RAPHAEL HENRIQUE TEIXEIRA DA SILVA

REGRESSÃO MULTIVARIADA PARA DETERMINAÇÃO DE SACAROSE NA PRESENÇA DE CACAU USANDO DIFERENTES INSTRUMENTOS DE ESPECTROSCOPIA NIR

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

VIÇOSA
MINAS GERAIS – BRASIL
2019

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

T
S586r
2019

Silva, Raphael Henrique Teixeira da, 1991-
Regressão multivariada para determinação de sacarose na presença de cacau usando diferentes instrumentos de espectroscopia NIR / Raphael Henrique Teixeira da Silva. – Viçosa, MG, 2019.
ix, 31f : il. (algumas color.) ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.29-31.

1. Análise multivariada. 2. Predição. 3. Espectroscopia de infravermelho . I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.535

RAPHAEL HENRIQUE TEIXEIRA DA SILVA

REGRESSÃO MULTIVARIADA PARA DETERMINAÇÃO DE SACAROSE NA PRESENÇA DE CACAU USANDO DIFERENTES INSTRUMENTOS DE ESPECTROSCOPIA NIR

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

APROVADA: 19 de fevereiro de 2019.

Reinaldo Francisco Teofilo

Bruno Ricardo de Castro L. Júnior

Luiz Alexandre Peternelli
(Orientador)

Aos meus pais,
pelo sacrifício para que eu chegasse até aqui e
à minha noiva, Sara,
por todo amor incondicional e amparo nos momentos difíceis,
DEDICO

AGRADECIMENTOS

Primeiramente a Deus, por sempre iluminar meus caminhos e me conceder sabedoria.

Aos meus pais, pelas orações e por todo carinho e amor.

À minha noiva Sara, que teve uma participação muito especial neste trabalho, dando-me força nos momentos mais difíceis, contribuindo com o seu tempo para me auxiliar e por ter acreditado em mim. O seu amor e dedicação comigo foram essenciais para tornar esta etapa mais leve.

Ao professor Luiz Alexandre Peternelli, pela orientação competente, pela amizade, pelos ensinamentos teóricos, pela descontração, pela ajuda e compreensão que muito contribuíram para a elaboração desse trabalho.

Aos amigos conquistados em Viçosa, em especial, a Patrícia e o Bruno, pelos momentos de diversão e por serem grandes parceiros. A Jussara, por toda ajuda para a execução deste trabalho.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade concedida.

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

BIOGRAFIA

Raphael Henrique Teixeira da Silva, filho de Cristiane Magda Teixeira da Silva e Giovanni Bráz da Silva, nasceu em Ponte Nova, Minas Gerais, em 04 de novembro de 1991.

Em dezembro de 2014, graduou-se em Engenharia de Produção pela Universidade Federal de São João Del Rei, São João Del Rei - MG.

Em fevereiro de 2017 ingressou no curso de Mestrado em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

Em agosto de 2017, especializou-se em Engenharia de Segurança do Trabalho pela Faculdade de Ciências e Tecnologia de Viçosa, Viçosa - MG.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

SILVA, Raphael Henrique Teixeira, M.Sc., Universidade Federal de Viçosa, fevereiro de 2019. **Regressão multivariada para determinação de sacarose na presença de cacau usando diferentes instrumentos de espectroscopia NIR.** Orientador: Luiz Alexandre Peternelli.

O objetivo deste trabalho foi realizar um estudo comparativo entre os dois tipos de instrumentos da técnica NIR (o NIR de bancada, com maior resolução (1000nm até 2500nm), e o portátil, com menor resolução (900nm até 1700nm)) e averiguar se o NIR portátil é um substituto ao NIR de bancada. A fim de elucidar a viabilidade, ou não, da utilização do NIR portátil, foram realizados experimentos de mistura entre cacau e sacarose (Experimento 1) e cacau, sacarose e frutose (Experimento 2). Para ambos os experimentos, observou-se que a diferença dos erros relativos entre os instrumentos (NIR portátil e de bancada) não foram tão expressivas, sendo, em média, 4% de diferença entre as amostras do NIR portátil e do bancada para o Experimento 1, e 6% para o Experimento 2. É importante dizer que quando se utiliza a mesma faixa espectral coincidente em ambos os instrumentos (1000nm até 1700nm), pode-se concluir que o NIR portátil é recomendado para estudos de mistura de cacau e sacarose, e para os que envolvem frutose na mistura. Considerando os resultados para as amostras dos produtos industrializados, observou-se que os modelos dos Experimentos 1 e 2, utilizando o NIR portátil, foram capazes de predizer, de forma significativa, os percentuais de sacarose correspondente nas embalagens dos produtos de interesse. Desta forma, o instrumento portátil apresentou ser uma boa alternativa para realizar as análises para predição de sacarose, considerando o custo-benefício, podendo-se reduzir custos com aquisição de instrumento e proporcionar rapidez e maior mobilidade para análises.

ABSTRACT

SILVA, Raphael Henrique Teixeira, M.Sc., Universidade Federal de Viçosa, February, 2019**. Multivariate regression for the determination of sucrose in the presence of cocoa using different NIR spectroscopy instruments.** Adviser: Luiz Alexandre Peternelli.

The objective of this work was to perform a comparative study between the two types of instruments of the NIR technique (the bench NIR, with higher resolution (1000nm to 2500nm), and the portable, with lower resolution (900nm to 1700nm)) and Portable NIR is a replacement for benchtop NIR. In order to elucidate the feasibility or otherwise of the use of portable NIR, experiments were performed between cocoa and sucrose (Experiment 1) and cocoa, sucrose and fructose (Experiment 2). For both experiments, it was observed that the difference of the relative errors between the instruments (portable and bench NIR) were not as expressive, being, on average, 4% difference between the portable and bench NIR samples for the Experiment 1, and 6% for Experiment 2. It is important to say that when using the same coincident spectral range in both instruments (1000nm to 1700nm), it can be concluded that portable NIR is recommended for studies of cocoa and sucrose , and for those involving fructose in the blend. Considering the results for the samples of the industrialized products, it was observed that the models of Experiments 1 and 2, using portable NIR, were able to predict, in a significant way, the corresponding percentages of sucrose in the packages of the products of interest. In this way, the portable instrument presented a good alternative to carry out the analysis for sucrose prediction, considering the cost-benefit, being able to reduce costs with instrument acquisition and to provide speed and greater mobility for analysis.

1. INTRODUÇÃO

1.1. Considerações iniciais

A quimiometria é a subárea da química que utiliza ferramentas matemáticas e estatísticas com a finalidade de retirar informações elementares de dados analíticos. Uma das técnicas comumente utilizada é a espectroscopia na região do infravermelho próximo (NIR). A espectroscopia NIR abrange modelos de identificação e quantificação mais robustos e estende sua aplicabilidade, enquanto apresenta novos desafios à quimiometria que motivam o aprimoramento de muitas de suas técnicas (PASQUINI, 2003).

Nesse sentido, a espectroscopia NIR é empregada em diversas áreas do conhecimento, devido a diversos benefícios quando comparada aos métodos convencionais de análises. Dentre estes benefícios podem-se destacar: menor tempo necessário para preparação de amostras e execução, alta precisão, não utilização de reagentes e solventes químicos tóxicos, além de não provocar descarte de resíduos. Desta forma é considerada como um método ecologicamente correto, que está de acordo com os princípios da química verde (SILVA et al., 2017; CASCANT et al., 2017; CLAVAUD et al., 2016; YANG et al., 2017; OLAREWAJU et al., 2016). Estopa et al. (2017) e Nascimento et al. (2017), afirmam que a técnica NIR é promissora devido à necessidade de análises precoces e não destrutivas, em especial nas áreas de pesquisa do setor florestal. Além disso, tem sido amplamente aplicada para medir a qualidade de produtos alimentares, pois é não invasiva e pode fornecer rapidamente as informações físicas e químicas sobre as amostras (BAYE et al., 2006). Diante deste contexto, é notória a importância da realização de estudos que envolvem a técnica NIR.

Existem dois tipos de instrumentos para a técnica: o espectrômetro NIR de bancada, em que as análises são realizadas considerando a faixa de comprimento de onda (λ) de 1000 a 2.500 nm; e o portátil, que possui, em geral, λ = 900 a 1700 nm. A diferença entre ambos os instrumentos está relacionada à acurácia dos resultados obtidos, sendo o NIR de bancada considerado, em tese, o que promoverá maior poder de predição. Em contra partida, o NIR de bancada é um instrumento mais caro e, além disso, restrito a receber as amostras no laboratório onde se encontra (PASQUINI, 2003).

Por conseguinte, é importante a realização de trabalhos abordando estudos comparativos entre os instrumentos NIR de bancada e portátil, a fim de avaliar o desempenho dos dois instrumentos, e de verificar se o NIR portátil possui performance melhor ou igual ao NIR de bancada. Em caso afirmativo, pode-se indicar a substituição do NIR de bancada pelo portátil, visto que há uma possível redução do custo com aquisição do instrumento e maior praticidade para a realização das análises.

1.2. Justificativa e relevância

O objeto de estudo foram dois experimentos de mistura entre cacau, sacarose e frutose, que são similares aos achocolatados industriais. A escolha se justifica, pois, segundo Eduardo e Lannes (2004), os achocolatados são alimentos consumidos por diversas pessoas de todas as idades e de fácil acesso em todo o mundo. Além disso, este produto é bem aceito no mercado por causa de suas características sensoriais e nutricionais. São bem convenientes e práticos na rotina dos consumidores, o que acarreta em alta aceitação e consumo. É importante destacar que na sua apresentação mais simples, o achocolatado contém cerca de 70% de sacarose, misturado ou não a outros açúcares (como a frutose) e cerca de 30% de cacau em pó (EDUARDO e LANNES, 2004).

A Cromatografia Líquida de Alta Eficiência (HPLC) é um dos métodos convencionais de análise que pode ser utilizado para a determinação de sacarose em achocolatados. Porém, é um método que demanda muito tempo e é caro, o que se torna inviável para as indústrias, visto que é necessário realizar diversas análises em diferentes amostras (COLLINS, 1997).

Portanto, a espectroscopia NIR, associada a técnicas de calibração multivariada, surgiu como ferramenta para predição da sacarose nos experimentos propostos como possibilidade de aplicação prática no setor industrial ou de inspeção de qualidade e fraude dos produtos. É importante mencionar que Permanyer & Perez (1989) apresentaram um estudo em que a técnica de espectroscopia NIR poderia ser usada no controle de qualidade de produtos de cacau em pó, para umidade, gordura e sacarose, que é o foco deste trabalho.

1.3. Objetivos

1.3.1. Objetivo geral

O objetivo deste trabalho foi construir e validar modelos de calibração multivariada para análise de dois experimentos de misturas: (1) cacau e sacarose; (2) cacau, sacarose e frutose, usando espectrômetros portátil e de bancada.

1.3.2. Objetivos específicos

- Comparar os modelos de predição de sacarose nessas misturas por meio de recursos estatísticos.
- Comparar os desempenhos dos espectrômetros NIR portátil e de bancada.

1.4. Organização do texto

Este trabalho está estruturado em cinco seções. A primeira contém a introdução. A segunda possui a revisão de literatura, em que consta o embasamento teórico das ferramentas estatísticas utilizadas. A terceira seção contém os métodos. A quarta seção apresenta os resultados e as discussões. Por último, a quinta seção possui as conclusões.

2. REVISÃO BIBLIOGRÁFICA

2.1 Espectroscopia no infravermelho próximo (NIR)

O termo espectroscopia consiste no estudo da radiação eletromagnética emitida ou absorvida por um corpo. A espectroscopia no infravermelho está entre uma das técnicas mais importantes para a identificação e elucidação de compostos orgânicos e na obtenção de variáveis qualitativas. A região do espectro que corresponde ao infravermelho (Figura 1) está localizada depois da região do visível e está dividida em: infravermelho próximo – NIR (*Near Infrared*), infravermelho médio – MIR (*Middle Infrared*) e infravermelho distante – FIR (*Far Infrared*) (PASQUINI, 2003).

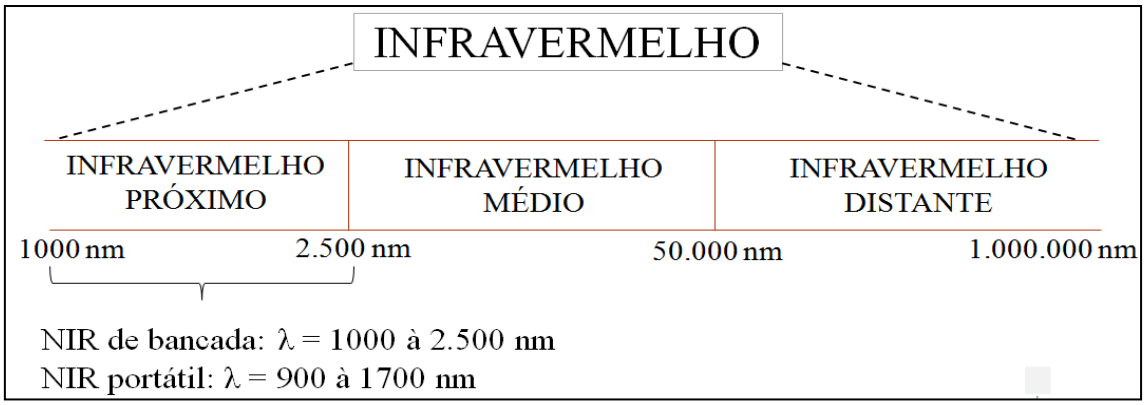


Figura 1– Esquema resumido sobre os tipos de infravermelho. O comprimento de onda apresentado corresponde aos instrumentos usados nessa pesquisa. FONTE: Adaptado de Pasquini (2003).

A espectroscopia no NIR, que é o foco deste trabalho, está presente em diversos setores, como por exemplo: nas áreas agrícola, alimentícia, médica, têxtil, de cosméticos, de polímeros, de tintas, ambiental, petroquímica e farmacêutica. A técnica tem como fundamento a espectroscopia vibracional e se resume na interação da radiação eletromagnética com a matéria, para se obter, assim, os espectros. A partir desses dados, pode-se gerar um modelo estatístico capaz de explicar a maioria das informações contidas no espectro (WILLIAMS e NORRIS, 2001).

A espectroscopia no NIR apresenta vantagens, das quais pode-se citar: (*i*) é rápido; (*ii*) não é destrutiva; (*iii*) é limpa, pois pode ser evitado o uso de solventes tóxicos e descarte de resíduos; (*iv*) exige pouco ou nenhum preparo de amostras; (*v*) possui parte instrumental simples, dentre outras.

Os dados provenientes do NIR são organizados em uma matriz $\mathbf{X}$ ($N \times p$) em que cada linha se refere a uma amostra i (i = 1 a N), ou espectro, e cada coluna j (j = 1 a p) aos comprimentos de ondas, ou variáveis. Apenas uma única amostra origina muitas variáveis, de modo que, em geral, o número de comprimentos de onda apurados no NIR é muito superior ao de amostras coletadas ($p >> N$). É importante destacar que para a execução deste trabalho, foi utilizada apenas a espectroscopia no infravermelho próximo (Figura 1), considerando dois instrumentos de medição: o de bancada e o portátil.

Uma vez que existe variabilidade (ruídos experimentais) em qualquer conjunto de dados (WOLD, 1995), o que prejudica a interpretação dos resultados, é necessário realizar tratamentos nos dados. As principais causas da ocorrência de ruído são: falta de completo controle das condições experimentais, a instabilidade do instrumento de medida e erros de modelo. Além disso, algumas variáveis podem representar a mesma

informação por existir uma forte relação entre elas (multicolinearidade). Estas situações são características em dados provenientes da técnica NIR. Desta forma, foram aplicados os pré-tratamentos, que visam minimizar os erros sistemáticos presentes nos dados, tornando a matriz de dados mais adequada para análise (FERREIRA, 2015).

2.2. Quimiometria

A quimiometria faz uso de recursos computacionais e das áreas da matemática e da estatística para aperfeiçoar procedimentos experimentais e de análise na área da química. O principal objetivo é retirar informações cruciais, por meio de análise de dados, e adquirir alto conhecimento sobre o sistema em estudo. Ou seja, é possível obter a informação química de um conjunto complexo de informações (FERREIRA, 2015).

Devido à utilização de recursos e ferramentas que permitem manusear e arquivar um grande volume de dados torna-se fundamental realizar tratamentos matemáticos e estatísticos mais complexos, com o intuito de alcançar informações mais completas e exatas. Nesse sentido, a quimiometria consiste em analisar conjuntos de dados multivariados nas variáveis explicativas de origem química com métodos estatísticos (FERREIRA, 2015).

2.2.1. Calibração multivariada

A calibração multivariada pode ser definida como o processo de construção de um modelo matemático que relaciona variáveis independentes a variáveis dependentes. O objetivo é relacionar respostas instrumentais à(s) propriedade(s) conhecida(s) de padrões. O modelo construído permite realizar predições das propriedades desconhecidas a partir de respostas obtidas de amostras (ROQUE, 2015).

A Figura 2 ilustra o processo de calibração resumido em três etapas básicas. Na primeira etapa, referente à construção do modelo, foi utilizado o método de Regressão por Quadrados Mínimos Parciais (PLS). A finalidade é reduzir o conjunto de dados original sem perder muita informação do sistema completo de variáveis e desta forma, maximizar o desempenho com o mínimo esforço de processamento (FERREIRA, 2002). A segunda etapa, referente à validação, que teve como objetivo separar as amostras em dois conjuntos, um de calibração (ou treinamento) e em outro de predição (ou teste), pré-definido no planejamento experimental. A terceira etapa consistiu em validação externa. Foi verificada a eficiência do modelo de calibração construído e sua capacidade preditiva.

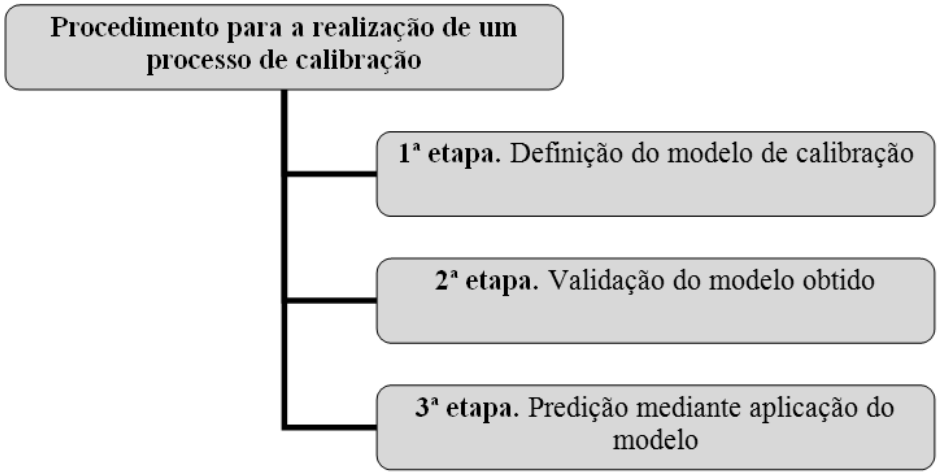


Figura 2– Realização de um processo de calibração, FONTE: Adaptado de Pasquini (2003).

Segundo Ferreira (2015), após a organização dos dados matricialmente, eles devem ser pré-tratados antes da análise quimiométrica. Este é um procedimento importante e vários métodos são testados para garantir que o pré-tratamento mais adequado seja empregado. O objetivo é reduzir as variações indesejadas que não foram retiradas durante a aquisição dos dados e que não são eliminadas naturalmente durante a análise, mas que podem impactar de forma negativa nos resultados finais.

Diante do exposto, foi necessário realizar os pré-tratamentos na matriz de dados antes da construção do modelo estatístico, pois estes minimizam os erros sistemáticos presentes nos dados, tornando a matriz de dados mais adequada para análise. Há dois tipos de pré-tratamentos: um deles é aplicado às linhas da matriz **X** (transformação) e o outro, às colunas de **X** (pré-processamento) (FERREIRA, 2015).

Foram aplicados à matriz de espectros diferentes pré-tratamentos (centragem na média, alisamento, derivação e correção multiplicativa de sinal (MSC)) (FERREIRA, 2015). Após, foi verificado quais foram os mais adequados para o conjunto de dados em estudo por meio de estatísticas que inferem sobre erros de predição. Mais detalhes e esclarecimentos sobre pré-tratamentos podem ser obtidos em Ferreira (2015).

2.2.2. Regressão Múltipla por Quadrados Mínimos Parciais (PLS)

Nesse trabalho foi adotado o método de regressão múltipla para a realização da calibração multivariada, que relacionou a resposta de interesse obtida em cada amostra com os espectros de infravermelho próximo obtidos nos dois instrumentos: o portátil e o de bancada. Em outras palavras, foi obtido um modelo que correlacionou a matriz **X**

(matriz de variáveis explicativas) com o vetor **y** (variável resposta). Para que o modelo seja promissor, é necessário estabilidade entre o modelo a ser escolhido e sua capacidade preditiva (FERREIRA, 2015). Afinal, caso seja obtido um modelo que não descreve de forma satisfatória o comportamento dos dados, ou um com sobre ajuste aos dados, isto é, inclusão de excesso de informação no modelo, que pode ser aleatória ou estar relacionada à presença de erros sistemáticos, pode acarretar em perda da sua capacidade preditiva.

O exemplo de um modelo de regressão linear múltipla está representado pela equação (1).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1,2,\cdots,N \tag{1}$$

em que, x_{ij} é o valor da variável independente x_j na *i-ésima* observação com *j=1,...,p*; β_k, $k = 0,1,2...,p$, são os coeficientes da regressão (parâmetros), y_i é a variável dependente na *i-ésima* observação e ε_i é erro na *i-ésima* observação. Ou também, o modelo pode ser representado conforme em (2).

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim N(\mathbf{0}; \sigma^2 \boldsymbol{I}) \tag{2}$$

em que, **y** é um vetor *N* x 1 em que os elementos correspondem às *N* observações; **X** é uma matriz de dimensão $N \times (p+1)$, denominada matriz de incidência; **β** é um vetor $(p+1) \times 1$ em que os componentes são os coeficientes de regressão, e **Ɛ** é um vetor de dimensão $N \times 1$ onde os elementos correspondem aos erros.

Primeiramente, os valores dos coeficientes de regressão são estimados para que se possa ajustar um modelo de regressão múltipla. Em geral, é preferível representar a equação (2) em notação matricial, conforme em (3). Os dados são dispostos em uma matriz **X** com *N* linhas e *p+1* colunas.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix} \tag{3}$$

Os dados provenientes da espectroscopia NIR apresentam, em geral, multicolinearidade. A multicolinearidade é definida como a presença de um alto grau de correlação entre as variáveis independentes e resulta em estimativas duvidosas dos parâmetros de regressão, quando estimados pelo método dos quadrados mínimos (FERREIRA, 2015). O método via quadrados mínimos parciais (PLS) foi desenvolvido

com o intuito de suprir a necessidade de trabalhar com dados fortemente correlacionados. O PLS é um método fundamentado na compressão dos dados. Visa reduzir o espaço das medidas originais mantendo apenas as informações mais importantes, gerando assim alguns subespaços. Além disso, o PLS pode ser utilizado em dados com ruídos experimentais (BEEBE & KOWALSKI, 1987; WOLD et al., 2001).

Devido a correlação entre as variáveis independentes nos dados NIR, existem várias colunas com informações equivalentes de variância. Portanto, o PLS tem, como objetivo, reduzir estas colunas semelhantes para apenas uma. Consequentemente, há a redução do conjunto de dados original sem perda de informações cruciais do sistema completo de variáveis (FERREIRA, 2015).

Dentre os algoritmos para a execução do PLS, foi utilizado, neste trabalho, o algoritmo bidiagonal, baseado na decomposição de valores singulares (SVD). Segundo Martins et al. (2010), citado por Roque (2015), este é o que possui menor tempo de processamento computacional e também é o mais intuitivo.

O algoritmo bidiagonal é baseado na decomposição da matriz **X** em três matrizes, conforme em (4).

$$\mathbf{y} \rightarrow \mathbf{X} = \mathbf{URV^t} \tag{4}$$

em que, **UR** é a matriz de escores e **V** a matriz de *loadings*. O PLS considera tanto as informações presentes na matriz **X**, como também, as informações do vetor **y** para construção das variáveis latentes (WOLD; SJÖSTRÖM; ERIKSSON, 2001, citado por ROQUE, 2015).

Segundo Roque (2015), as colunas da matriz **U** e as linhas da matriz $\mathbf{V^t}$ criam os novos subespaços (variáveis latentes), que possuem informações presentes na matriz de dados **X** e no vetor **y**. Geralmente, as primeiras variáveis latentes (entre 2 a 10) informam praticamente toda (aproximadamente 100%) informação da matriz de dados **X** original.

O algoritmo para a execução do PLS está representado de forma sucinta a seguir (obtido de Martins, Teofilo & Ferreira, 2010, citado por Ferreira (2018) e Roque (2015).

1. Inicialize o algoritmo para a primeira componente:

$$\mathbf{X} = \mathbf{y}\mathbf{v}_\mathbf{1}^\mathbf{t}$$

$$\mathbf{X^t} = \mathbf{v_1}\mathbf{y^t}$$

$$\mathbf{X^t}\mathbf{y} = \mathbf{v_1}\mathbf{y^t}\mathbf{y}$$

$$\mathbf{X^t y(y^t y)^{-1}} = \mathbf{v_1(y^t y)(y^t y)^{-1}}$$

$$\mathbf{v_1} = \mathbf{X^t y(y^t y)^{-1}} \quad (5)$$

Usando a norma euclidiana, $\|v_1\| = \sqrt{v_1^t v_1}$, pode-se normalizar o vetor. Para isto, divida-o por sua norma. O vetor normalizado de (5), será:

$$\mathbf{v_1} = \frac{\mathbf{X^t y}}{\|\mathbf{X^t y}\|} \; ; \; \alpha_1 \mu_1 = \mathbf{X v_1} \quad (6)$$

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

2.1 $y_{i-1}\mathbf{v_1} = \mathbf{X^t}\mu_{\mathbf{i-1}} - \alpha_{i-1}\mathbf{v_{i-1}}$

2.2 $\alpha_i \mu_1 = \mathbf{X v_i} - y_{i-1}\mu_{\mathbf{i-1}}$

$\mathbf{V_h} = (\mathbf{v_1}, \ldots, \mathbf{v_h})$ $\mathbf{U_h} = (\mu_1, \ldots, \mu_i)$ e $\mathbf{R_h} = \begin{pmatrix} \alpha_1 & y_1 & & \\ & \ddots & & \\ & & \alpha_{k-1} & y_{k-1} \\ & & & \alpha_k \end{pmatrix}$

em que *h* é o número de variáveis latentes escolhido, prova-se que $\mathbf{U_h R_h} = \mathbf{X V_h}$ e então $\mathbf{R_h = U_h^t X V_h}$**.** Portanto, obtidas as matrizes **U**, **V** e **R** pode-se estimar a pseudo-inversa de Moore-Penrose de **X**, e estimar o modelo conforme em (7).

$$\mathbf{y = X\beta} \rightarrow \mathbf{X = U_h R_h V_h^t} \rightarrow \mathbf{y = U_h R_h V_h^t \beta} \rightarrow \mathbf{\widehat{\beta} = V_h R_h^{-1} U_h^t y} \quad (7)$$

O sistema estudado deve ser conduzido por um pequeno número de *h*, pois esta é a suposição básica para qualquer modelo PLS. Desta forma, a etapa de escolha do número de variáveis latentes é crucial para evitar sub ou sobre ajuste do modelo. Isto se justifica porque as matrizes reconstruídas dependem do valor de *h*. O sub ajuste ocorre quando a modelagem de dados é insuficiente para explicar toda informação e o sobre ajuste é quando há a inclusão de excesso de informação no modelo, que pode ser aleatória ou estar relacionada à presença de erros sistemáticos (BRERETON, citado por ROQUE, 2015).

Para definir o *h* pode-se aplicar o método da validação cruzada, que é alicerçado no procedimento de reamostragem. Primeiramente, um gráfico relacionando os erros nesta reamostragem versus *h* é construído. O ponto com menor erro é, então, selecionado. Normalmente o cálculo do erro utilizado é a raiz quadrada do erro quadrático médio de validação cruzada (RMSECV) (ROQUE, 2015), conforme apresentado em (8).

$$\mathrm{RMSECV} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \quad (8)$$

em que *N* é o número de amostras da validação, y_i é o *i-ésimo* valor referente ao vetor de referência **y**, o $\hat{y}_i$ é o *i-ésimo* valor previsto pelo modelo. O método de validação cruzada utilizado foi o *leave-one-out*, que remove uma amostra de cada vez (BRERETON, 2007).

2.2.3. Construção e validação do modelo

Para realizar o processo de construção e validação do modelo é importante seguir as seguintes etapas que estão representadas pela Figura 3:

i. Aplicar as transformações e os pré-processamentos no conjunto de calibração;
ii. Empregar o método de redução da dimensionalidade PLS e escolher o *h* a partir da validação cruzada;
iii. Construir o modelo de calibração;
iv. Validar o modelo: consiste em verificar sua capacidade preditiva no conjunto de predição. Isto é, aplicar os mesmos pré-tratamentos utilizados no conjunto de calibração e aplicar o modelo construído a esses dados, obtendo, então, o valor predito ($\hat{y}_p$).
v. Realizar a validação externa, em que é verificada a eficiência do modelo de calibração construído e sua capacidade preditiva através de estatísticas que inferem sobre o erro de predição.

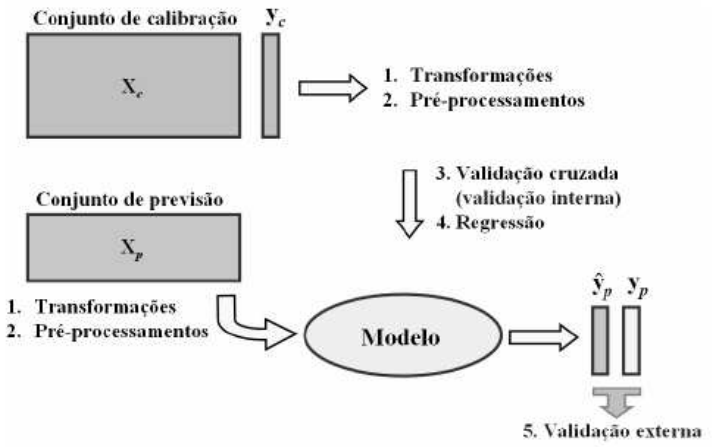


Figura 3 - Esquema usado para construir e validar o modelo (FONTE: TEÓFILO, 2007).

2.2.4. Comparação dos modelos

Para verificar a eficiência do modelo ajustado foram utilizados: o coeficiente de correlação (r) entre os valores preditos e os valores reais, que mede o grau de associação entre as variáveis $(-1 \leq \mathrm{r} \leq 1)$. A raiz quadrada do erro quadrático médio (RMSE) que mede as diferenças individuais entre os valores previstos pelo modelo $(\hat{y}_i)$ e os observados (y_i) (FERREIRA, 2015).

Esses parâmetros estatísticos, r e RMSE, podem ser calculados pelas equações (9) e (10), respectivamente. Segundo Ferreira (2015), o ideal são valores maiores para r e menores para RMSE.

$$r = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}} \tag{9}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}} \tag{10}$$

em que, $\hat{y}$ é o valor estimado, $\bar{\hat{y}}$ é o valor médio estimado, y representa os valores observados e $\bar{y}$ os valores médios observados. *N* é o número de amostras pertencente a cada subconjunto (teste ou treinamento) em que está analisando.

3. MATERIAL E MÉTODOS

3.1. Planejamento Experimental

O conjunto de dados foi obtido de dois experimentos de mistura entre cacau e sacarose (Experimento 1), e posteriormente cacau, sacarose e frutose (Experimento 2). Estes experimentos foram conduzidos no laboratório de análise de dados químicos multivariados (MCDALab) do Departamento de Química da Universidade Federal de Viçosa.

Inicialmente, foram testadas diferentes massas de cacau em uma embalagem de polietileno de baixa densidade, com o instrumento NIR portátil. O intuito foi verificar qual seria a massa mínima limite que poderia ser utilizada, de forma que o feixe de luz do instrumento não ultrapassasse através da embalagem. O valor obtido e fixado foi de quatro gramas de cacau.

Foram preparadas diferentes amostras para cada experimento. Em cada uma delas o valor em gramas de sacarose e frutose foram diferentes (Tabelas 1 e 2).

Foi definido um total de 29 amostras. Depois, utilizou-se uma margem de aproximadamente 25% do total das amostras para predição. Portanto, restaram 22 amostras para calibração e 7 para predição (Tabela 1 e 2).

Para o planejamento do Experimento 1, foi estipulado uma faixa percentual de 11,25 até 90% de sacarose para calibração, e de 20 até 85% para predição. Posteriormente, foi calculado o incremento para calibração e predição e obtidas as massas de sacarose (Tabela 1).

As massas de sacarose foram obtidas pela equação (11).

$$Massas\ de\ sacarose = \frac{-\%sacarose\ x\ massa\ cacau}{(\%sacarose - 100)} \quad (11)$$

Após, foi realizado o planejamento do Experimento 2. Como foi necessário acrescentar a massa de frutose, a massa total correspondeu à soma das três massas (cacau, sacarose e frutose). O percentual obtido de frutose nos alimentos industrializados, em geral, é de 0 a 20%. Portanto, o próximo passo foi calcular o incremento, e obter os valores de porcentagem de frutose para as 22 amostras de calibração e depois para as 7 de predição.

Tabela 1. Massa e percentual de cacau e sacarose utilizados no Experimento 1 para as etapas de calibração e predição.

Experimento 1				
Calibração				
Mistura	**Sacarose (g)**	**Cacau (g)**	**Sacarose (%)**	**Cacau (%)**
1	0,5070	4,001	11,25	88,75
2	0,7059	4,000	15,00	85,00
3	0,9231	4,001	18,75	81,25
4	1,1613	4,002	22,50	77,50
5	1,4237	4,001	26,25	73,75
6	1,7143	4,000	30,00	70,00
7	2,0377	4,000	33,75	66,25
8	2,4022	3,999	37,50	62,50
9	2,8085	3,999	41,25	58,75
10	3,2727	4,001	45,00	55,00
11	3,8049	4,000	48,75	51,25
12	4,4211	4,000	52,50	47,50
13	5,1429	4,000	56,25	43,75
14	6,0000	4,000	60,00	40,00
15	7,0345	4,002	63,75	36,25
16	8,3077	4,000	67,50	32,50
17	9,9130	4,000	71,25	28,75
18	12,0087	4,001	75,00	25,00
19	14,8235	4,000	78,75	21,25
20	18,8571	4,001	82,50	17,50
21	25,0909	4,000	86,25	13,75
22	36,0000	4,000	90,00	10,00
Predição				
Mistura	**Sacarose (g)**	**Cacau (g)**	**Sacarose (%)**	**Cacau (%)**
1	1,0043	4,000	20,00	80,00
2	1,5556	4,001	28,00	72,00
3	2,5574	4,000	39,00	61,00
4	5,5238	4,001	58,00	42,00
5	10,8148	4,002	73,00	27,00
6	16,0785	4,000	80,00	20,00
7	22,6667	3,999	85,00	15,00

Tabela 2. Massa e percentual de cacau, sacarose e frutose utilizados no Experimento 2 para as etapas de calibração e predição.

Experimento 2						
Calibração						
Mistura	**Sacarose (g)**	**Cacau (g)**	**Frutose (g)**	**Sacarose (%)**	**Cacau (%)**	**Frutose (%)**
1	0,5070	4,001	1,3061	8,72	68,81	22,47
2	0,7059	4,000	0,1623	14,50	82,17	3,33
3	0,9231	4,001	0,5818	16,77	72,66	10,57
4	1,1613	4,002	0,1156	22,01	75,80	2,19
5	1,4237	4,001	2,1719	18,74	52,66	28,59
6	1,7143	4,000	1,8212	22,75	53,08	24,17
7	2,0377	4,000	0,4082	31,61	62,06	6,33
8	2,4022	3,999	0,2716	35,97	59,96	4,07
9	2,8085	3,999	0,4895	38,48	54,81	6,71
10	3,2727	4,001	1,5385	37,14	45,40	17,46
11	3,8049	4,000	3,9929	32,25	33,90	33,84
12	4,4211	4,000	0,000	52,50	47,50	0,00
13	5,1429	4,000	0,336	54,26	42,20	3,54
14	6,0000	4,000	0,9483	54,80	36,54	8,66
15	7,0345	4,002	0,0349	63,55	36,14	0,32
16	8,3077	4,000	0,6871	63,93	30,78	5,29
17	9,9130	4,000	2,6173	59,97	24,20	15,83
18	12,0087	4,001	0,0733	74,66	24,89	0,46
19	14,8235	4,000	10,000	51,43	13,88	34,69
20	18,8571	4,001	5,1312	67,37	14,29	18,33
21	25,0909	4,000	6,8978	69,72	11,11	19,17
22	36,0000	4,000	0,8081	88,22	9,80	1,98
Predição						
Mistura	**Sacarose (g)**	**Cacau (g)**	**Frutose (g)**	**Sacarose (%)**	**Cacau (%)**	**Frutose (%)**
1	1,0043	4,000	0,1971	19,24	76,97	3,79
2	1,5556	4,001	3,9362	16,39	42,14	41,47
3	2,5574	4,000	2,4900	28,27	44,21	27,52
4	5,5238	4,001	1,3399	50,85	36,82	12,33
5	10,8148	4,002	0,0857	72,58	26,84	0,57
6	16,0785	4,000	0,7512	77,10	19,28	3,62
7	22,6667	3,999	0,3286	83,97	14,82	1,22

Para o cálculo das porcentagens de frutose, iniciou-se em zero, e depois foi adicionado o incremento. Com a massa de cacau e sacarose conhecida e com o percentual de frutose desejado, foi possível calcular as massas de frutose necessárias conforme a fórmula (12).

$$Massas\ de\ frutose = \frac{-\%frutose\ (massa\ cacau + massa\ sacarose)}{(\%frutose - 100)} \quad (12)$$

É importante mencionar que as massas de frutose foram aleatorizadas entre as amostras, para ficar linearmente independente (isento de correlação entre frutose com cacau e sacarose).

3.2. Preparo das Amostras

Com base nas Tabelas 1 e 2 foi possível realizar o preparo das amostras. Inicialmente, foram preparadas as amostras do Experimento 1. As massas foram pesadas por meio de uma balança semi-analítica. A pesagem ocorria de forma simultânea para cada uma das amostras, isto é, foi pesado o cacau, para a primeira amostra e depois a sacarose. As massas foram maceradas em um almofariz com auxílio de um pistilo. Após obter homogeneidade, a mistura foi inserida e identificada em embalagens de polietileno de baixa densidade (Figura 4 – Experimento 1). No final deste ciclo, foi realizado o mesmo procedimento para a segunda amostra e assim sucessivamente até a amostra 29.

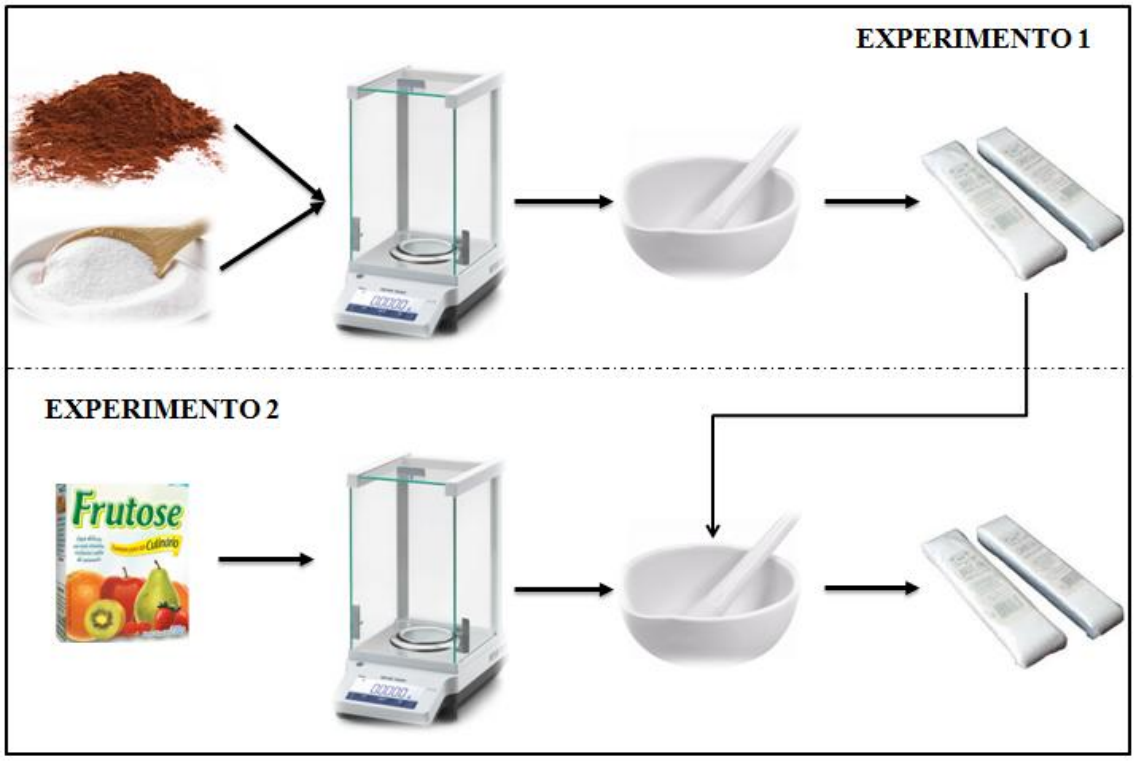


Figura 4– Esquema ilustrativo do processo de pesagem das amostras.

O procedimento adotado para as amostras do Experimento 2 foi realizado da seguinte forma: no planejamento experimental (Tabela 1) as massas das amostras do Experimento 1 foram aproveitadas para criar o Experimento 2. Portanto, foi necessário pesar apenas as massas de frutose na balança semi-analítica. À medida que eram pesadas, eram maceradas com as respectivas massas de cacau e sacarose das amostras do Experimento 1. Posteriormente, foram inseridas e identificadas em novas embalagens de polietileno de baixa densidade (Figura 4 – Experimento 2).

3.3. Análise Espectral das Amostras

As análises espectrométricas foram realizadas na parte externa das embalagens de polietileno de baixa densidade de cada uma das amostras. Os espectros foram obtidos a partir de dois instrumentos NIR. O primeiro instrumento foi o NIR de bancada Thermo Scientific Antaris II, avaliado em, aproximadamente, $ 64000.00 (USD). Para este instrumento, os espectros foram medidos por um espectrômetro de infravermelho próximo de transformada de Fourier (FT-NIR) controlado com o software TQ Analysis. Os espectros foram coletados entre 1000 e 2500 nm. Além disso, as leituras foram realizadas em triplicada. O segundo foi o NIR portátil, modelo Texas Instruments DLP NIRscan® Nano EVM, avaliado em, aproximadamente, $ 1000.00 (USD). Os espectros foram coletados entre 900 nm e 1700 nm. As leituras também foram realizadas em triplicada.

Os espectros foram exportados dos softwares de cada instrumento. Uma matriz de dados para cada conjunto de espectros foi montada e denominada matriz **X**, que contém as variáveis independentes. As linhas da matriz **X** correspondem às amostras e as colunas correspondem às variáveis (número de onda ou comprimento de onda). Um vetor contendo os respectivos valores determinados para as porcentagens de sacarose foi construído e denominado **y**, que é a variável dependente. O vetor **y** possui um número de linhas igual ao número de amostras (linhas) na matriz **X**.

3.4. Construção dos Modelos de Calibração Multivariada

Para a construção dos modelos foi utilizada a regressão PLS. Transformações e pré-processamentos foram realizadas nas matrizes **X** com o objetivo de encontrar o melhor modelo para predição. Os pré-tratamentos executados foram: (1) primeira derivada; (2) segunda derivada e (3) correção multiplicativa de sinal (MSC), (4) centragem na média, (5) alisamento.

Todas as rotinas dos métodos empregados foram implementadas no software R versão 3.5.0, utilizando programação própria e auxílio de pacotes disponíveis. Em especial, para o ajuste dos modelos, por meio do método PLS, foi utilizada a função *spls* do pacote SPLS (CHUNG; CHUN; KELES, 2018) do software R.

A qualidade dos modelos foi avaliada pela raiz quadrada do erro quadrático médio (RMSE), a qual pode ser calculada de acordo com a Equação (10). Pela Equação (9) calcula-se o coeficiente de correlação (r).

Para selecionar o número de variáveis latentes (*h*) do modelo, foi utilizado o método de validação cruzada *leave-one-out* (KOHAVI, 1995). O gráfico número de componentes versus RMSE foi utilizado para escolher o número de *h*.

Os valores de RMSE e r dos valores medidos e preditos foram utilizados como parâmetros de verificação do ajuste dos modelos. Como a regressão PLS foi aplicada, o conjunto de predição foi utilizado para verificar a capacidade preditiva do modelo e desta forma validar o modelo. Os valores de RMSE e r na predição, como também os valores dos erros relativos das amostras, foram considerados como parâmetro para verificar a capacidade de predição dos modelos construídos.

4. RESULTADOS E DISCUSSÃO

4.1. Resultados das análises dos Experimentos 1 e 2 utilizando o NIR de bancada e o portátil

Como pré-definido pelo planejamento experimental, a matriz **X** foi dividida em treinamento ($\mathbf{X_{tre}}$) e teste ($\mathbf{X_{tes}}$). Posteriormente, foram testadas diferentes transformações (alisamento (grau do polinômio = 2; janela = 5), MSC e primeira e segunda derivadas). Em todos os testes, as colunas de $\mathbf{X_{tre}}$ foram centradas na média (CM). Estas etapas foram realizadas para todas as análises realizadas.

Os espectros NIR das amostras dos Experimentos 1 e 2 e dos instrumentos NIR de bancada e portátil são apresentados na Figura 5.

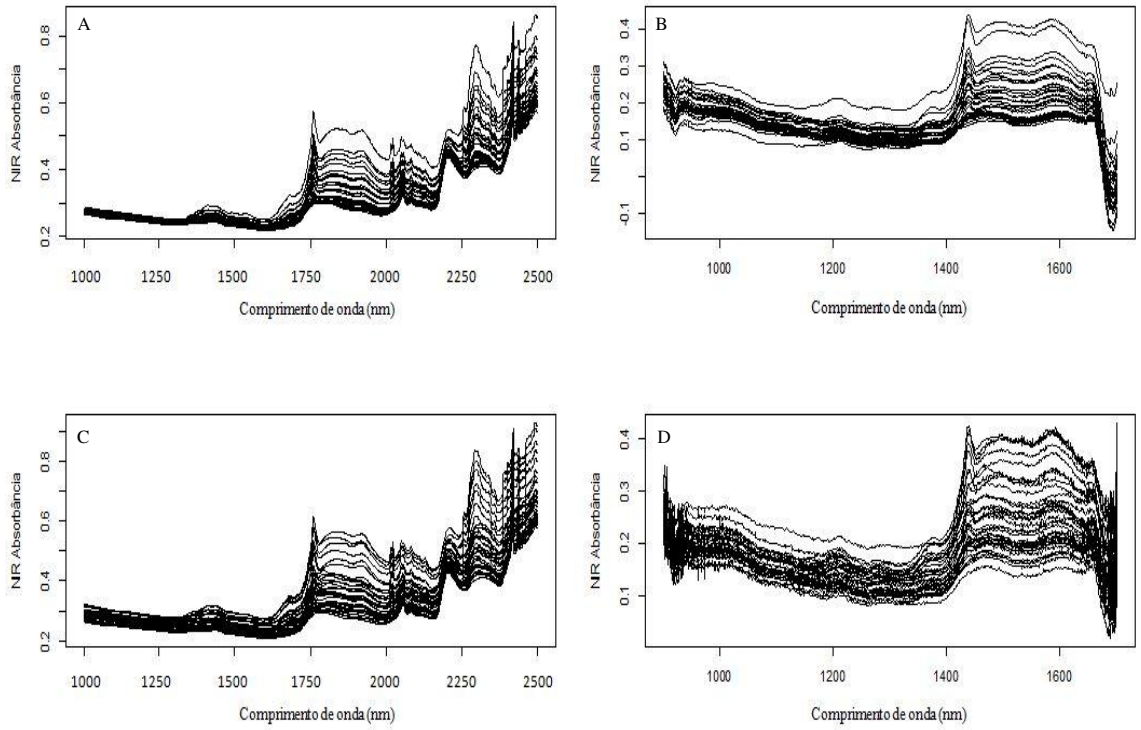


Figura 5 – Espectros obtidos com os dados originais do **(A)** NIR de bancada para Cacau e Sacarose; **(B)** NIR portátil para Cacau e Sacarose; **(C)** NIR de bancada para Cacau, Sacarose e Frutose; **(D)** NIR portátil para Cacau, Sacarose e Frutose.

Os parâmetros estatísticos calculados e os pré-tratamento selecionados para todos os modelos estão apresentados na Tabela 3.

Tabela 3. Resultados obtidos dos parâmetros estatísticos e dos pré-tratamentos obtidos para os Experimentos 1 e 2 utilizando os instrumentos NIR de bancada e portátil.

	Experimento 1 (Sac+Cac) Bancada	**Experimento 1 (Sac+Cac) Portátil**	**Experimento 2 (Sac+Cac+Fru) Bancada**	**Experimento 2 (Sac+Cac+Fru) Portátil**
H	5	10	6	5
RMSECV	0,0200	0,0400	0,0300	0,0500
Pré-tratamentos escolhidos	Alis, MSC e cm	Alis e cm	Alis, MSC e cm	Alis, MSC e cm
r_p	0,9971	0,9898	0,9949	0,9903
r^2_p	0,9943	0,9797	0,9898	0,9806
RMSEP	0,0597	0,0811	0,0773	0,0748

Sac = Sacarose; Cac = Cacau; Fru = Frutose; h = Número de variáveis latentes; RMSECV = Raiz quadrada do erro quadrático médio de validação cruzada; r_p = Coeficiente de correlação da predição; r^2_p = Coeficiente de determinação da predição; RMSEP = Raiz quadrada do erro quadrático médio de predição; MSC = Correção multiplicativa de sinal; Alis = Alisamento de Savitski-Golay; cm = Centragem na média.

A partir da análise dos valores de RMSECV, verificou-se que os melhores pré-tratamentos foram: alisamento, MSC e centragem na média, para o Experimento 1 (NIR bancada) e também para o Experimento 2 (NIR bancada e portátil). Já para o Experimento 1 (NIR portátil) os melhores foram: alisamento e centragem na média. Logo, os modelos foram construídos com base nestas transformações e pré-tratamentos (Tabela 3).

Analisando a Tabela 3, os valores de RMSECV e RMSEP estão muito próximos para todos os modelos, assim como os valores dos coeficientes de correlação (r_p). Além disso, pela Figura 6, percebe-se que há um ajuste linear entre os valores para todos os casos estudados.

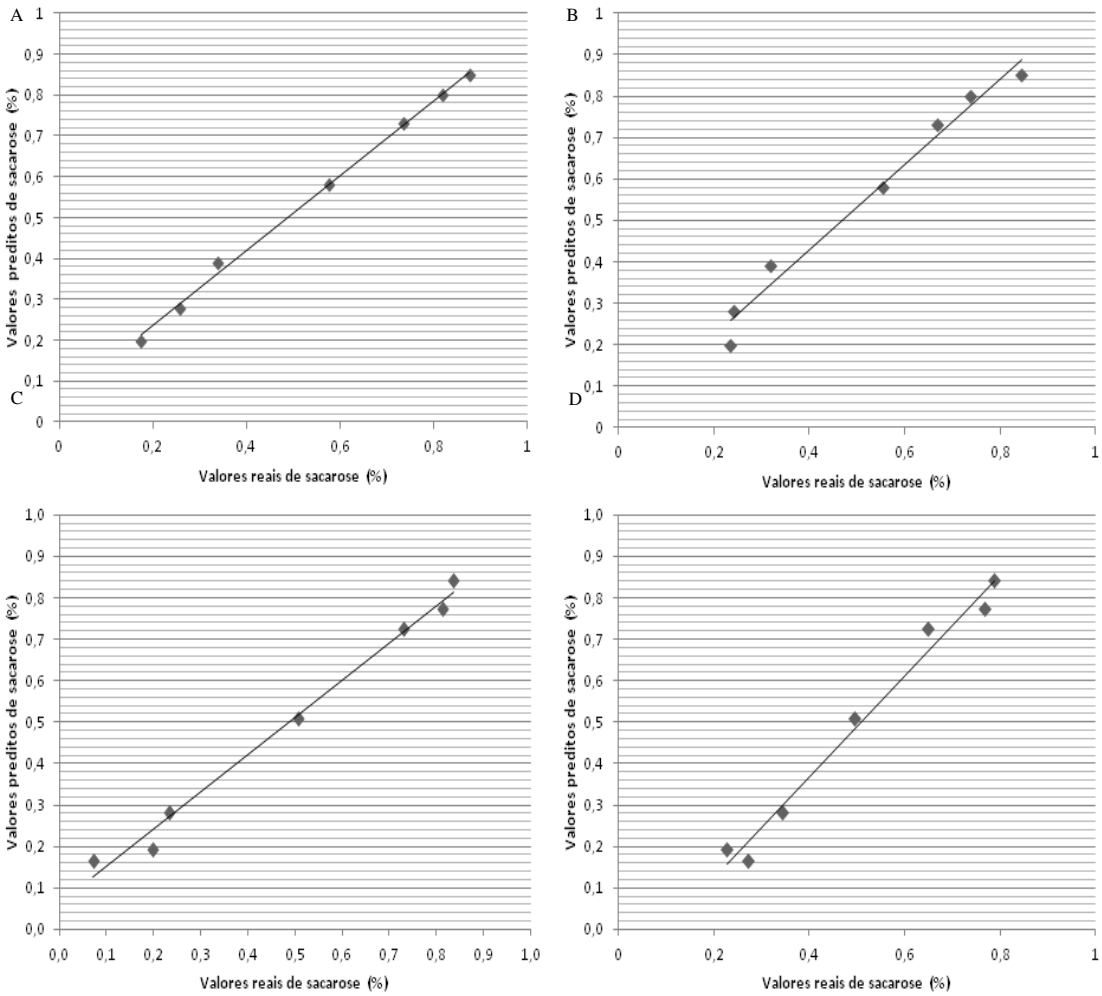


Figura 6 – Valores reais versus os preditos de sacarose para o conjunto de predição do **(A)** NIR de bancada para Cacau e Sacarose ($y = 0{,}0535 + 0{,}9152x$). **(B)** NIR portátil para Cacau e Sacarose ($y = 0{,}0162 + 1{,}031x$). **(C)** NIR de bancada para Cacau, Sacarose e Frutose ($y = 0{,}063 + 0{,}8974x$). **(D)** NIR portátil para Cacau, Sacarose e Frutose ($y = 0{,}1065 - 0{,}803x$).

Observando a Figura 7, nota-se que para o NIR de bancada e portátil para Cacau e Sacarose, os erros relativos estão altos para três amostras (1, 2 e 3), quando comparadas às demais. O estudo envolvendo o NIR de bancada para Cacau, Sacarose e Frutose, apresentou os erros relativos altos para as amostras 2 e 3. Já o NIR portátil para Cacau, Sacarose e Frutose, apresentou as amostras 1, 2 e 3 com o erro relativo mais alto.

Realizando um estudo comparativo entre os erros relativos, pela Figura 8, entre o NIR portátil com o de bancada para o Experimento 1 (cacau e sacarose), pode-se verificar que o NIR portátil apresentou percentual de erro relativo menor em apenas uma amostra (7) de um total de sete. Agora, realizando a comparação dos erros relativos entre o NIR portátil com o de bancada para o Experimento 2 (cacau, sacarose e frutose)

(Figura 8), pode-se verificar que o NIR portátil apresentou percentual de erro relativo menor para uma amostra (6) de um total de sete.

Levando em consideração: custo de instrumento, facilidade de manuseio e os resultados obtidos, o NIR portátil pode ser indicado para realização das análises para predição de sacarose. Por mais que a maioria das amostras apresentaram erros relativos maiores para o NIR portátil, em ambos os experimentos, estes valores foram muito próximos ao NIR de bancada, para ambos os Experimentos, apresentando pouca diferença (Figuras 7 e 8).

Realizando a diferença das médias dos valores em módulo dos erros relativos, das sete amostras, de cada instrumento e experimento, pode-se concluir que: a diferença das médias do portátil com o de bancada, foi de 4% para o Experimento 1 e de 6% para o Experimento 2. O que leva a concluir que o NIR portátil errou apenas 4%, em média, a mais no Experimento 1 e 6%, em média, a mais no Experimento 2. Sendo estes, valores irrelevantes levando em consideração ao custo do NIR portátil ser bem inferior ao de bancada e também a praticidade.

Para corroborar este resultado, foi aplicado o teste de hipóteses para proporção, com o intuito de avaliar a ocorrência dos erros relativos ao longo das amostras. O critério foi: se o erro relativo do NIR portátil foi menor do que o de bancada, foi atribuído a esta situação o número 1 (sucesso), caso contrário, 0 (fracasso). Foi utilizado um nível de significância igual a 5%, $\pi = 0{,}5$, e as hipóteses: H_0: $\pi = 0{,}5$ e H_1: $\pi \neq 0{,}5$. O p-valor obtido para o Experimento 1 e 2 foi exatamente 0,125, o que leva a concluir que os instrumentos, em ambos os experimentos, apresentam resultados de erro relativo estatisticamente iguais.

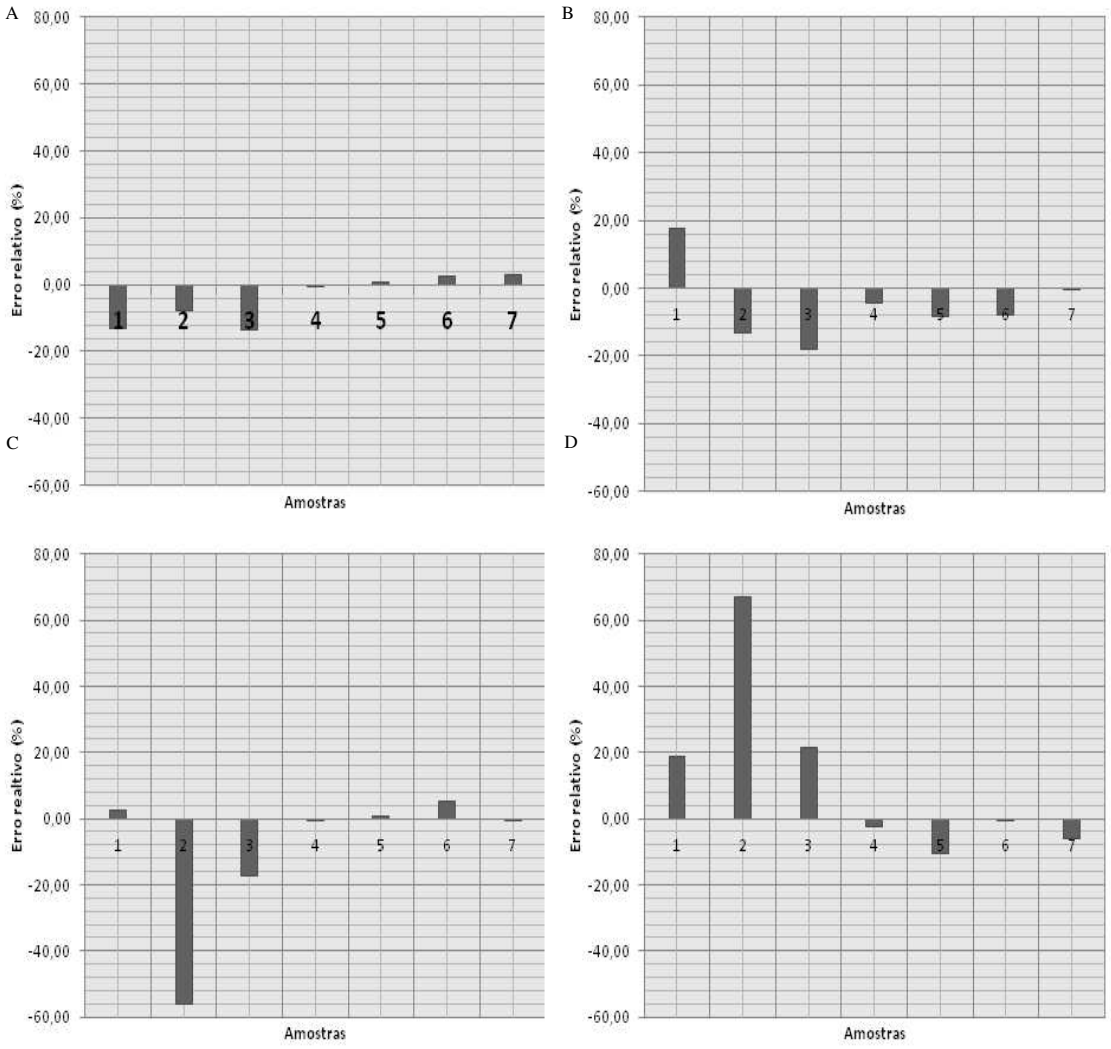
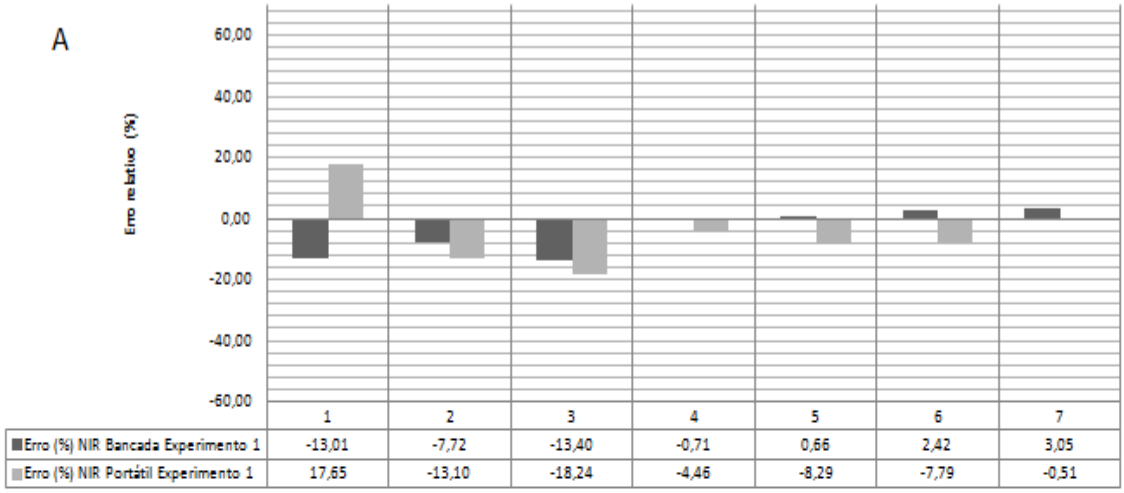


Figura 7 - Erros relativos no conjunto de predição (teste) para o **(A)** NIR de bancada para Cacau e Sacarose; **(B)** NIR portátil para Cacau e Sacarose; **(C)** NIR de bancada para Cacau, Sacarose e Frutose; **(D)** NIR portátil para Cacau, Sacarose e Frutose.



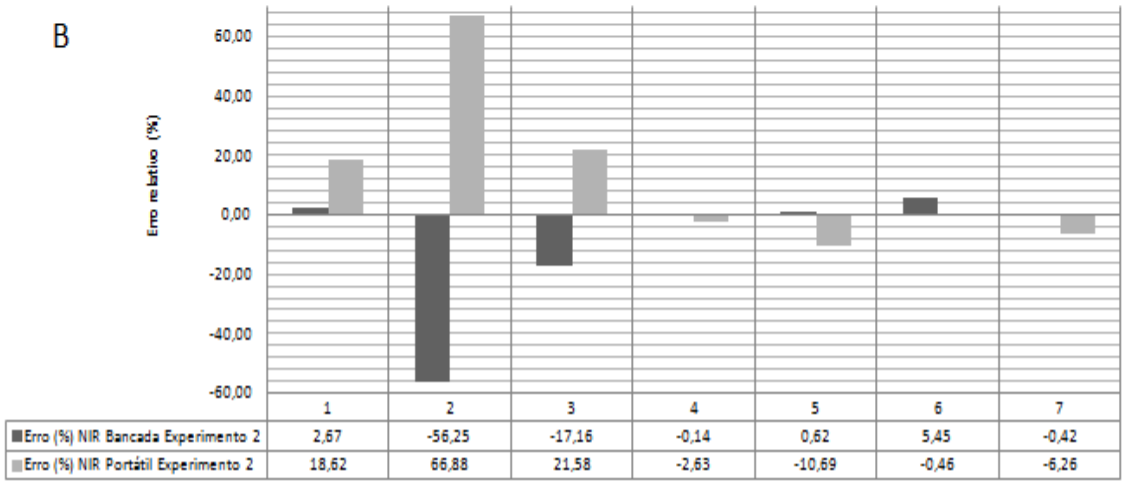


Figura 8 – Erros relativos das sete amostras entre os instrumentos NIR de bancada e portátil, no conjunto de predição (teste) para: **(A)** Experimento 1 (Cacau e Sacarose); **(B)** Experimento 2 (Cacau, Sacarose e Frutose).

4.2. Considerando a mesma faixa

As seções abaixo apresentam o estudo realizado com o mesmo banco de dados utilizado na seção acima. Porém, foi proposto o estudo entre os instrumentos utilizando a mesma faixa espectral conforme ilustrado pela Figura 9. Desta forma, as matrizes **X** dos Experimentos 1 e 2, dos dois instrumentos, foram reduzidas para incluir os comprimentos de onda conforme a especificação da Figura 9. Posteriomente foram realizados os passos da seção 2.2.3 para a construção dos modelos, para realizar o estudo comparativo entre os instrumentos em ambos os experimentos.

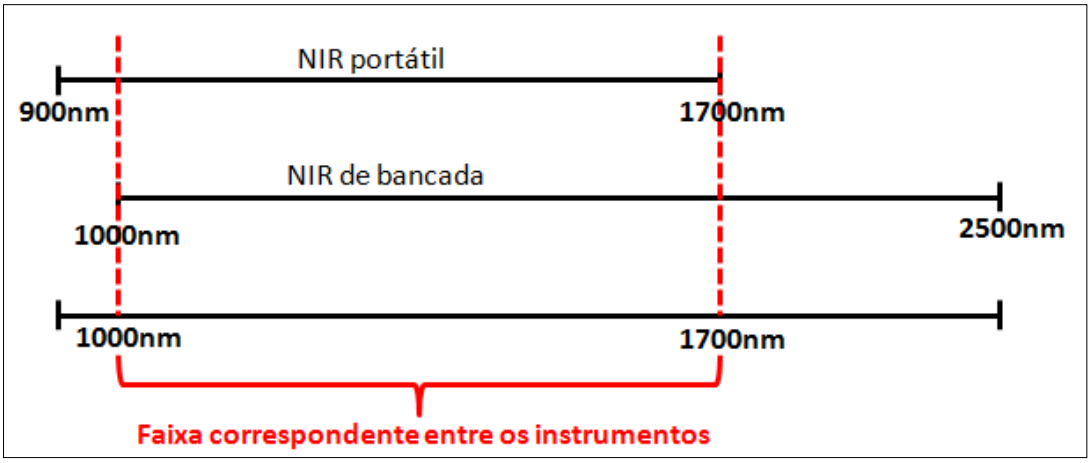


Figura 9 – Comprimentos de onda dos instrumentos NIR de bancada e NIR portátil e indicação da faixa de comprimento de onda comum a ambos os instrumentos.

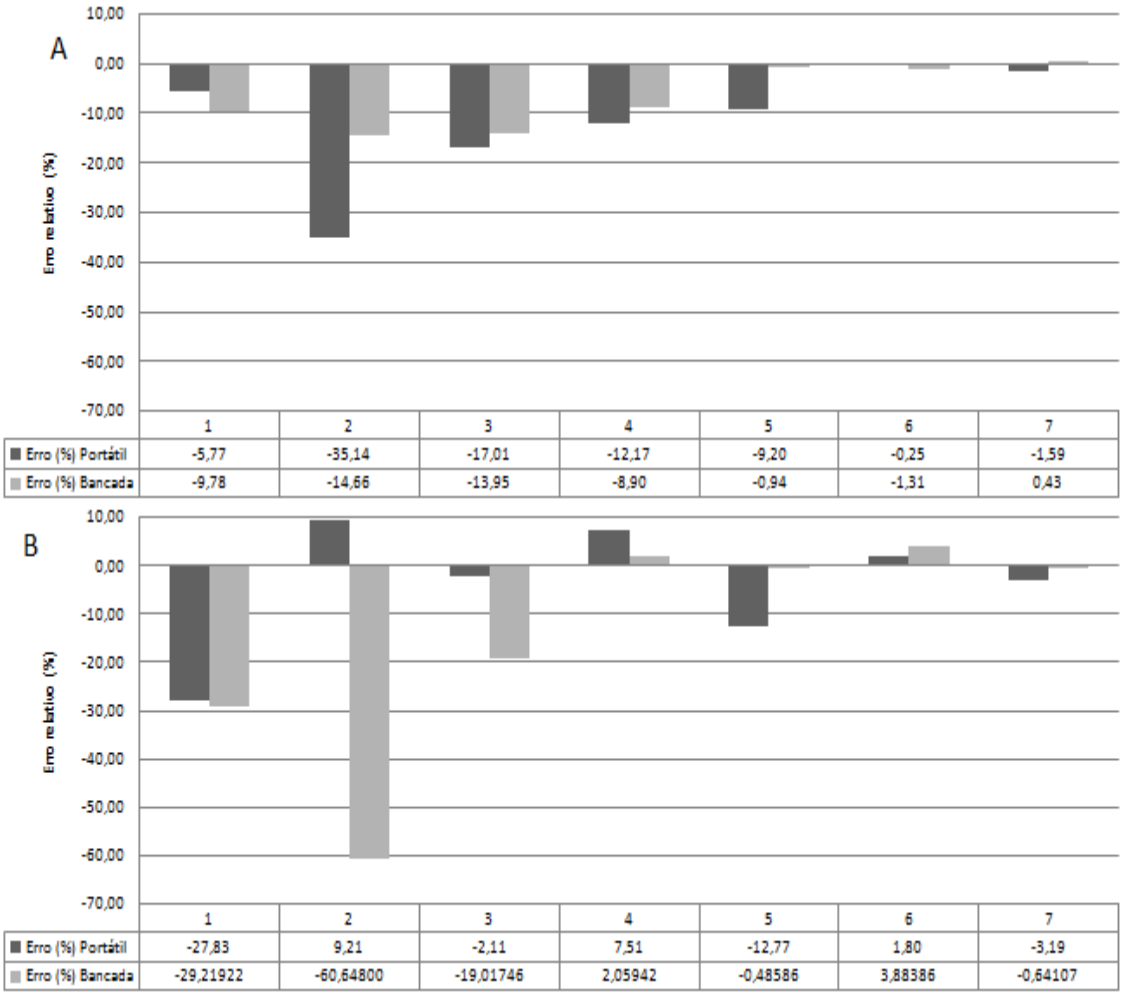


Figura 10 – Comparação entre os erros relativos no conjunto de predição (teste) para o NIR portátil e de bancada para: **(A)** o Experimento 1 (Cacau e Sacarose); **(B)** Experimento 2 (Cacau, Sacarose e Frutose), considerando a mesma faixa dos instrumentos.

Analisando a Figura 10 (A), os valores dos erros relativos foram todos negativos, isso se justifica porque ocorreu a predição de cada amostra abaixo do valor real esperado. Já a Figura 10 (B), aparesentou os erro relativos da predição oscilando entre valores negativos e positivos. Ou seja, houveram predições menores e maiores do que os valores reais.

Para o Experimento 1 é possível verificar que comparando os resultados obtidos dos coeficientes de correlação do NIR portátil (r = 0,9920) com o de bancada (r = 0,9977), ambos apresentam bons modelos de predição. Além disso, o portátil apresentou percentual de erro relativo mais baixo para duas amostras (1 e 6) (Figura 10 - A). Portanto, quando se utiliza a mesma faixa entre ambos os instrumentos, em princípio, o NIR portátil não foi o mais indicado para a predição de sacarose, o que era esperado, pois a resolução do NIR de bancada é maior na mesma faixa espectral. Porém, realizando o teste de proporção, com um nível de significância igual a 5%, $\pi = 0{,}5$, e as hipóteses: H_0: $\pi = 0{,}5$ e H_1: $\pi \neq 0{,}5$, o p-valor obtido foi 0,453. Isto leva a concluir que os instrumentos, em ambos os experimentos, apresetam resultados para erro relativo estatisticamente iguais. Desta forma, o NIR portátil pode ser escolhido para substituir o de bancada.

Com relação ao Experimento 2, é possível concluir que comparando os resultados obtidos dos coefieicnetes de correlação do NIR portátil (r = 0,9874) com o de bancada (r = 0,9974), ambos apresentam bons modelos de predição. Como também, pela Figura 10 (B), o portátil apresentou percentual de erro relativo mais baixo para quatro amostras (1, 2, 3 e 6). Portanto, quando se utiliza a mesma faixa entre ambos os instrumentos, o NIR portátil pode ser indicado para a predição de sacarose, o que não era esperado, considerando os motivos apresentandos acima. Realizando o teste de proporção, o p-valor obtido foi 1, isto é, há forte evidência de que os resultados obtidos dos erros relativos dos dois instrumentos são estatisticamente iguais, o que corrobora a escolha no NIR portátil como melhor.Afinal, possui um custo bem menor e mais prático na rotina das análises.

4.3. Predição para amostras industrializadas

Sete amostras de diferentes marcas de achocolatados industrializados foram coletadas para a predição de sacarose, utilizando os modelos obtidos nos Experimentos

1 e 2 com o NIR portátil da seção 4.1. O NIR portátil foi escolhido para esta análise, visto os resultados obtidos e pelas considerações citadas.

Analisando a Tabela 4 e a Figura 11, percebe-se que os modelos obtidos para os Experimentos 1 e 2 apresentaram valores de r consideravelmente altos e valores de RMSE consideravelmente baixos, o que indica que tem boa capacidade preditiva. Portanto, os modelos podem ser utilizados em predições de sacarose em achocolatados.

Tabela 4. Resultados obtidos para as amostras industrializadas, utilizando os modelos dos Experimentos 1 e 2 e o NIR portátil.

	r	**r^2**	**RMSE**
Experimento 1	0,9568	0,9154	0,1079
Experimento 2	0,9644	0,9301	0,1478

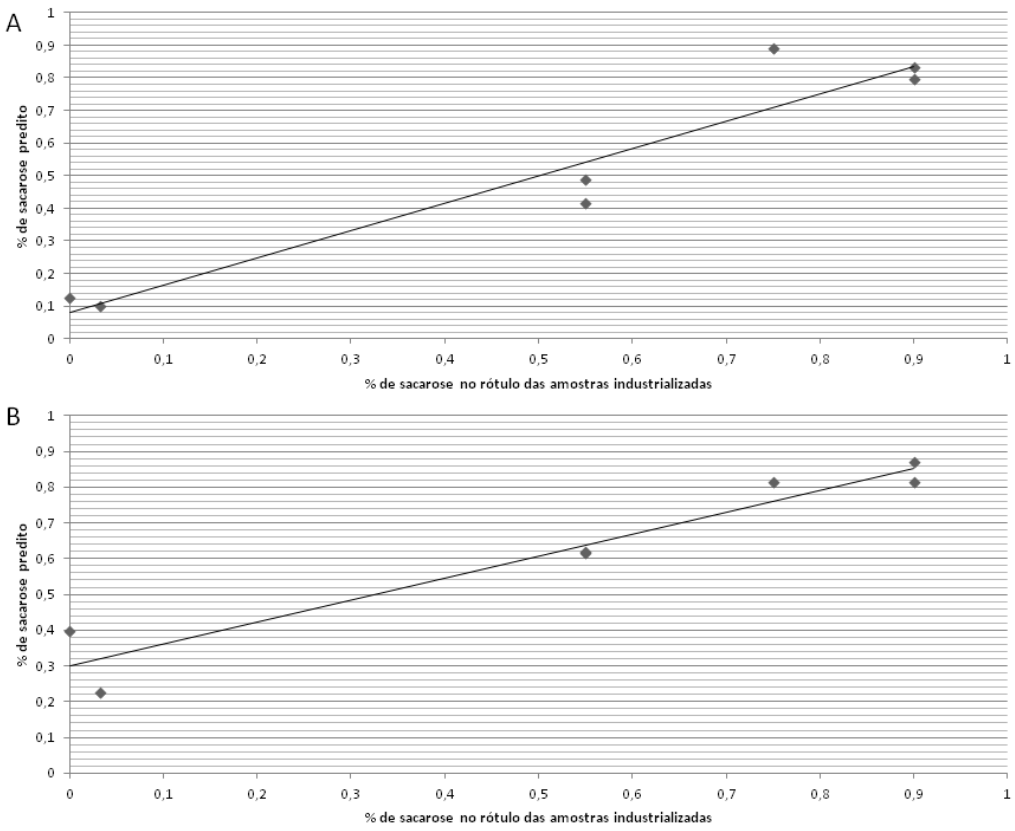


Figura 11 – Percentual de sacarose predito versus percentual de sacarose de acordo com as informações nutricionais contidas nos rótulos dos produtos industrializadas utilizando: **(A)** o modelo do Experimento 1 (Cacau e Sacarose), **(B)** o modelo do Experimento 2 (Cacau, Sacarose e Frutose), com o NIR portátil.

Pela Figura 11 é possível verificar que há um ajuste linear entre os valores de sacarose preditos e a informação do percentual de sacarose indicada nos rótulos dos produtos industrializados, para ambos os modelos construídos. O que é corrobora a eficiência dos modelos.

5. CONCLUSÕES

Para ambos os experimentos, observou-se que a diferença dos erros entre os instrumentos (NIR portátil e de bancada) não foram tão expressivas. Desta forma, o portátil apresentou ser uma boa alternativa para realizar as análises para predição de sacarose, considerando o custo-benefício.

Ao utilizar a mesma faixa espectral dos instrumentos, pode-se concluir que o NIR portátil foi o recomendado para predição de sacarose em estudos de mistura de cacau e sacarose, e para estudos que envolvem frutose na mistura.

Considerando os resultados para as amostras dos produtos industrializados, observou-se que os modelos propostos pelos Experimentos 1 e 2 são capazes de predizer, de forma acurada, os valores de sacarose correspondente nas embalagens dos produtos de interesse. Cabe dizer que para o Experimento 2 (Cacau, Sacarose e Frutose), os resultados para predição de sacarose foram melhores.

Este trabalho permitiu indicar o NIR portátil, com relação a custos e praticidade de manuseio, para as indústrias de achocolatados que praticam o controle estatístico de processo e que desejam garantir a qualidade de seus produtos. Não obstante, por mais que o NIR portátil possa aparentar ser uma alternativa pertinente em substituição ao NIR de bancada, mais estudos devem ser realizados em outros conjuntos de dados para garantir maior consistência nas comparações.

6. REFERÊNCIAS

BAYE, T. M.; PEARSON, T. C.; SETTLES, A. M. Development of a calibration to predict maize seed composition using single kernel near infrared spectroscopy. **Journal of Cereal Science**, v. 43, n. 2, p. 236–243, 2006.

BEEBE, K. R.; KOWALSKI, B. R. An introduction to multivariate calibration and analysis. **Analytical Chemistry**. v. 59, n. 17, p. 1007 A–1017 A, 1987.

BRERETON, R. G. **Applied chemometrics for scientists**. England: John Wiley & Sons, Ltd., 2007.

CASCANT, M. M.; RUBIO, S.; GALLELLO, G.; PASTOR, A.; GARRIGUES, S.; GUARDIA, M. Burned bones forensic investigations employing near infrared spectroscopy. **Vibrational Spectroscopy**. v. 90, p. 21–30, 2017.

CATELANI, T. A.; SANTOS, J. R.; PÁSCOA, R. N. M. J.; PEZZA, L.; PEZZA, H. R.; LOPES, J. A. Real-time monitoring of a coffee roasting process with near infrared spectroscopy using multivariate statistical analysis: a feasibility study. **Talanta**. v.179, p. 292-299, 2018.

CHUNG, D.; CHU, H.; KELES, S. **spls**. Package 'spls'. CRAN, 2018. Disponível em:< https://cran.r-project.org/web/packages/spls/spls.pdf >. Acesso em: Jan. 2019.

CLAVAUD, M.; ROGGO, Y.; D´EGARDIN, K.; SACR´E, P.; HUBERT, P.; ZIEMONS, E. Moisture content determination in an antibody-drug conjugate freeze-dried medicine by near-infrared spectroscopy: acase study for release testing. **Journal of Pharmaceutical and Biomedical Analysis**. p. 01-32, 2016.

COLLINS, C. H. **Princípios básicos de cromatografia. In: Introdução a métodos cromatográficos**. 7 ed. Campinas: Editora da UNICAMP, 1997. p.11-27.

EDUARDO, M. F.; LANNES, S. C. S. Achocolatados: análise química. **Brazilian Journal of Pharmaceutical Sciences**. v. 40, n. 3, jul./set., 2004.

ESTOPA, R. A.; MILAGRES, R. F.; GOMES, F. J. B.; AMARAL, C. A. S. Caracterização química da madeira de eucalyptus benthamii por meio de espectroscopia NIR. **O PAPEL**. v. 78, n. 2, p. 75 – 81, 2017.

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

FERREIRA, M. M. C. **Quimiometria – Conceitos, Métodos e Aplicações.** Campinas, SP: Editora Unicamp, 2015. 493 f.

FERREIRA, R. A. **Comparação de métodos de seleção de variáveis em regressão aplicados a dados genômicos e de espectroscopia NIR**. 2018. 53 f. Dissertação (Mestrado em Estatística Aplicada e Biometria) -Universidade Federal de Viçosa, Viçosa, 2018.

HAIR JR, J. F.; BLACK, C. W.; BABIN, B. J.; ANDERSON, R. E. **Multivariate Data Analysis**. 7th Edition, Prentice Hall. 2009.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. **Ijcai**, v. 14, n. 10, p. 1137-1145, 1995.

NASCIMENTO, C. C.; BRASIL, M. M.; NASCIMENTO, C; S.; BARROS, S. V. S. Estimativa da densidade básica da madeira de Eschweilera odora (Poepp.) Miers por espectroscopia no infravermelho próximo. **Ciência da Madeira (Brazilian Journal of Wood Science)**. v. 8, n.1, p. 42-53, 2017

OLAREWAJUA, O. O.; BERTLING, I.; MAGWAZA, L. S. Non-destructive evaluation of avocado fruit maturity using near infrared spectroscopy and PLS regression models. **Scientia Horticulturae**. v. 199, p. 229–236, 2016.

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **Journal of the Brazilian Chemical Society**, v. 14, p. 198 219, 2003.

PERMANYER, J. J.; PEREZ, M. L. Compositional Analysis of Powdered Cocoa Products by Near Infrared Reflectance Spectroscopy. **Journal of Food Science.** v. 54, n. 3, 1989.

R Core Team (2017). **R: A language and environment for statistical computing. R Foundation for Statistical Computing**, Vienna, Austria. Disponível em: http://www.R-project.org/. Acesso em: Nov. 2018.

ROQUE, J. V. **Desenvolvimento de modelos de regressão multivariada para determinação de ésteres de forbol em sementes de jatropha curcas L. usando**

espectroscopia e quimiometria. 2015. 84 f. Dissertação (Mestrado em Agroquímica) - Universidade Federal de Viçosa, Viçosa, 2015.

SILVA, E. E.; SILVA, L. M.; WADT, P. G. S.; MARCHAO, R. L. Espectroscopia de infravermelho próximo na predição de propriedades químicas e físicas de solos de Roraima. **Biota Amazônia**. v.7, n.2, p.31-35, 2017.

TEÓFILO, R. F. **Métodos quimiométricos em estudos eletroquímicos de fenóis sobre filmes de diamante dopado com boro**. 2007. 329 f: Tese (Doutorado em Química) - Universidade Estadual de Campinas, 2007.

WILLIAMS, P.; NORRIS, K. **Near-infrared technology**. 2nd ed. Saint Paul: American Association of Cereal Chemistry, 2001. 296 p.

WOLD, S. Chemometrics: what do we mean with it, and what do we wantfrom it? **Chemometrics and Intelligent Laboratory Systems**. p. 30-109, 1995.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**. v.58, p.109-130, 2001.

YANG, Z.; NIE, G.; PAN, L.; ZHANG, Y.; HUANG, L.; MA, X.; ZHANG, X. Development and validation of near-infrared spectroscopy for the prediction of forage quality parameters in Lolium multiflorum. **PeerJ**. v. 5, 2017.