuição dos erros que é esperada para novas amostras), modelos de calibração multivariada, na maioria das vezes, necessitam de uma equação que fornece uma estimativa de variância específica para cada nova amostra que está

sendo prevista. Este intervalo pode ser determinado através de um teste T e uma estimativa para a variância dos erros de previsão (V(PE)).

A V(PE) pode ser determinada através da teoria dos erros nas variáveis (EIV – do inglês, Error in Variables) também conhecida como método de propagação de erros[87], que apresenta uma simplificação abordada pela norma E1655-00 da ASTM[61], cuja equação para o cálculo da variância dos erros de previsão é expressa como:

$$V(PE) = \left(1 + h_i + \frac{1}{n}\right) MSEC_p \tag{74}$$

em que, $MSEC_p$ (do inglês, Pseudo Mean Square Error of Calibration) é o pseudo erro médio quadrático da calibração calculado a partir da equação:

$$MSEC_p = \sum_{i=1}^{n} \frac{(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{n_{GL}} \tag{75}$$

em que, n_{GL} corresponde ao número de graus de liberdade determinado pela abordagem dos pseudo-graus de liberdade proposto por Van der Voet[88].

O número efetivo de graus de liberdade envolvidos no cálculo do MSEC consiste em mais uma dificuldade para o cálculo dos intervalos de confiança. Nesse sentido H. Van der Voet[88] definiu o conceito de pseudograus de liberdade (PDF – do inglês, Pseudo-Degrees of Freedom), que leva em consideração a diferença entre o erro médio quadrático de calibração estimado por validação cruzada (MSECV – do inglês, Mean Square Error of Cross Validation) e pela previsão das próprias amostras de calibração (MSEC – do inglês, Mean Square Error of Calibration), de modo que quanto maior for essa diferença, menor será n_{GL}:

$$n_{GL} = n\left(1 - \sqrt{\frac{MSEC}{MSECV}}\right) \tag{76}$$

Após o cálculo da variância, com um número apropriado de graus de liberdade, os limites de confiança (ϕ) podem ser obtidos através de:

$$\phi_i = \pm t_{1-\alpha/2nGL} \sqrt{V(PE)_i} \quad (77)$$

em que, $t_{1-\alpha/2nGL}$ é o parâmetro estatístico da distribuição t-student com probabilidade $(1-\alpha)/2nGL$ de recobrimento.

Os limites calculados admitem que os erros correspondentes às concentrações estimadas não são correlacionados e seguem uma distribuição normal. É importante que os limites de confiança estimados sejam consistentes e cubram o intervalo esperado para aquele nível de probabilidade, ou seja, para o nível de confiança de 95 %, por exemplo, 95 de 100 amostras devem ter o valor verdadeiro da propriedade dentro do intervalo de confiança calculado.

4.3.13. Teste para erros sistemáticos (Bias)

De acordo com a definição da IUPAC, erros sistemáticos são calculados pela diferença entre a média da população e o valor verdadeiro e são todas as componentes de erro que não são aleatórias[67]. A existência desse tipo de erro afeta a precisão, exatidão e a determinação dos intervalos de confiança[89,90]. A norma E1655-00 da ASTM[61] aborda a investigação de erros sistemáticos em modelos de calibração multivariada através de um teste-t, para as amostras de validação no nível de 95% de confiança para avaliar quantitativamente se o bias presente no modelo é significativo. Para este teste, primeiro um bias médio para o conjunto de validação é calculado pela equação[61]:

$$\text{bias} = \frac{\sum_{i=1}^{I_v} (y_i - \hat{y}_i)}{I_v} \quad (78)$$

em que, I_v corresponde ao número de amostras do conjunto de validação.

A seguir, o desvio padrão dos erros de validação (SDV – do inglês, Standard Deviation of Validation) é obtido por[61]:

$$SDV = \sqrt{\frac{\sum[(y_i - \hat{y}_i) - bias]^2}{I_v - 1}} \tag{79}$$

Por fim, o valor de t é dado por[61]:

$$t_{bias} = \frac{|bias|\sqrt{I_v}}{SDV} \tag{80}$$

Caso o valor de t_{bias} apresentar resultado maior do que o valor de t crítico para I_v-1 graus de liberdade, onde I_v é o número de amostras da validação, com 95% de confiança, isso é uma evidência de que erros sistemáticos presentes no modelo multivariado são significativos. No entanto, se o valor de t_{bias} calculado apresentar valor menor do que o valor crítico, então, o erro sistemático incluído no modelo pode ser considerado insignificante e desprezado.

CAPÍTULO 5
Aplicação

5. Aplicação

5.1. Objetivos

Essa Dissertação tem por objetivos:

- A construção de modelos de calibração multivariada empregando o método de regressão por PLS para a quantificação do Brix, Pol e AR utilizando espectroscopia no infravermelho próximo.
- Seleção de variáveis em calibração multivariada através do método iPLS para a quantificação do Brix, Pol e AR utilizando espectroscopia no infravermelho próximo.
- Validação dos modelos de calibração multivariada pela determinação das figuras de mérito.
- Comparação dos modelos construídos a partir do espectro todo com os modelos construídos a partir das variáveis selecionadas pelo iPLS.

5.2. Parte experimental

A parte experimental deste trabalho foi realizada na usina de álcool da Cocamar – Cooperativa Agroindustrial, localizada na cidade de São Tomé, no estado do Paraná - Brasil.

A cana-de-açúcar madura chega na indústria transportada por caminhões e é amostrada através de uma sonda amostradora horizontal. A amostra é decomposta (na indústria utiliza-se o termo desintegrada) e segue para o laboratório. Quando se trata de amostras de cana-de-açúcar verde para ensaios de pré-colheita as amostras são coletadas na lavoura por técnicos especializados e na unidade industrial são decompostas antes de seguir para o laboratório.

No laboratório, as amostras são prensadas a 250Kgf/cm^2 em uma prensa hidráulica por um período de 1 minuto, resultando o caldo da cana para as análises subseqüentes.

Os espectros foram coletados em espectrômetro de infravermelho próximo NIRSystems, marca FOSS Perstorp Analytical, modelo 5000 Monocromador, tendo como referência interna uma placa de poliestireno e equipado com uma

fonte de luz de filamento de tungstênio, cubeta de quartzo, caminho ótico de 1 mm e detector PbS. A aquisição dos espectros deu-se a partir das amostras do caldo da cana na faixa de 1100 – 2500 nm, de 2 em 2 nm, através do software ISIScan. As amostras sofreram um pré-tratamento de filtragem em algodão para eliminar partículas suspensas.

Um total de 1381 amostras de caldo de cana-de-açúcar foram coletadas. Cada amostra foi submetida às análises convencionais, em triplicata, para determinação do Brix (por densímetro), Pol (por sacarímetro) e AR (por titulação de oxidação-redução), seguido por aquisição do espectro na região do infravermelho próximo. Na determinação do Pol as amostras de caldo de cana-de-açúcar foram clarificadas com uma mistura de acetato e hidróxido de chumbo, (na indústria conhecida como subacetato de chumbo ($Pb(CH_3COO)_2.Pb(OH)_2$)) e filtradas em papel de filtro pregueado. O grau de polarização das amostras (%caldo) foi determinado no caldo de cana clarificado através da leitura sacarimétrica (LS) utilizando um sacarímetro digital e a equação (1)[4]:

Na determinação dos açúcares redutores (AR) por metodologia padrão utilizou-se o método proposto por Eynon & Lane[4,5] que consiste na titulação de oxidação-redução do licor de Fehling pelo caldo de cana filtrado. As substâncias redutoras do caldo de cana (glicose e frutose) reduzem do cobre de Cu^{2+} para Cu_2O do licor Fehling, tendo como indicador do ponto final da titulação o azul de metileno a 1%. O teor de açúcares redutores (% caldo) presente em cada amostra foi obtido de acordo com a equação (2).

O conjunto de dados, composto por 1381 amostras, foi dividido em conjuntos de calibração e validação através do algoritmo de Kennard-Stone[48]. Os espectros sofreram pré-processamento para ficarem centrados na média, seguida pela eliminação de uma banda intensa na região de 1900 nm (1890-2046 nm), correspondente à absorção da água[16]. A seleção de variáveis foi realizada através do método iPLS. Modelos de calibração para os parâmetros Brix, Pol e AR foram desenvolvidos através do software MatLab 6.5 usando o pacote PLS-Toolbox 2.1 baseado no método PLS1 tanto para modelos desenvolvidos com o espectro inteiro (623 variáveis) como para os modelos desenvolvidos com as variáveis do

espectro selecionadas pelo iPLS (125 variáveis). Os conjuntos de calibração foram otimizados pela eliminação de amostras anômalas com *leverage* extremo[30,61] determinado pela equação (48), amostras anômalas identificados através dos resíduos espectrais[30] calculados pelas equações (49) e (50) e amostras anômalas com resíduos significativos na variável dependente[30] determinados pela equação (51). Os conjuntos de validação foram otimizados a partir dos conjuntos de calibração já otimizados. Para esses conjuntos de validação além do *leverage* extremo e dos resíduos espectrais, foram identificados os resíduos da repetibilidade espectral[61] de amostras em três níveis de concentração, cobrindo a faixa útil do modelo, com sete replicatas em cada nível, que foram identificados através das equações (52) e (53).

Para a validação dos modelos desenvolvidos determinou-se as figuras de mérito: "*bias"*, exatidão, precisão, sensibilidade, sensibilidade analítica, seletividade, linearidade, ajuste e ajuste através do NAS, razão sinal/ruído, limite de detecção, limite de quantificação e intervalo de confiança.

5.3. Resultados e discussão

Para os resultados do método padrão, realizado com três replicatas para cada amostra, utilizou-se o valor médio de cada resultado para compor o vetor **y** de valores de referência. A matriz **X** foi composta pelas absorbâncias para cada amostra, onde cada linha da matriz é representada por uma amostra diferente e as colunas representam as respectivas absorbâncias para a amostra em questão.

A separação das amostras entre os conjuntos de calibração e validação realizado pelo algoritmo de Kennard-Stone, resultou em um total de 1000 amostras para o conjunto de calibração e 378 amostras para o conjunto de validação. Três amostras foram acrescentadas ao final do conjunto de calibração, sendo que estas ultimas não foram selecionadas pelo algoritmo, e foram introduzidas somente devido ao teste de *outliers* baseado na repetibilidade espectral para a validação, conforme descrito na norma E1655-00 da ASTM, o qual é baseado em amostras em três níveis de concentração conforme a equação (53)[61].

A Figura 7 ilustra os espectros, na região do infravermelho próximo, do caldo de cana-de-açúcar para todas as amostras, enquanto que a Figura 8 mostra os referidos espectros após a eliminação da faixa entre 1890 a 2046 nm a qual é referente à forte absorção da água. Para essa região eliminada, observou-se uma baixa relação sinal/ruído que contribui para má previsão dos modelos de calibração multivariada.

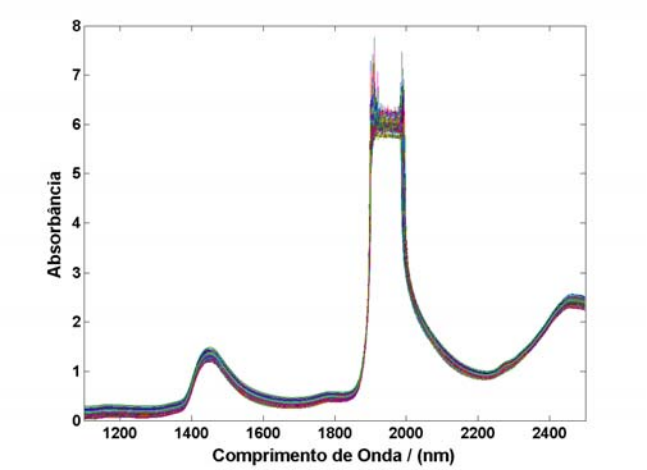


Figura 7. Espectros para as amostras de caldo de cana-de-açúcar

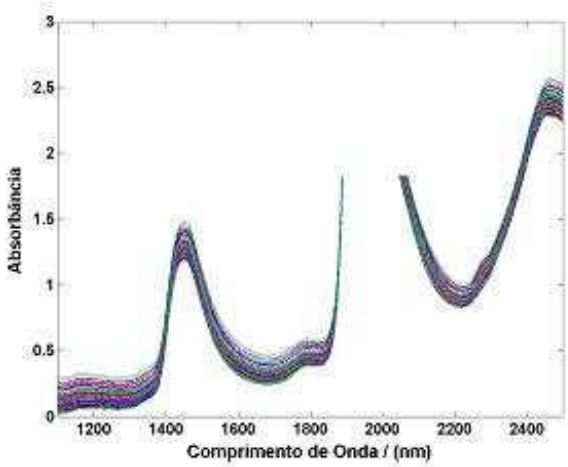


Figura 8. Espectros para as amostras de caldo de cana-de-açúcar após a eliminação da região compreendida entre 1890 – 2046 nm

A partir de então, quando se faz menção ao espectro inteiro, na realidade refere-se ao espectro cuja região mencionada foi eliminada.

A otimização dos conjuntos de calibração e validação pela eliminação das amostras anômalas resultaram em 897, 924 e 857 amostras de calibração e 362, 358 e 368 amostras de validação para os parâmetros Brix, Pol e AR dos modelos PLS construídos com o espectro inteiro, respectivamente. A Tabela 2 mostra os resultados para os testes de identificação de amostras anômalas nos conjuntos de calibração e validação bem como os valores de RMSEC e RMSEP para os modelos à medida que os *outliers* são eliminados (Mod1= primeiro modelo; Mod2= segundo modelo; Mod3= terceiro modelo). Foram realizados os testes para a identificação de *outliers* até o segundo modelo de calibração e as anomalias da validação foram avaliadas com base no terceiro modelo de calibração tido como otimizado. Os valores para o *leverage* limite (h_{lim}) obtidos para cada modelo como $3\frac{A+1}{n}$ (onde, A é o número de variáveis latentes e n o número de amostras da calibração) são apresentados, mostrando que já para os primeiros modelos de calibração os valores para o h_{lim} são inferiores a 0,5 que é o valor apresentado como limite, segundo a norma E1655-00 da ASTM, para o critério de parada dos testes de identificação de outliers no segundo modelo. Nas amostras de calibração quando se calcula os valores para os resíduos na variável dependente **y** conforme a equação (51) tem-se como valor limite (3RMSEC). Estes valores, para o resíduo na variável dependente, são apresentados na Tabela 2 como o valor médio dos resíduos na variável dependente ($\overline{y}_{lim}$) para cada modelo. Para as amostras de validação, quando o erro absoluto da previsão é maior do que o valor de $\overline{y}_{lim}$ do terceiro modelo de calibração, tido como otimizado, estas amostras são eliminadas do conjunto. É possível observar que os valores para a raiz do erro médio quadrático da calibração (RMSEP) e validação (RMSEP) diminuem significativamente nos modelos otimizados, indicando que os modelos tornaram-se mais eficientes apresentando maior exatidão.

Tabela 2. Resultados para os testes de identificação de outliers para os modelos $PLS_{espectro\ inteiro}$

Testes / Modelos	Nº Amostras	Nº anomalias baseadas no Leverage	h_{lim}	Nº anomalias baseadas no resíduo espectral	Nº anomalias baseadas no Resíduo da variável dependente	$\bar{y}_{lim}$	Nº anomalias baseadas na repetibilidade espectral	Nº total de anomalias descartadas	RMSEC	RMSEP
Brix Mod1	1003	17	0,0150	36	15	1,9212	n.a.	63	0,6404	0,7563
Brix Mod2	940	18	0,0160	17	12	1,0457	n.a.	43	0,3486	0,7629
Brix Mod3 Otimizado	897	9	0,0167	7	7	0,9105	n.a.	22	0,3035	0,7613
$Brix_{Validação}$	378	2	0,0167	0	14	0,9105	0	16	0,3035	0,7613
$Brix_{Validação}$ Otimizado	362	0	0,0167	0	0	0,9105	0	0	0,3035	0,2805
Pol Mod1	1003	17	0,0209	7	20	2,5215	n.a.	44	0,8405	0,9584
Pol Mod2	959	16	0,0219	4	18	1,4975	n.a.	35	0,4992	0,9705
Pol Mod3 Otimizado	924	11	0,0227	2	10	1,3111	n.a.	23	0,4370	0,9727
$Pol_{Validação}$	378	1	0,0227	0	19	1,3111	0	20	0,4370	0,9727
$Pol_{Validação}$ Otimizado	358	0	0,0227	0	0	1,3111	0	0	0,4370	0,4167
AR Mod1	1003	25	0,0150	47	17	1,0226	n.a.	85	0,3409	0,3278
AR Mod2	918	21	0,0163	35	8	0,8788	n.a.	61	0,2929	0,3264
AR Mod3 Otimizado	857	7	0,0175	29	2	0,8292	n.a.	38	0,2702	0,3263
$AR_{Validação}$	378	3	0,0175	0	7	0,8292	0	10	0,2702	0,3263
$AR_{Validação}$ Otimizado	368	0	0,0175	0	0	0,8292	0	0	0,2702	0,2601

Rep. = Repetibilidade; n.a.= não se aplica

A seleção de variáveis realizada pelo método iPLS resultou no intervalo de 1600 a 1850 nm que corresponde às variáveis de 251 a 375, conforme pode ser observado na Figura 9. Nessa região é predominante a absorção relativa ao primeiro sobretom dos grupos C-H presente nos açúcares[13].

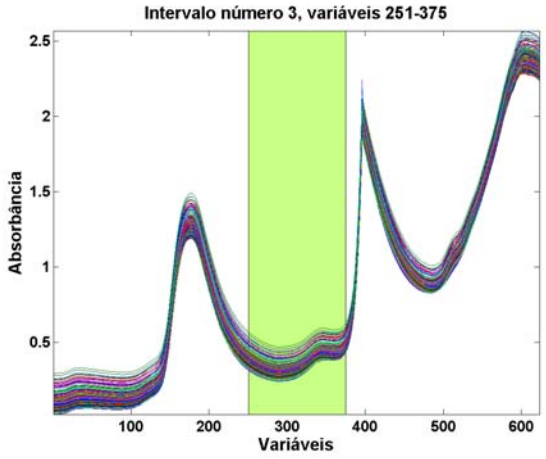


Figura 9. Variáveis selecionadas pelo método iPLS

Os conjuntos de calibração e validação, formado pelas variáveis selecionadas pelo iPLS, foram otimizados pela eliminação das amostras anômalas resultando em 893, 914 e 891 amostras de calibração e 358, 353 e 369 amostras de validação para os parâmetros Brix, Pol e AR, respectivamente. A Tabela 3 mostra os resultados para os testes de identificação de anomalias nos conjuntos de calibração e validação bem como os valores de RMSEC e RMSEP para os modelos à medida que as amostras anômalas são eliminadas. Foram realizados os testes para a identificação de anomalias até o segundo modelo de calibração e as amostras anômalas da validação foram avaliadas com base no terceiro modelo de calibração tido como otimizado. Os valores para h_{lim} e $\bar{y}_{lim}$ também são apresentados. Para os modelos construídos a partir das variáveis selecionadas pelo iPLS também foi possível observar que os valores para a raiz do erro médio quadrático da calibração (RMSEC) e validação (RMSEP) diminuem significativamente nos modelos otimizados, indicando que os modelos tornaram-se mais eficientes apresentando maior exatidão.

Tabela 3. Resultados para os testes de identificação de outliers para os modelos iPLS

Testes / Modelos	Nº Amostras	Nº anomalias baseadas no Leverage	h_{lim}	Nº anomalias baseadas no resíduo espectral	Nº anomalias baseadas no Resíduo da variável dependente	$\bar{y}_{lim}$	Nº anomalias baseadas na repetibilidade espectral	Nº total de anomalias descartadas	RMSEC	RMSEP
Brix Mod1	1003	15	0,0179	42	14	1,8840	n. a.	66	0,6280	0,7716
Brix Mod2	937	14	0,0192	22	13	0,9433	n. a.	44	0,3144	0,7779
Brix Mod3 Otimizado	893	5	0,0202	12	13	0,8151	n. a.	28	0,2717	0,7805
$Brix_{Validação}$	378	2	0,0202	0	18	0,8151	0	20	0,2717	0,7805
$Brix_{Validação}$ Otimizado	358	0	0,0202	0	0	0,8151	0	0	0,2717	0,2919
Pol Mod1	1003	26	0,0239	21	17	2,3574	n. a.	53	0,7858	0,9058
Pol Mod2	950	13	0,0253	10	15	1,1006	n. a.	36	0,3668	0,9176
Pol Mod3 Otimizado	914	12	0,0263	2	4	0,8858	n. a.	17	0,2953	0,9187
$Pol_{Validação}$	378	6	0,0263	0	21	0,8858	0	25	0,2953	0,9187
$Pol_{Validação}$ Otimizado	353	0	0,0263	0	0	0,8858	0	0	0,2953	0,2681
AR Mod1	1003	21	0,0179	45	11	0,9338	n. a.	69	0,3113	0,3029
AR Mod2	934	14	0,0193	23	6	0,8344	n. a.	43	0,2781	0,3136
AR Mod3 Otimizado	891	5	0,0202	8	1	0,8075	n. a.	14	0,2692	0,3181
$AR_{Validação}$	378	2	0,0262	0	7	0,8075	0	9	0,2692	0,3181
$AR_{Validação}$ Otimizado	369	0	0,0262	0	0	0,8075	0	0	0,2693	0,2555

Rep. = Repetibilidade; n.a.= não se aplica

Os resultados apresentados nas Tabelas 2 e 3 podem ser melhor visualizados através das Figuras apresentadas a seguir.

As Figuras 10, 11 e 12 ilustram os histogramas para o *Leverage* do primeiro modelo de calibração dos parâmetros Brix, Pol e AR nos modelos construídos com o espectro todo e nos modelos iPLS, respectivamente. A partir dos valores de h_{lim} apresentados nas Tabelas 2 e 3 as amostras foram excluídas por se tratarem de anomalias do conjunto de calibração.

a) b)

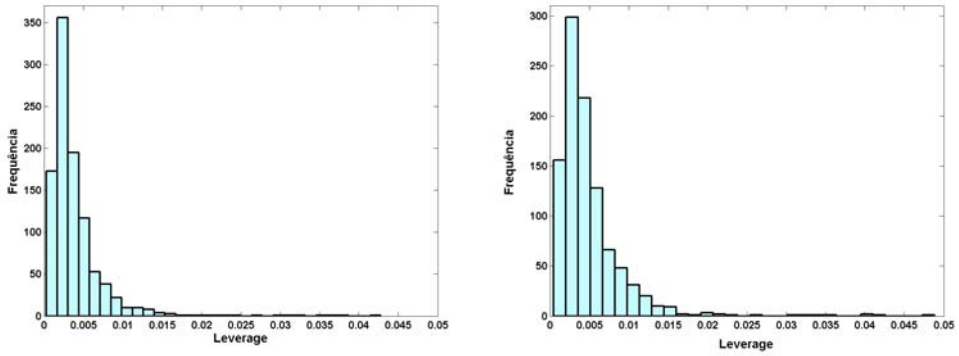


Figure 10. Valores de *Leverage* para o Brix no primeiro modelo.
(a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

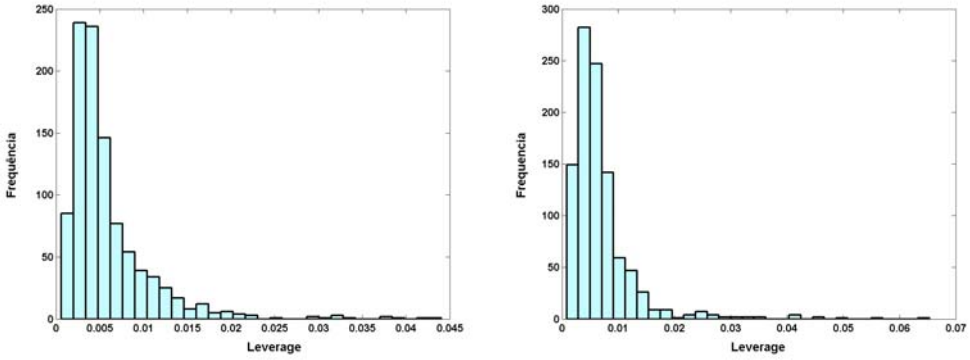


Figure 11. Valores de *Leverage* para o Pol no primeiro modelo.
(a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

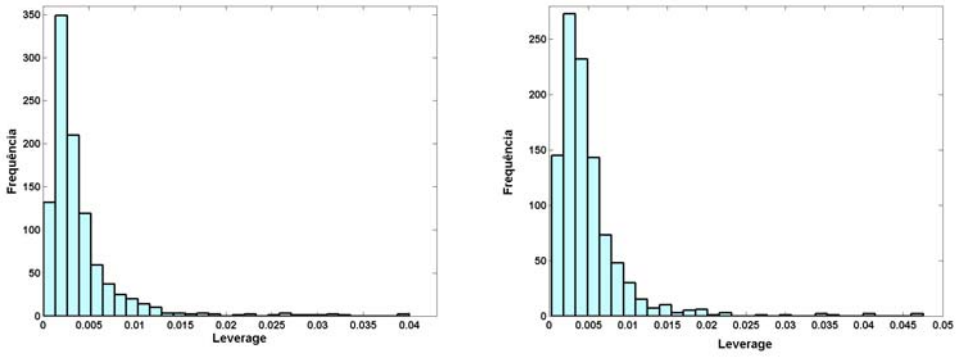


Figure 12. Valores de *Leverage* para o AR no primeiro modelo.
(a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

As Figuras 13, 14 e 15 ilustram as amostras identificadas como anômalas com base nos testes de *Leverage* e do resíduo espectral, enquanto que as Figuras 16, 17 e 18 mostram as anomalias identificadas com base na variável dependente no primeiro modelo construído com o espectro inteiro e no primeiro modelo construído com as variáveis selecionadas pelo iPLS, para os parâmetros Brix, Pol e AR, respectivamente. São mostrados os valores limites do Leverage, do resíduo espectral total e dos resíduos na variável dependente através das linhas em verde nas Figuras. As amostras que ficaram acima destes limites, em ambos os casos, foram eliminadas do conjunto de calibração por serem consideradas como anomalias.

a) b)

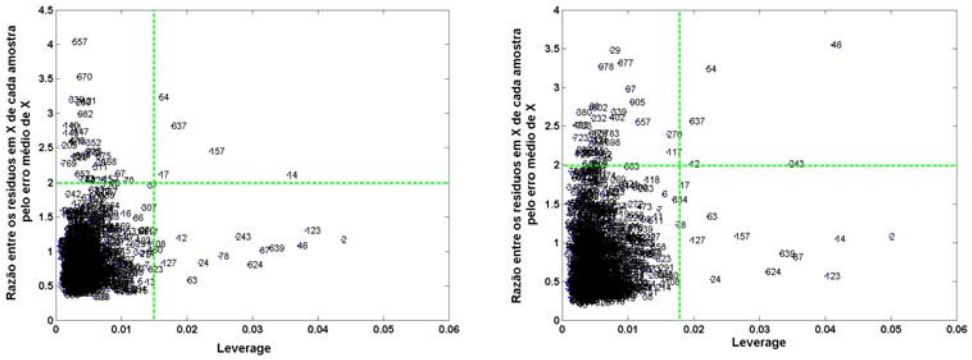


Figure 13. Amostras anômalas do primeiro modelo de calibração do parâmetro Brix identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

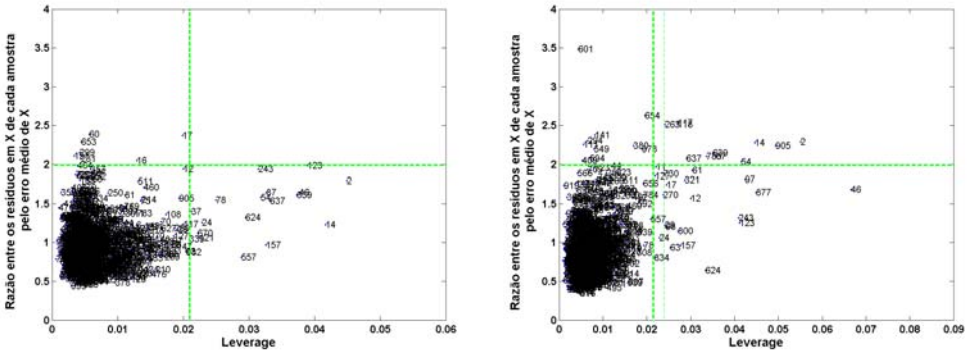


Figure 14. Amostras anômalas do primeiro modelo de calibração do parâmetro Pol identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

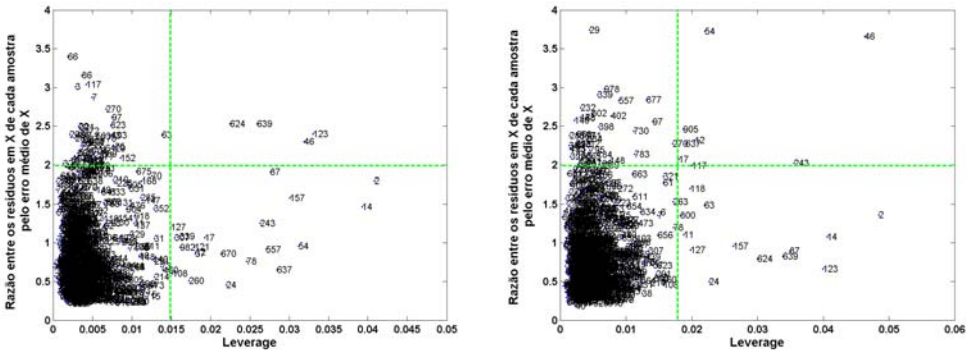


Figure 15. Amostras anômalas do primeiro modelo de calibração do parâmetro AR identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

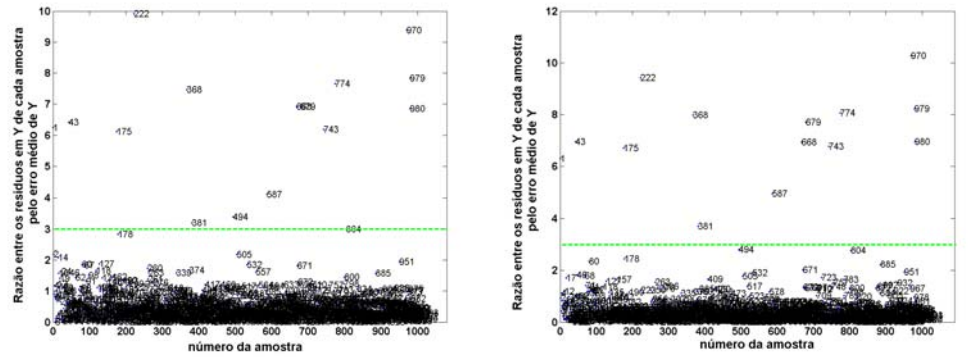


Figure 16. Amostras anômalas do primeiro modelo de calibração do parâmetro Brix identificados com base na variável dependente. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

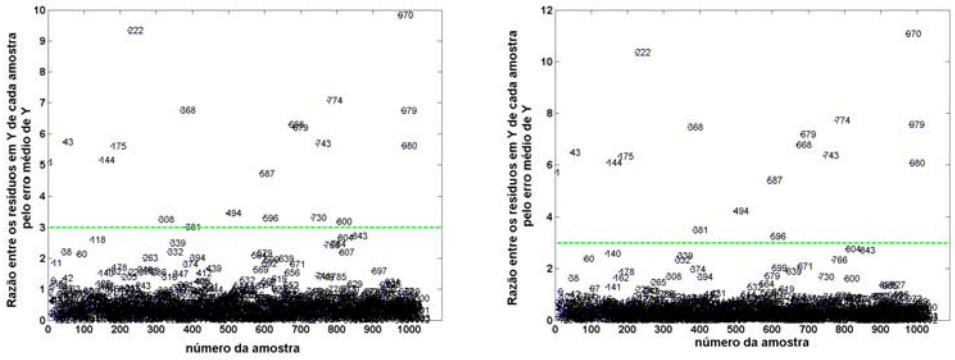


Figure 17. Amostras anômalas do primeiro modelo de calibração do parâmetro Pol identificados com base na variável dependente. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

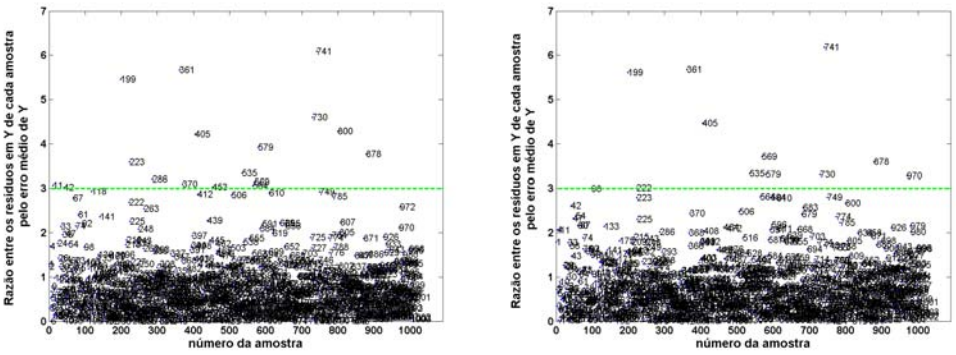


Figure 18. Amostras anômalas do primeiro modelo de calibração do parâmetro AR identificados com base na variável dependente. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

Para as amostras anômalas do conjunto de validação, as Figuras 19, 20 e 21 ilustram sua identificação com base no *leverage* e no resíduo espectral, enquanto que as Figuras 22, 23 e 24 mostram as amostras identificadas como anômalas com base no teste de repetibilidade espectral para os parâmetros Brix, Pol e AR, respectivamente. Para a identificação das amostras anômalas da validação a previsão das amostras foi realizada com base no terceiro modelo de calibração tido como otimizado. São mostrados os valores limites do leverage, do resíduo espectral total e dos resíduos da repetibilidade espectral através das linhas em verde nas Figuras. As amostras que ficaram acima destes limites foram eliminadas do conjunto de validação, em ambos os casos, por serem consideradas como anomalias.

a) b)

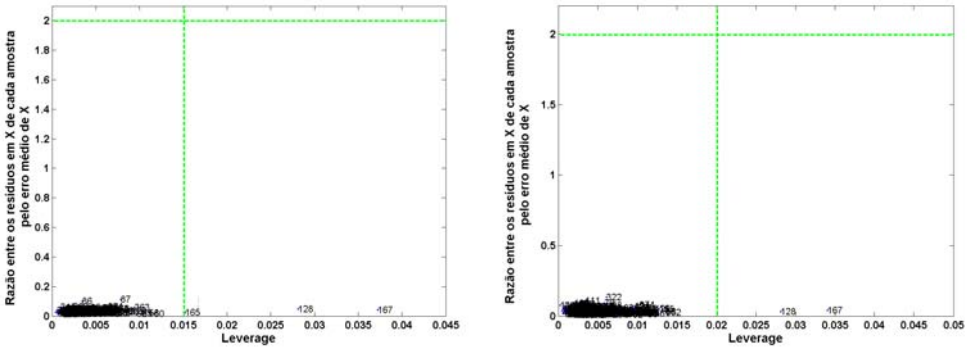


Figure 19. Amostras anômalas do conjunto de validação do parâmetro Brix identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

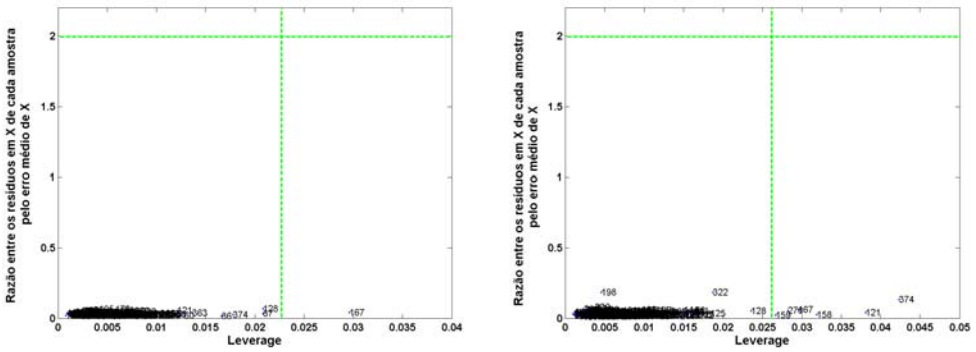


Figure 20. Amostras anômalas do conjunto de validação do parâmetro PoI identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

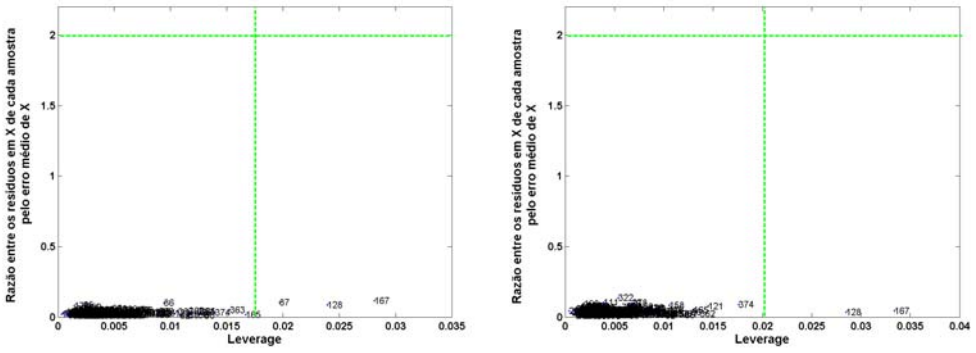


Figure 21. Amostras anômalas do conjunto de validação do parâmetro AR identificados com base no *Leverage* e no resíduo espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)

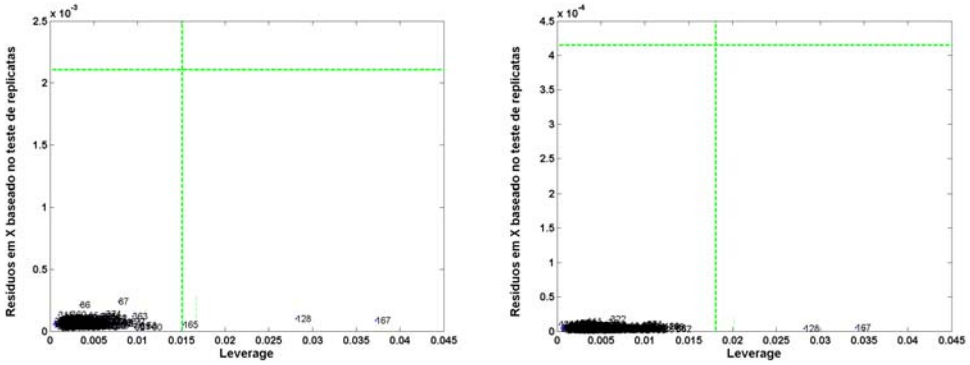
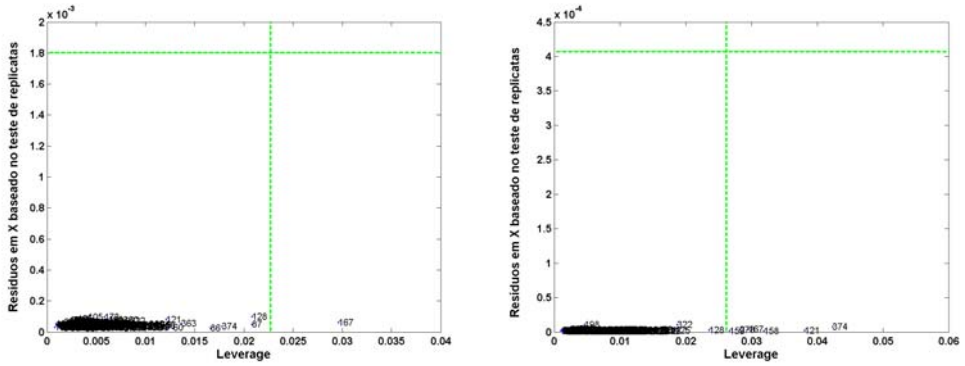


Figure 22. Amostras anômalas do conjunto de validação do parâmetro Brix identificados com base na repetibilidade espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

a) b)



(a) (b)

Figure 23. Amostras anômalas do conjunto de validação do parâmetro Pol identificados com base na repetibilidade espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

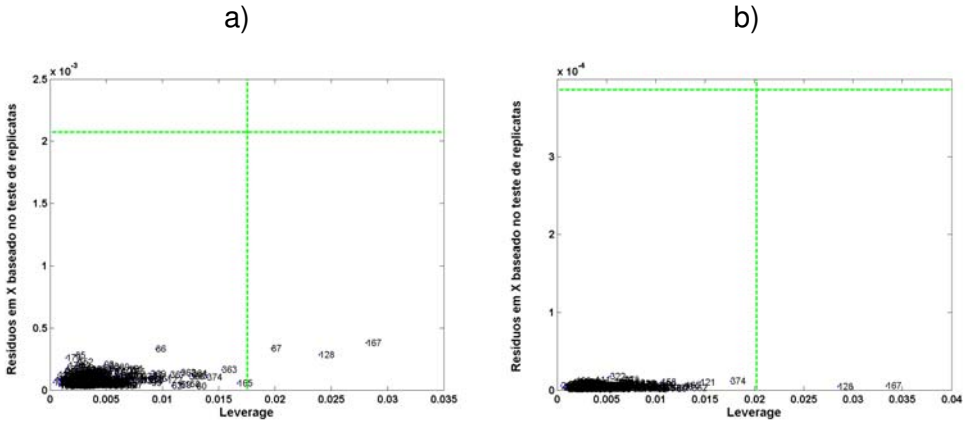


Figure 24. Amostras anômalas do conjunto de validação do parâmetro AR identificados com base na repetibilidade espectral. (a) – Modelo com o espectro inteiro e (b) – Modelo iPLS

O número de variáveis latentes (VLs) para cada modelo de calibração multivariada, foi determinado através dos resultados da raiz quadrada do erro quadrático médio de validação cruzada (RMSECV) para as amostras de calibração, obtido por validação cruzada em blocos contínuos de 10 amostras e levando-se em consideração a presença de bias relevante, testada com a previsão dos resultados para as amostras de validação através do teste-t sugerido pela norma E1655-00 da ASTM, para o número de VLs determinado. Assim, um total de 4, 6 e 4 VLs para os modelos utilizando o espectro inteiro, e, 5, 7 e 5 VLs para os modelos iPLS, para os parâmetros Brix, Pol e AR, respectivamente, foram necessários para conservar uma variância significativa dos dados e evitar elevado bias. Os valores de t_{bias} calculados pela equação (80) apresentaram valores de 2,07, 1,37 e 2,17 para o espectro inteiro e 1,94, 2,08 e 2,20 para os modelos iPLS dos parâmetros Brix, Pol e AR, respectivamente. Comparando esses resultados com o valor crítico de 1,96 com 95% de confiança, verifica-se que a seleção de variáveis produz um modelo com bias não significativo para o caso do parâmetro Brix. Para o parâmetro Pol a presença de bias irrelevante é observada no modelo construído a partir do espectro inteiro. Já para o parâmetro AR o bias é significativo em ambos os modelos. A presença de bias significativo para o parâmetro AR pode ser devido a erros inerentes ao método de referência que é

baseado na titulação de oxidação-redução, ou então a não linearidades no modelo de calibração.

Os resultados das figuras de mérito são mostrados na Tabela 4. Os valores de exatidão são representados pela raiz quadrada do erro quadrático médio da calibração (RMSEC) e da previsão (RMSEP) e revelam que os valores estimados pelos modelos multivariados apresentaram uma boa concordância com os métodos de referência.

A precisão, no nível de repetibilidade, foi calculada pela análise de amostras em três níveis de concentração, cobrindo a faixa de concentração utilizada para a construção dos modelos, com seis replicatas cada nível, e todas as determinações foram realizadas no mesmo dia. Os resultados para Brix e Pol mostraram que a repetibilidade dos modelos multivariados foram melhores do que os valores regulamentados por normas de avaliação da qualidade da cana-de-açúcar que é de 0,3 % caldo para o Brix e de 0,6 % caldo para o Pol. Para o AR também foi observado um bom resultado, no entanto, não existe uma norma que regulamenta a precisão deste parâmetro, isto porque para o pagamento do fornecedor de cana na indústria, este parâmetro não é determinado experimentalmente, sendo apenas estimado pela equação (4) que leva em consideração os parâmetros Brix e Pol.

A sensibilidade e sensibilidade analítica apresentaram bons resultados para os três parâmetros considerando-se a faixa de concentrações utilizada no trabalho. Como a sensibilidade analítica é mais simples e informativa para comparações e julgamento de um método analítico, uma vez que esse parâmetro apresenta, de forma direta, a sensibilidade do método em termos da unidade de concentração que é utilizada, o inverso desse parâmetro permite estabelecer a menor diferença de concentração entre amostras que pode ser distinguida pelo método. Neste sentido, por exemplo para o parâmetro Brix, é possível fazer a distinção de amostras com diferença de concentração em torno de $0{,}22 \times 10^{-2}$ % de Brix no caldo em modelos construídos com o espectro inteiro. O inverso da sensibilidade analítica para os modelos construídos com as variáveis selecionadas pelo iPLS mostraram valores maiores do que os apresentados por modelos

construídos com o espectro inteiro. A sensibilidade reduzida para os modelos construídos utilizando as variáveis selecionadas com o iPLS, pode ser justificada pelos baixos valores de absorbância da região espectral selecionada.

Os valores de seletividade referem-se à parte do sinal que é perdida devido à sobreposição entre o sinal do analito de interesse com outros componentes presentes na amostra. Levando em consideração a definição do NAS, o vetor $\hat{\mathbf{x}}_{k,i}^{nas}$ é a parte do sinal que é ortogonal à matriz de interferentes ($\hat{\mathbf{X}}_{A,-k}$) e os valores de seletividade apresentados na Tabela 4 foram determinados através da razão entre o escalar $nâs_k$ e a norma do vetor de resposta instrumental representando quanto do sinal é utilizado para a quantificação dos parâmetros. Os valores para δx, tomados em absorbância, obtidos pelo desvio padrão do sinal de referência, que para o equipamento utilizado consiste em uma placa de poliestireno, foram de $1{,}4152 \times 10^{-4}$ e $1{,}4277 \times 10^{-4}$ para o espectro inteiro e para a região espectral utilizada na construção dos modelos iPLS, respectivamente. Assim, por exemplo, o valor de seletividade de 0,30 para o parâmetro Brix no modelo construído com o espectro inteiro, indica que 70% do sinal foi perdido por não ser ortogonal ao sinal referente ao parâmetro Brix, neste caso. Os valores de seletividade mostraram um decréscimo para os modelos iPLS em relação aos modelos construídos com o espectro inteiro, indicando que a região selecionada ainda contém uma grande quantidade de informação irrelevante.

Tabela 4. Figuras de mérito

Figuras de Mérito		Brix	Pol	AR
Exatidão[a]	**RMSEC** espectro inteiro	0,30	0,44	0,28
	RMSEC modelo iPLS	0,27	0,29	0,27
	RMSEP espectro inteiro	0,28	0,42	0,26
	RMSEP modelo iPLS	0,29	0,27	0,25
Precisão[a] espectro inteiro		0,02	0,08	0,08
Precisão[a] modelo iPLS		0,02	0,03	0,01
Sensibilidade[b] espectro inteiro		0,06	0,02	0,32
Sensibilidade[b] modelo iPLS		$0,23x10^{-2}$	$0,58x10^{-3}$	$0,51x10^{-2}$
Sensibilidade Analítica[a] espectro inteiro		$0,22x10^{-2}$	$0,87x10^{-2}$	$0,23x10^{-3}$
Sensibilidade Analítica[a] modelo iPLS		$6,33x10^{-2}$	0,24	$2,80x10^{-2}$
Seletividade espectro inteiro		0,30	$9,56x10^{-2}$	0,27
Seletividade modelo iPLS		$4,99x10^{-2}$	$1,58x10^{-2}$	$2,37x10^{-2}$
Ajuste	**Inclinação** espectro inteiro	0,99±0,01	0,99±0,01	0,76±0,01
	Inclinação modelo iPLS	0,99±0,01	0,99±0,01	0,76±0,01
	Intercepto espectro inteiro	0,18±0,06	0,23±0,02	0,19±0,01
	Intercepto modelo iPLS	0,15±0,06	0,10±0,04	0,19±0,01
	Coef. Corr. (R^2) espectro inteiro	0.99	0,99	0,76
	Coef. Corr. (R^2) modelo iPLS	0,99	0,99	0,76
Ajuste NAS	**Inclinação** espectro inteiro	15,00±0,06	56,00±4,51	2,90±0,05
	Inclinação modelo iPLS	$(4,30 \pm 0,01) x10^2$	$(1,60 \pm 0,05) x10^2$	$(1,40 \pm 0,02) x10^2$
	Intercepto espectro inteiro	8,90±0,04	4,40±1,01	-0,21±0,02
	Intercepto modelo iPLS	8,70±0,04	4,00±0,04	-0,13±0,02
	Coef. Corr. (R^2) espectro inteiro	0,99	0,99	0,81
	Coef. Corr. (R^2) modelo iPLS	0,99	0,99	0,81
Razão Sinal/Ruído	**Máx.** espectro inteiro	$6,69x10^3$	$2,18x10^3$	$6,05x10^3$
	Máx. modelo iPLS	233,61	77,04	114,54
	Min. espectro inteiro	218,99	24,58	952,75
	Min modelo iPLS	10,13	1,98	6,73
Limite Detecção[a] espectro inteiro		$0,69x10^{-2}$	$2,62x10^{-2}$	$0,13x10^{-2}$
Limite Detecção[a] modelo iPLS		0,19	0,74	$8,41x10^{-2}$
Limite Quantificação[a] espectro inteiro		0,02	0,09	$0,44x10^{-2}$
Limite Quantificação[a] modelo iPLS		0,63	2,45	0,28

a = % caldo e b= % caldo^{-1}

O ajuste dos modelos construídos foi avaliado com base nos gráficos dos valores para os parâmetros estimados pelos modelos PLS contra os valores de referência. Outra maneira de avaliar o ajuste de modelos de calibração multivariada é através dos gráficos dos valores escalares do sinal analítico líquido (NAS), determinado pela norma do vetor de sinal analítico líquido em função do valor de referência para cada parâmetro. Esta última refere-se à representação pseudo-univariada dos modelos de calibração multivariada. A inclinação, o intercepto e o coeficiente de correlação para os modelos são mostrados na Tabela 4. As Figuras 25, 26 e 27 mostram os ajustes (referência versus estimado e ajuste NAS) para os modelos construídos com os espectros inteiros, enquanto as Figuras 28, 29 e 30 ilustram os ajustes para os modelos iPLS dos parâmetros Brix, Pol e AR, respectivamente. Através dos gráficos e dos resultados apresentados na Tabela 4, observa-se que os modelos para Brix e Pol apresentam um ajuste similar e claramente superior ao observado para o AR. O ajuste inferior para os açúcares redutores pode ser um indicativo de que este parâmetro apresenta uma relação não linear com os dados espectrais.

a) b)

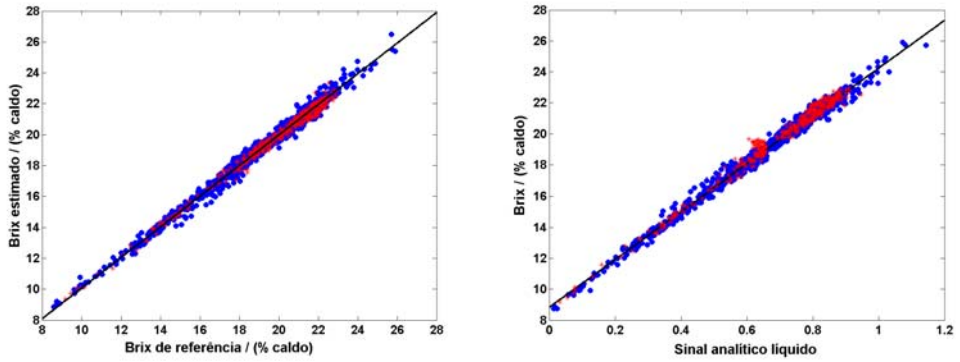


Figure 25. Ajuste para o parâmetro Brix do modelo construído com o espectro inteiro. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

a) b)

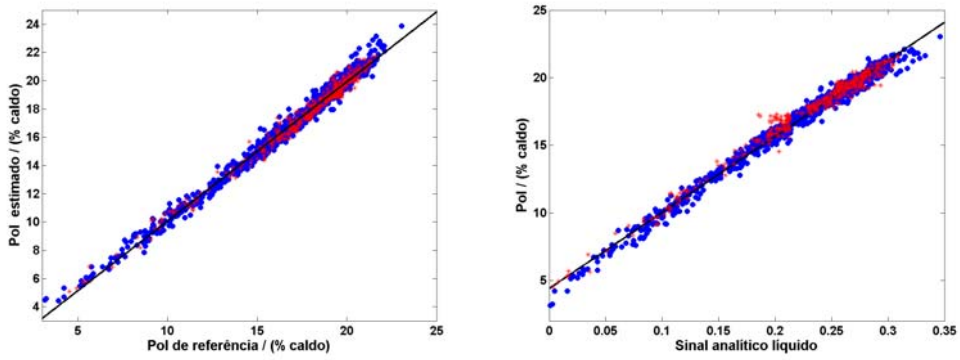


Figure 26. Ajuste para o parâmetro Pol do modelo construído com o espectro inteiro. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

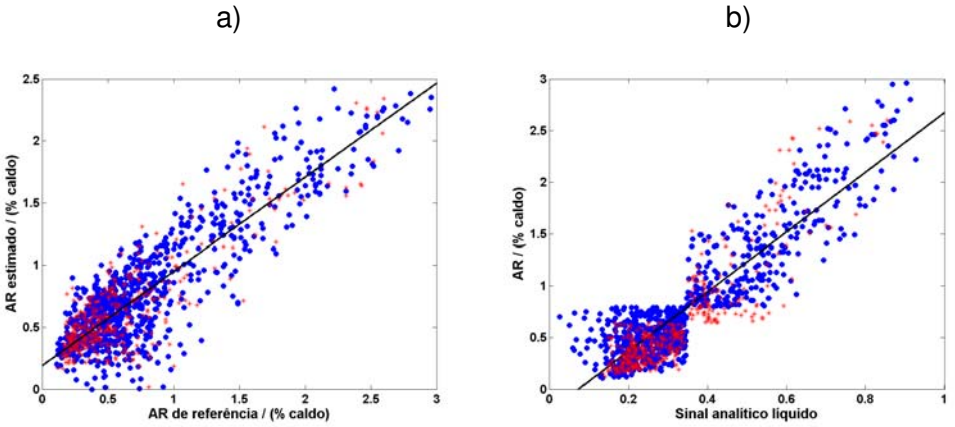


Figure 27. Ajuste para o parâmetro AR do modelo construído com o espectro inteiro. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

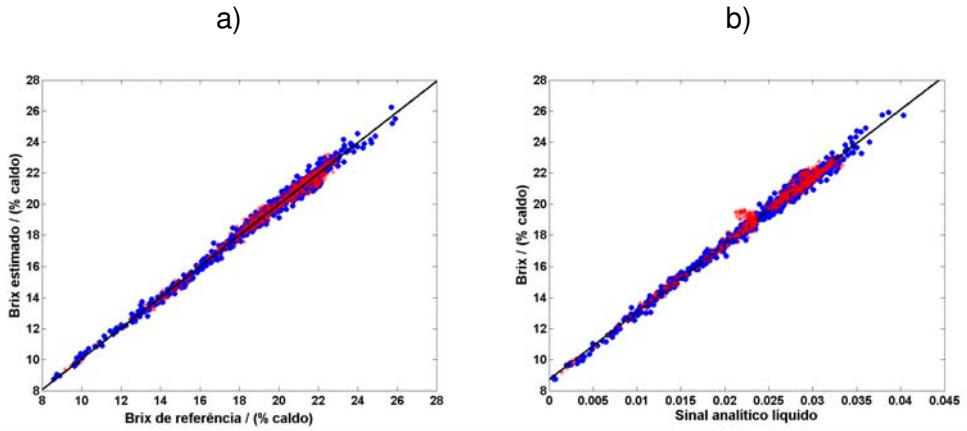


Figure 28. Ajuste para o parâmetro Brix do modelo iPLS. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

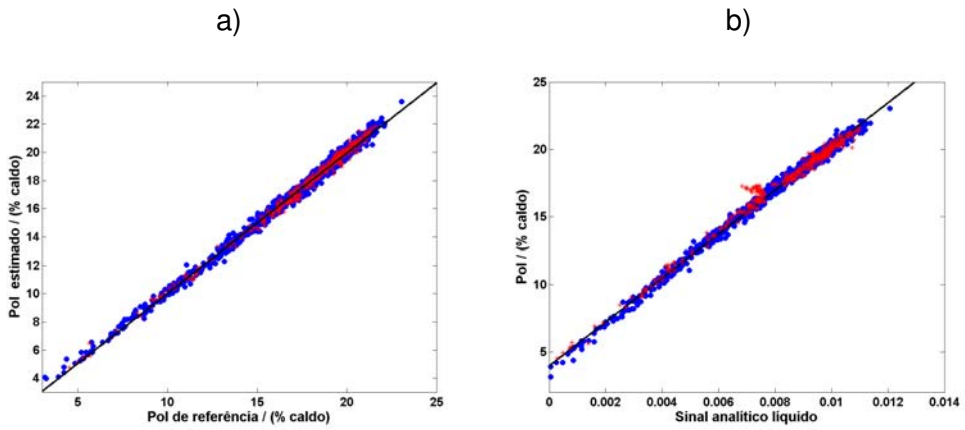


Figure 29. Ajuste para o parâmetro Pol do modelo iPLS. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

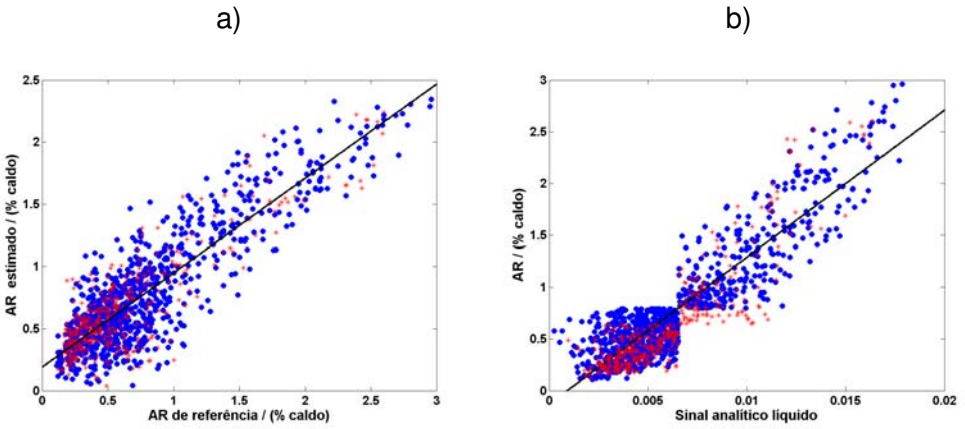


Figure 30. Ajuste para o parâmetro AR do modelo iPLS. (a) – Referência versus estimado e (b) – Ajuste NAS. Amostras de calibração (•) e validação (*).

Para os gráficos que representam o ajuste NAS observa-se uma lacuna que pode ser explicada devido aos modelos estarem centrados na média. Quando os modelos são centrados na média, metade dos resultados ficam abaixo e metade ficam acima da média. Assim, quando o vetor de sinal analítico líquido é representado por um escalar todos os resultados tornam-se positivos, uma vez que esta representação consiste no cálculo da norma euclidiana e os gráficos apresentam uma forma de V que é corrigida antes do cálculo dos coeficientes de regressão ($\hat{b}_{nas}$) de forma a evitar o erro de sinal que é introduzido pelo uso da norma euclideana. Esta correção é feita multiplicando-se os valores do escalar "nas" por –1. Quando o modelo apresenta um bom ajuste, esta correção funciona bem, como, por exemplo, para os parâmetros Brix e Pol. Por outro lado, para o parâmetro AR observa-se um ajuste inferior, uma falha nessa correção representa a lacuna observada no gráfico do ajuste NAS deste parâmetro.

As Figuras 31, 32 e 33 representam os resíduos da calibração e validação para as amostras dos parâmetros Brix, Pol e AR. Qualitativamente, estes gráficos podem indicar se os dados seguem ou não um comportamento linear. A distribuição aleatória desses resíduos é um indicativo de comportamento linear[30]. A distribuição dos erros para Brix e Pol apresentam um comportamento aleatório, no entanto, para o parâmetro AR observa-se uma certa tendência, que está de acordo com o valor de bias observado para o parâmetro e reforçam a suspeita de não-linearidade do modelo.

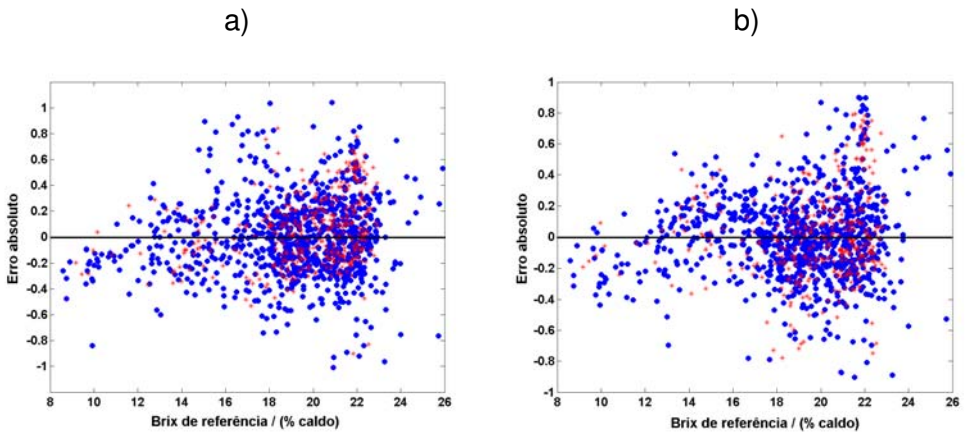


Figure 31. Resíduos do parâmetro Brix. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

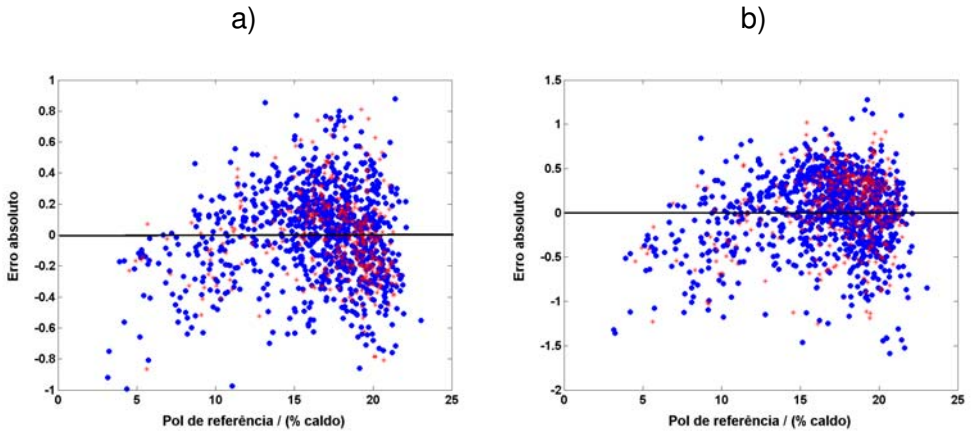


Figure 32. Resíduos do parâmetro Pol. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

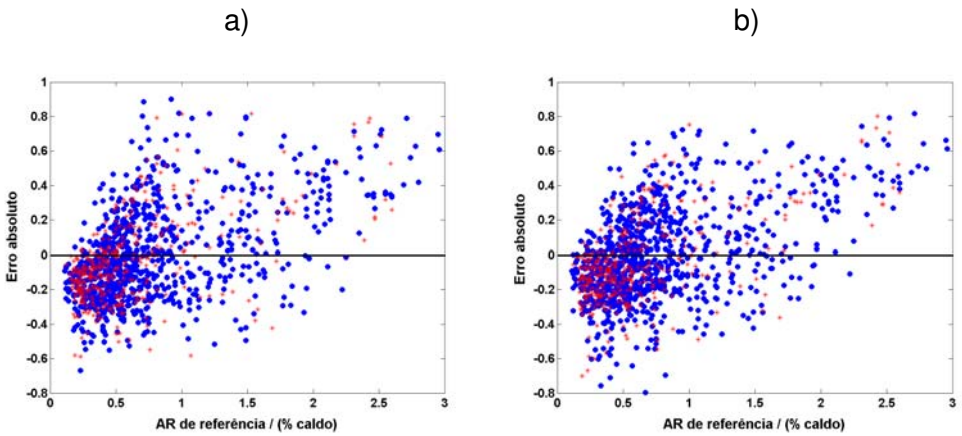


Figure 33. Resíduos do parâmetro AR. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

Os valores para a razão sinal/ruído apresentados na Tabela 4 mostram o quanto o escalar "nas" está acima do desvio padrão da flutuação do sinal instrumental da referência. Para os modelos construídos com o espectro inteiro estes valores são elevados e decrescem para os modelos construídos a partir das variáveis selecionadas pelo iPLS. Esta observação pode ser justificada pelos baixos valores de absorbância da região selecionada pelo iPLS.

Os limites de detecção e quantificação para os modelos construídos com todo o espectro mostraram resultados coerentes com as quantidades medidas para todos os parâmetros. Para os modelos iPLS esses limites apresentaram resultados inferiores em relação aos modelos obtidos com todas as variáveis do espectro, para todos os parâmetros. A justificativa para os resultados obtidos pode ser formulada pelo fato da sensibilidade ter valores inferiores para os modelos construídos a partir das variáveis selecionadas pelo iPLS. Entretanto, para Brix e Pol os valores para os limites de detecção e quantificação do modelo iPLS são condizentes com as quantidades determinadas para estes parâmetros, pois a faixa utilizada na construção dos modelos foi de aproximadamente 8,00 a 26,00 % caldo para o Brix e 3,00 a 24,00 % caldo para o Pol. Para o parâmetro AR os resultados obtidos mostram que o método NIR não consegue quantificar amostras com quantidades de açúcares redutores abaixo de 0,28 % caldo sendo que a faixa utilizada para a construção dos modelos foi aproximadamente 0,10 a 4,00 %

caldo. Dessa forma, a maioria das amostras utilizadas na determinação do AR está abaixo dos valores possíveis de se quantificar.

A Tabela 5 mostra os resultados para as percentagens de recobrimento dos intervalos de confiança estimados nos níveis de probabilidade de 99,0, 95,0 e 90,0%. Esta percentagem de recobrimento representa a percentagem das amostras que possuem o valor verdadeiro dentro dos limites de confiança estimados. Assim, por exemplo, no nível de 95% de confiança, no parâmetro Brix para os modelos construídos com o espectro inteiro, a percentagem de recobrimento de 96,9% indica que 96,9% das amostras apresentam os valores de referência dentro do intervalo calculado. Para todos os modelos e níveis de probabilidade estimados, os resultados mostraram que a equação para o cálculo da variância apresentado pela ASTM fornece intervalos de recobrimento próximos aos esperados teoricamente.

Tabela 5. Percentagem de recobrimento dos intervalos de confiança

Níveis de Probabilidade (%)	Brix (%) Espectro Inteiro/ Modelo iPLS	Pol (%) Espectro Inteiro/ Modelo iPLS	AR (%) Espectro Inteiro/ Modelo iPLS
99,0	99,2 / 96,4	99,2 / 98,6	98,4 / 98,6
95,0	96,9 / 90,5	95,8 / 96,0	95,6 / 95,1
90,0	90,6 / 86,9	93,3 / 93,5	92,7 / 91,9

Os limites médios estimados dos intervalos de confiança nos níveis de probabilidade de 99,0, 95,0 e 90,0% são apresentados na Tabela 6. Uma característica dos limites estimados obtidos usando as variâncias calculadas pela equação (74) sugerida pela ASTM, é que se obtém um limite para cada amostra em cada nível de probabilidade estimado. Isto porque o *leverage* é um termo que é específico para cada amostra o que influencia diretamente a estimativa dos limites de confiança. Para as amostras de calibração e amostras futuras, espera-se que o leverage não apresente diferença significativa. Assim, os limites para todas as amostras devem apresentar praticamente os mesmos valores. Esta situação é similar à calibração univariada, onde os limites de confiança são os

mesmos para todas as amostras. Uma hipótese para este resultado é que o erro instrumental representa uma pequena contribuição para o erro total. Neste sentido, por exemplo, no parâmetro Brix do modelo com o espectro inteiro, um valor de 20,0 %caldo significa que o resultado pode estar entre 19,4 e 20,6 %caldo.

Tabela 6. Limites médios dos intervalos de confiança estimados

Níveis de Probabilidade (%)	Brix (% caldo) Espectro Inteiro/ Modelo iPLS	Pol (% caldo) Espectro Inteiro/ Modelo iPLS	AR (% caldo) Espectro Inteiro/ Modelo iPLS
99,0	0,78/0,70	1,14/0,76	0,72/0,70
95,0	0,60/0,53	0,87/0,58	0,54/0,53
90,0	0,50/0,45	0,73/0,49	0,46/0,44

As Figuras 34, 35 e 36 mostram as barras de erro para a previsão de cada amostra dos parâmetros Brix, Pol e AR nos modelos construídos com o espectro inteiro e nos modelos iPLS, respectivamente. Estes gráficos ilustram a incerteza para cada parâmetro que, no caso do Brix e do Pol, mostra-se aceitável, sendo que, para o AR uma grande incerteza é observada. Entretanto, uma incerteza elevada também é observada para o AR estimado através da equação (4) em relação ao método de titulação. Isto pode ser comprovado através valor de RMSEP de 0,38 % caldo para previsões usando a referida equação.

a) b)

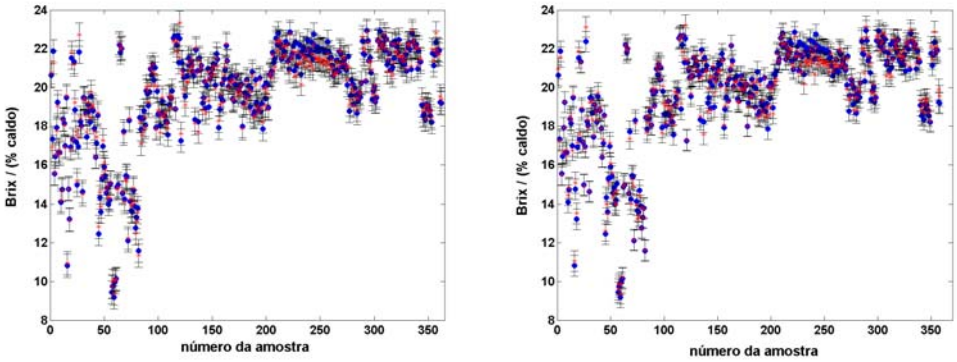


Figure 34. Barras de erro para o parâmetro Brix. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

a) b)

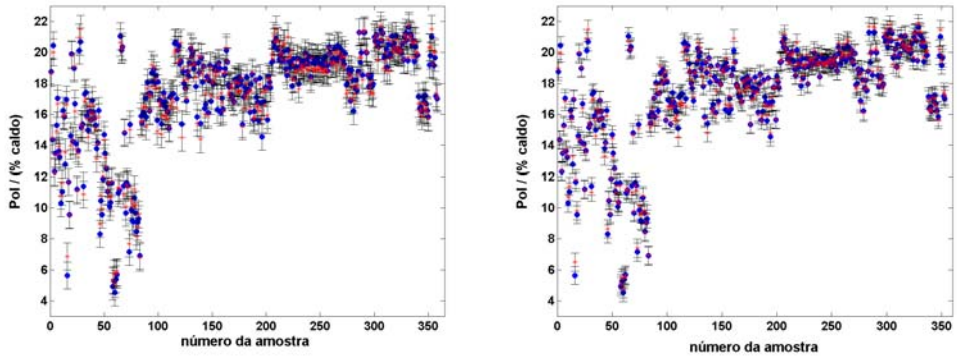


Figure 35. Barras de erro para o parâmetro Pol. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

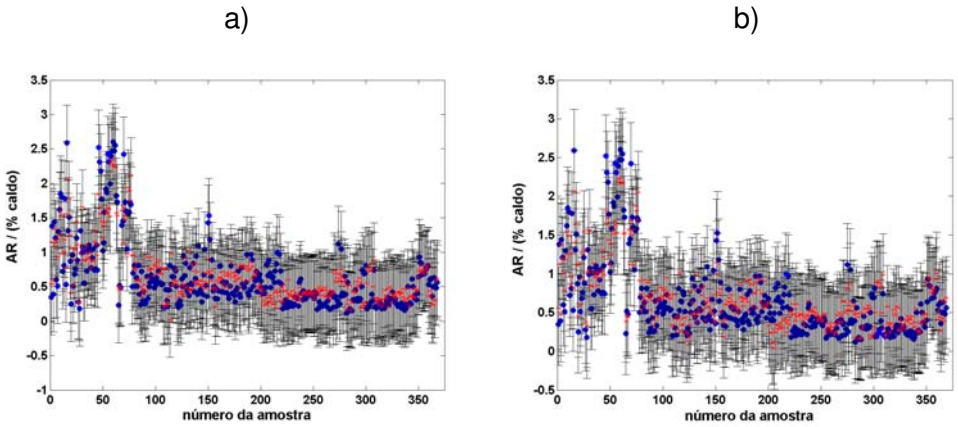


Figure 36. Barras de erro para o parâmetro AR. (a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS. Amostras de calibração (•) e validação (*).

Outro fato que confirma a consistência dos limites de confiança estabelecidos são os resíduos Studentizados das amostras de validação que são apresentados nas Figuras 37, 38 e 39, para os parâmetros Brix, Pol e AR, nos modelos construídos com todo o espectro e nos modelos iPLS, respectivamente. Estes resíduos seguem a distribuição-t com desvio padrão de 0,90, 0,93 e 0,93 para Brix, Pol e AR, respectivamente, nos modelos construídos com o espectro inteiro e 1,07, 0,90 e 0,94 para Brix, Pol e AR, respectivamente, nos modelos iPLS. Os valores alcançados de desvio padrão para todos os parâmetros são próximos ao valor teórico unitário.

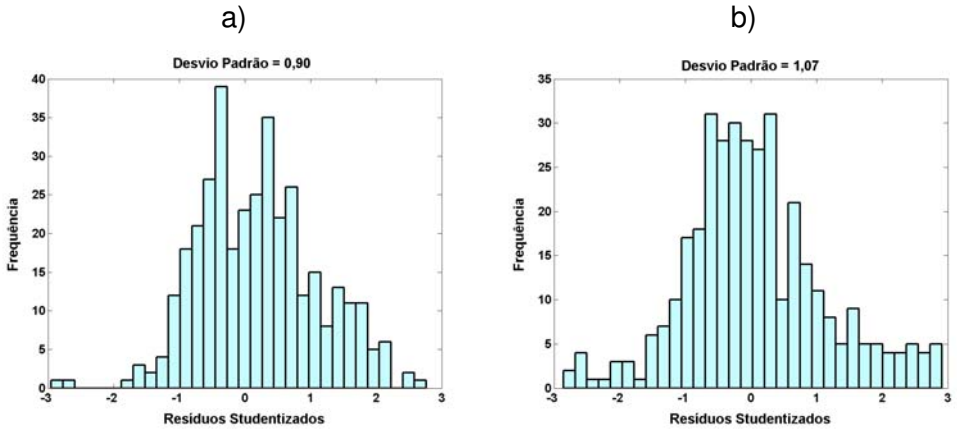


Figure 37. Resíduos Studentizados para o parâmetro Brix.

(a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS.

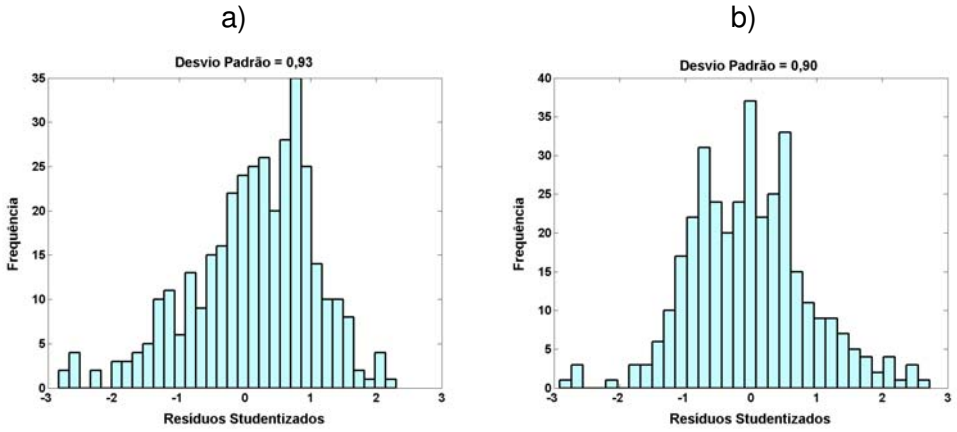


Figure 38. Resíduos Studentizados para o parâmetro Pol.

(a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS.

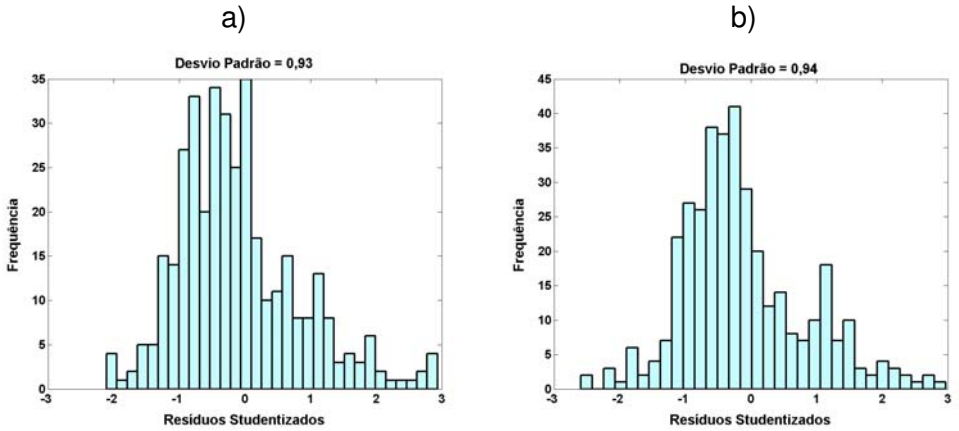


Figure 39. Resíduos Studentizados para o parâmetro AR.
(a) – Modelo construído com o espectro inteiro e (b) – Modelo iPLS.

O teste de robustez para os modelos desenvolvidos pode ser considerado satisfatório, uma vez que as amostras utilizadas para a validação dos referidos modelos foram consistentes com toda diversidade industrial como para canas colhidas em tempo seco, com chuva, a partir de colheita manual (cana queimada), a partir de colheita mecanizada, com horas de queima dentro do limite e fora deste, entre outros fatores. Assim, observou-se que os modelos foram capazes de realizar boas previsões, não se mostrando sensível às variações mencionadas.

A validação dos modelos desenvolvidos apresentaram resultados satisfatórios para o parâmetro Brix no modelo iPLS e para o parâmetro Pol no modelo com o espectro todo, pois nestes modelos o bias incluído não foi significativo. No caso do parâmetro AR, seria necessário um estudo de viabilidade levando em consideração a estimativa dos erros para o parâmetro obtido pela indústria através da equação (4) em relação ao método padrão de titulação. A Figura 40 ilustra a comparação entre os valores de AR obtidos pela titulação e os valores obtidos através da equação (4) que calcula o AR levando em consideração os valores de Brix e Pol. Por outro lado, a Figura 41 mostra os erros absolutos para o método de titulação em relação à estimativa do AR pela equação (4), de onde se pode verificar que os erros são superiores quando comparados aos do NIR em relação à titulação (Figura 33).

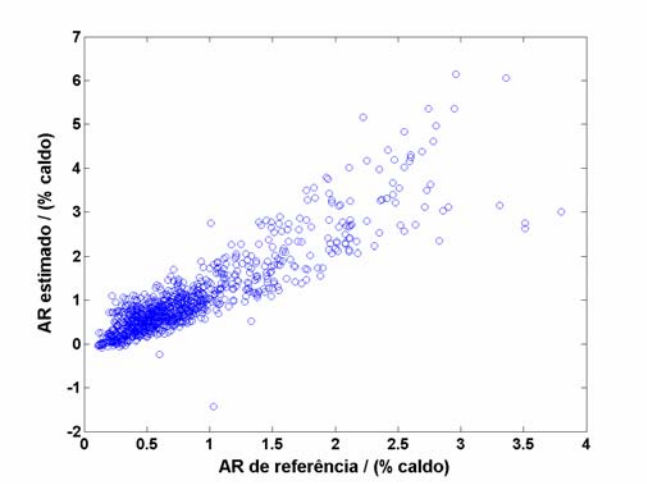


Figura 40. Comparação entre os valores de AR obtidos através do método de titulação e os valores estimados obtidos pela equação (4).

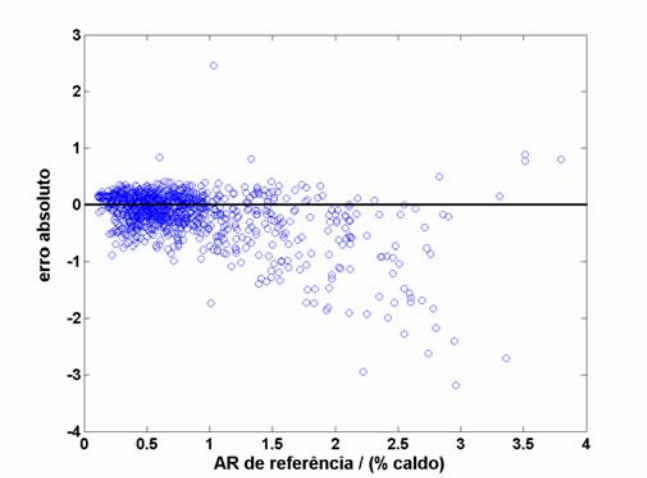


Figura 41. Resíduos dos valores de AR de referência obtidos através do método de titulação e os valores estimados obtidos pela equação (4).

CONCLUSÕES

Conclusões

De acordo com os resultados apresentados para a aplicação abordada, conclui-se que a validação de modelos de calibração multivariada construídos pelo método PLS1 pode ser realizada com base no cálculo das figuras de mérito, sendo estas condizentes com os resultados experimentais e demonstrando-se útil na comparação de modelos de calibração multivariada com métodos de referência, bem como, na caracterização da capacidade de previsão e performance dos referidos modelos frente às exigências de órgãos e normas de fiscalização. Algumas figuras de mérito não apresentaram maiores dificuldades na determinação, como por exemplo, a exatidão, precisão, robustez e o bias, sendo que, a determinação destas figuras de mérito para os modelos de calibração multivariada pode ser estimada de maneira bastante similar aos métodos de calibração univariada. Por outro lado, a linearidade, sensibilidade, razão sinal/ruído, ajuste, seletividade e intervalos de confiança não podem ser determinados, para os modelos de calibração multivariada, de forma similar à que é realizada para os modelos univariados. A seletividade e a razão sinal/ruído, para modelos de calibração multivariada, somente podem ser estimadas mediante o cálculo do sinal analítico líquido (NAS) para os parâmetros que estão sendo quantificados.

Os erros de previsão para Brix e Pol, foram muito inferiores aos apresentados como limites toleráveis por normas que regulamentam o controle de qualidade da cana-de-açúcar. Para o parâmetro AR não há registro acerca de qualquer regulamentação.

No parâmetro Brix a seleção de variáveis pelo iPLS produziu um modelo com melhor desempenho, enquanto que, para o parâmetro Pol melhores resultados foram obtidos nos modelos construídos a partir de todas as variáveis do espectro, considerando que, estes modelos apresentam bias não significativo.

A sensibilidade dos modelos iPLS foi inferior em relação aos modelos construídos com todo o espectro. Isto ocorreu porque as variáveis selecionadas apresentam baixos valores de absorbância, trazendo como conseqüência uma

redução nos limites de detecção e quantificação, uma vez que estes são calculados levando em consideração a sensibilidade. Entretanto, os valores obtidos para os limites de detecção e quantificação, embora inferiores nos modelos iPLS, são ainda condizentes com as quantidades determinadas para os parâmetros Brix e Pol.

Os limites de confiança estimados mostraram boa concordância com a probabilidade de recobrimento esperada em que a consistência dos resultados pôde ser atestada através dos resíduos studentizados.

Os modelos apresentaram elevada sensibilidade para os parâmetros estudados, sendo capazes de diferenciar amostras com pequenas diferenças de concentração e a capacidade de diferenciação entre amostras foi satisfatória para todos os modelos.

Os valores obtidos para exatidão, precisão e demais figuras de mérito mostraram resultados promissores, indicando que os modelos desenvolvidos por espectroscopia no infravermelho próximo para os parâmetros Brix e Pol podem ser utilizados pela indústria alcooleira como uma alternativa à refratometria (ou areometria) e medidas de polarização, metodologias padrão para determinação de Brix e Pol, respectivamente.

Para o parâmetro AR, pôde-se verificar que o método NIR apresenta menor erro absoluto do que a estimativa que atualmente é empregada pela indústria. Assim, a implementação do método NIR pela indústria fica restrita a aprovação do conselho dos produtores de cana-de-açúcar, açúcar e álcool que é o órgão que faz regulamentação do setor.

REFERÊNCIAS BIBLIOGRÁFICAS

Referências Bibliográficas

1. Fernandes, A. C. *Cálculos na agroindústria da cana-de-açúcar.* 2. ed. São Paulo: STAB, 2003. Site (URL): http://www.stab.org.br.

2. Payne, J. H. *Operações unitárias na produção de açúcar de cana.* São Paulo: Nobel/STAB, 1989.

3. Técnicos do Instituto Cubano de Pesquisa dos Derivados da cana-de-açúcar. *Manual dos derivados da cana-de-açúcar.* São Paulo: ABIPTI, 1999. Site (URL): http://www.abipti.org.br.

4. Conselho dos produtores de cana-de-açúcar, açúcar e álcool do estado do Paraná, CONSECANA-PR. *Normas operacionais de avaliação da qualidade da cana-de-açúcar.* 1. ed. Curitiba: FAEP, 2000. Site (URL): http://www.faep.com.br/consecana/normasop.htm.

5. Lane, H.; Eynon, L.; Determination of reducing sugar by means of Fehling's solution with methylene blue as internal indicator, *Journal of the Society of Chemistry Industry.* 1923, 42, 32T-37T.

6. Shreve, R. N.; Brink Jr, J. A. *Indústria de processos químicos.* Rio de Janeiro: Guanabara Dois, 1980.

7. Johnson, T. P. Cane juice analysis by near infrared (NIR) to determine grower payment, *International Sugar Journal.* 2000, 102, 1223, 603-609.

8. Cadet, F.; Bertrand, D.; Robert, P.; Maillot, J.; Dieudonné, J.; Rouch, C. Quantitative determination of sugar cane sucrose by multidimensional statistical analysis of their mid-infrared attenuated total reflectance spectra, *Applied Spectroscopy.* 1991, 45, 2, 166-172.

9. Cadet, F.; Robert, C.; Offmann, B. Simultaneous determination of sugar by multivariate analysis apllied to mid-infrared spectra of biological samples, *Applied Spectroscopy.* 1997, 51, 3, 369-375.

10. Baunsgaard, D.; Norgaard, L.; Godshall, M. A. Fluorescence of raw cane sugars evaluated by chemometrics, *Journal of Agricultural and Food Chemistry.* 2000, 48, 4955-4962.

11. Irudayaraj, J.; Xu, F.; Tewari, J. Rapid determination of invert cane sugar adulteration in honey using FTIR spectroscopy and multivariate analysis, *Journal of Food Science.* 2003, 68, 6, 2040-2045.

12. Sivakesava, S.; Irudayaraj, J. Prediction of inverted cane sugar adulteration of honey by Fourier transform infrared spectroscopy, *Journal of Food Science.* 2001, 66, 7, 972-978.

13. Tewari, J.; Mehrotra, R.; Irudayaraj, J. Direct near infrared analysis of sugar cane clear juice using a fibre-optic transmittance probe, *Journal of Near Infrared Spectroscopy.* 2003, 11, 351-356.

14. Filho, P. A. C.; Poppi, R. J. Use of near infrared spectroscopy for rapid estimation of sugar cane juice quality components, *Proceedings of the 9TH Internation Conference on Near Infrared Spectroscopy.* 1999, 897-902.

15. Salgo, A.; Nagy, J.; Mikó, É. Application of near infrared spectroscopy in the sugar industry, *Journal of Near Infrared Spectroscopy.* 1998, 6, A101-A106.

16. Burns, D. A.; Ciurczak, E. W. *Handbook of Near-Infrared Analysis.* New York: Marcel Dekker, 2001.

17. Workman Jr, J. J. Interpretative spectroscopy for near infrared, *Applied Spectroscopy Review.* 1996, 31, 3, 251-320.

18. Coates, J. A review of current new technology: Used in instrumentation for industrial vibrational spectroscopy, *Spectroscopy.* 1999, 14, 10, 21-34.

19. Williams, P.; Norris, K. *Near-Infrared technology in the agriculturam and food industries.* Minnesota: AACC, 1998.

20. Skoog, D. A.; Holler, F. J.; Nieman, T. A. *Princípios de análise instrumental.* 5. ed. São Paulo: Bookman, 2002.

21. Bokobza, L. Near infrared spectroscopy, *Journal of Near Infrared Spectroscopy.* 1998, 6, 3-17.

22. Wilks, P. A. IR filtometers for todays analytical requirements, *American Laboratory.* 1994, 26, 18, 42-45.

23. Frant, M. S.; LaButti, G. Process infrared measurements, *Analytical Chemistry.* 1980, 52, 12, A1331-A1344.

24. McClure, W. F. Near-infrared spectroscopy – the giant is running strong, *Analytical Chemistry.* 1994, 66, 1, A43-A53.

25. Noble, D. Illuminating near-IR spectroscopy, *Analytical Chemistry.* 1995, 67, 23, A735-A740.

26. Otto, M. *Chemometrics.* Weinheim: Wiley, 1999.

27. Brereton, R. G. Introduction to multivariate calibration in analytical chemistry, *Analyst.*2000, 125, 2125-2154.

28. Eurachem/Citac – Work group. Quality for research and development and non-routine analysis. 1. ed. 1998.

29. Helland, I. S. On the structure of partial least square regression, *Communications in statistics – simulation and computation.* 1998, 17, 2.

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

31. Sekulic, S.; Seasholtz, M. B.; Wang, Z.; Kowalski, B. R. Nonlinear multivariate calibration methods in analytical chemistry, *Analytical Chemistry*. 1993, 65, 19, A835-A845.

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

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

34. Chui, Q. S. H.; Zucchini, R. R.; Lichtig, J. Qualidade de medições em química analítica. Estudo de caso: determinação de cádmio por espectrofotometria de absorção atômica com chama, *Química Nova.* 2001, 24, 3, 374-380.

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

36. Beebe, K. R.; Kowalski, B. R. An Introduction to multivariate calibration and analysis, *Analytical Chemistry.* 1987, 59, 17, A1007-A1017.

37. Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; Jing, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B.* Amsterdan: Elsevier, 1998.

38. Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A practical guide.* Wiley, 1998.
39. Geladi, P.; Kowalski, B. R. Partial least square regression: a tutorial, *Analytica Chimica Acta.* 1986, 185, 1-17.
40. Oliveira, F. C.; Souza, A. T. P. C.; Dias, J. A.; Dias, S. C. L.; Rubim, J. C. A escolha da faixa espectral no uso combinado de métodos espectroscópicos e quimiométricos, *Química Nova.* 2004, 27, 2, 218-225.
41. Osborne, S. D.; Jordan, R. B.; Künnemeyer, R. Method of wavelength selection for partial least square, *Analyst.* 1997, 122, 1531-1537.
42. Costa Fillho, P. A.; Poppi, R. J. Algoritmo genético em química, *Química Nova.* 1999, 22, 3, 405-411.
43. Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L; Engelsen, S. B. Interval partial least-square regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, *Applied Spectroscopy.* 2000, 54, 3, 413-419.
44. Centner, V.; Massart, D. Elimination of uniformative variables for multivariate calibration, *Analytical Chemistry.* 1996, 68, 21, 3851-3858.
45. Filho, H. A. D.; Galvão, R. K. H.; Araújo, M. C. U.; Silva, E. C.; Saldanha, T. C. B.; José, G. E.; Pasquini, C.; Raimundo Jr.; I. M.; Rohwedder, J. J. R. A strategy for selecting calibration samples for multivariate modelling, *Chemometrics and Intelligent Laboratory Systems.* 2004, 72, 83-91.
46. Wang, Y. D.; Veltkamp, D. J.; Kowalski, B. R. Multivariate instrument standardization, *Analytical Chemistry.* 1991, 63, 23, 2750-2756.
47. Bouveresse, E.; Massart, D. L. Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration, *Chemometrics and Intelligent Laboratory Systems.* 1996, 32, 2, 201-213.
48. Kennard, R. W.; Stone, L. A. Computer aided desing of experiments, *Technometrics.* 1969, 11, 1, 137-148.

49. Honigs, D. E.; Hieftje, G. M.; Mark, H. L.; Hirschfeld, T. B. Unique-sample selection via near-infrared spectral subtraction, *Analytical Chemistry.* 1985, 57, 12, 2299-2303.

50. Puchwein, G. Selection of calibration samples for near-infrared spectrometry by factor analysis of spectra, *Analytical Chemistry.* 1988, 60, 569-573.

51. Navarro-Villoslada, F.; Pérez-Arribas, L. V.; Leon-González, M. E.; Pólo-Diez, L. M. Selection of calibration mixtures and wavelengths for different multivariate calibration methods, *Analytica Chimica Acta.* 1995, 313, 93-101.

52. Feudale, R. N.; Woody, N. A.; Tan, H.; Myles, A. J.; Brown, S. D.; Ferré, J. Transfer of multivariate calibration models: A review, *Chemometrics and Intelligent Laboratory Systems.* 2002, 64, 181-192.

53. Maesschalck, R. De; Jouan-Rimbaud, D.; Massart, D. L. The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems.* 2000, 50, 1-18.

54. Hoy, M.; Steen, K.; Martens, H. Review of partial least squares regression prediction error in Unscrambler, *Chemometrics and Intelligent Laboratory Systems.* 1998, 44, 123-133.

55. Fernández Pierna, J. A.; Wahl, F.; Noord, O. E.; Massart, D. L. Methods for outlier detection in prediction, *Chemometrics and Intelligent Laboratory Systems.* 2002, 63, 27-39.

56. Fernández Pierna, J. A.; Jin, L.; Daszykowski, M.; Wahl, F.; Massart, D. L. A methodology to detect outliers/inliers in prediction with PLS, *Chemometrics and Intelligent Laboratory Systems.* 2003, 68, 17-28.

57. Jouan-Rimbaud, D.; Bouveresse, E.; Massart,D. L.; Noord, O. E. Detection of prediction outliers and inliers in multivariate calibration, *Analytica Chimica Acta.*1999, 388, 3, 283-301.

58. Egan, W. J.; Morgan, S. L. Outlier detection in multivariate calibration chemical data, *Analytical Chemistry.* 1998, 70, 11, 2372-2379.

59. Pell, R. J. Multiple outlier detection for multivariate calibration using robust staticak techniques, *Intelligent and Laboratory Systems.* 2000, 52, 87-104.

60. Ilie, M. Comparison of different modalities of outlier treatment for qualitative near infrared spectra, *Journal of Near Infrared Spectroscopy.* 1998, 6, A175-A179.

61. Annual Book of ASTM Standards. *Standards practices for infrared, multivariate, quantitative analysis,* E1655, vol 03.06. ASTM International, West Conshohocken, Pennsylvania, USA, 2000.
62. Eurachem/Citac – Work Group. *Guide of quality in analytical chemistry – An aid to accreditation.* 2. ed. 2002.
63. Eurachem/Citac – Work Group. *Quantifying Uncertainty in Analytical Measurement.* 2. ed. 2000.
64. Danzer, K; Otto, M. Currie, L. Guidelines for calibration in analytical chemistry. Part 2. Multispecies calibration (IUPAC Technical Report). *Pure Applied Chemistry.* 2004, 76, 6, 1215-1225.
65. Braga, J. W. B.; Poppi, R. J. Figures of merit for determination of the polymorphic purity of carbamazepine by infrared spectroscopy and multivariate calibration, *Journal of Pharmaceutical Science.* 2004, 93, 8, 2124-2134.
66. Braga, J. W. B.; Poppi, R. J. Validação de modelos de calibração multivariada: uma aplicação na determinação de pureza polimórfica de carbamazepina por espectroscopia no infravermelho próximo, *Química Nova.* 2004, 27, 6, 1004-1011.
67. Currie, L. A. Nomenclature in evaluation of analytical methods including detection and quantification capabilities (IUPAC Recommendations 1995), *Analytica Chimica Acta.*1999, 391, 105-126.
68. Neto, B. B.; Pimentel, M. F.; Araújo, M. C. U. Recomendações para calibração em química analítica – Parte 1. Fundamentos e calibração com um componente (Calibração Univariada), *Química Nova.* 2002, 25, 5, 856-865.
69. Boqué, R.; Larrechi, M. S.; Rius, F. X. Multivariate detection limits with fixed probabilities of error, *Chemometrics and Intelligent Laboratory Systems.* 1999, 45, 397-408.
70. Faber, N. M.; Song, X. H.; Hopke, P. K. Sample-specific standard error of prediction for partial least squares regression, *Trends in Analytical Chemistry.* 2003, 22, 5, 330-334.
71. Espinosa-Mansilla, A.; Merás, I. D.; Gómez, M. J. R.; Muñoz de la Pena, A.; Salinas, F. Selection of the wavelength range and spectrophotometric

determination of leucovorin and methotrexate in human serum by na net analyte signal based methos, *Talanta.* 2002, 58, 255-263.

72. Faber, N. M.; Ferre, J.; Boqué, R.; Kalivas, J. H. Quantifying selectivity in spectrophotometric multicomponent analysis, *Trends in Analytical Chemistry.* 2003, 22, 6, 352-361.

73. Trafford, A. D.; Jee, R. D.; Moffat, A. C.; Graham, P. A rapid quantitative assay of intact paracetamol tablets by reflectance near-infrared spectroscopy, *Analyst.* 1999, 124, 163-167.

74. Bergmann, G.; Oepen, B. V.; Zinn, P. Improvement in the definitions of sensitivity and selectivity, *Analytical Chemistry.* 1987, 59, 2522-2526.

75. Faber, N. M. Notes on two competing definitions of multivariate sensitivity, *Analytica Chimica Acta.* 1999, 381, 103-109.

76. Lorber, A. Error propagation and figures of merit for quantification by solving matrix equations, *Analytical Chemistry.* 1986, 58, 6, 1167-1172.

77. Lorber, A.; Faber, K.; Kowalski, B. R. Net analyte signal calculation in multivariate calibration, *Analytical Chemistry.* 1997, 69, 8, 1620-1626.

78. Faber, N. M. Efficient computation of net analyte signal vector in inverse multivariate calibration models, *Analytical Chemistry.* 1998, 70, 23, 5108-5110.

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

80. Goicoechea, H. C.; Olivieri, A. C. A comparation of orthogonal signal correction and net analyte signal preprocessing methods. Theoretical and experimental study, *Chemometrics and Intelligent Laboratory Systems.* 1998, 56, 73-81.

81. Faber, N. M. Exact presentation of multivariate calibration model as univariate calibration graph, *Chemometrics and Intelligent Laboratory Systems.* 2000, 50, 107-114.

82. Faber, N. M. Mean centering and computation of scalar net analyte signal in multivariate calibration, *Journal of Chemometrics.* 1998, 12, 405-409.

83. Muñoz de la Pena, A.; Espinosa-Mansilla, A.; Acedo Valenzuela, M. I.; Goicoechea, H. C.; Olivieri, A. C. Comparative study of net analyte signal-based

methods and partial least squares for the simultaneous determination of amoxycillin and clavulanic acid by stopped-flow kinetic analysis, *Analytica Chimica Acta.* 2002, 463, 75-88.

84. Rodríguez, L. C.; Campanã, A. M. G.; Linares, C. J.; Ceba, M. R. Estimation of performance characteristics of an analytical method using the data set of the calibration experiment, *Analytical Letters.* 1993, 26, 6, 1243-1258.

85. Vessman, J.; Stefan, R. I.; Van Staden, J. F.; Danzer, K.; Lindner, W.; Burns, D. T.; Fajgelj, A.; Muller, H. Selectivity in analytical chemistry (IUPAC Recommendations 2001), *Pure an Applied Achemistry.* 2001, 73, 8, 1381-1386.

86. Boqué, R.; Rius, F. X. Multivariate detection limits estimators, *Chemometrics and Intelligent Laboratory Systems.* 1996, 32, 11-23.

87. Faber, N. M.; Bro, R. Standard error of prediction for multway PLS1. Background and a simulation study, *Chemometrics and Intelligent Laboratory Systems.* 2002, 61, 133-149.

88. Van der Voet, H. Pseudo-degrees of freedom for complex predictive models: The example of partial least squares, *Journal of Chemometrics.* 1999, 13, 195-208.

89. Morsing, T.; Ekman, C. Comments on construction of confidence intervals in connection with partial least squares, *Journal of Chemometrics.* 1998, 12, 295-299.

90. Faber, N. M.; Response to 'Coments on construction of confidence intervals in connection with partial least squares', *Journal of Chemometrics.* 2000, 14, 363-369.