UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Dissertação de Mestrado

Uma Estratégia de Validação Híbrida para Calibração Multivariada Baseada na Seleção de Amostras Fixadas pelo Algoritmo SPXY

João Batista de Sousa Costa

João Pessoa - PB - Brasil

Junho/2021

UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

Dissertação de Mestrado

Uma Estratégia de Validação Híbrida para Calibração Multivariada Baseada na Seleção de Amostras Fixadas pelo Algoritmo SPXY

João Batista de Sousa Costa

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba como parte dos requisitos para obtenção do título de Mestre em Química, área de concentração Química Analítica.

Orientador: Prof. Dr. Edvan Cirino da Silva

João Pessoa - PB - Brasil

Junho/2021

Catalogação na publicação
Seção de Catalogação e Classificação

C838e Costa, João Batista de Sousa. Uma estratégia de validação híbrida de calibração multivariada baseada na seleção de amostras fixadas pelo algoritmo SPXY / João Batista de Sousa Costa. João Pessoa, 2021.
72 f. : il.

Orientação: Edvan Cirino da Silva.
Dissertação (Mestrado) - UFPB/CCEN.

1. Química - Validação cruzada. 2. Amostras fixadas. 3. SPXY. 4. Calibração multivariada. 5. MLR-APS. I. Silva, Edvan Cirino da. II. Título.

UFPB/BC CDU 54(043)

Elaborado por WALQUELINE DA SILVA ARAUJO - CRB-15/514

Uma estratégia de validação híbrida de calibração multivariada baseada na seleção de amostras fixadas pelo algoritmo SPXY.

Dissertação de Mestrado apresentada pelo aluno **João Batista de Sousa Costa** e aprovada pela banca examinadora em 28 de junho de 2021.

Prof. Dr. Edvan Cirino da Silva
Departamento de Química-UFPB
Orientador/Presidente

Profa. Dra. Patrícia Kaori Soares
Escola de Ciências e Tecnologia-UFRN
Examinadora Externa

Prof. Dr. Ricardo Alexandre Cavalcanti de Lima
Departamento de Química-UFPB
Examinador Interno

Assinaturas da Banca realizadas em modo Webconferência em 28/06/2021, digitalizadas e certificadas pelo Prof. Dr. Edvan Cirino da Silva (SIAPE 1228750) em 28/06/2021.

Aos meus pais, Francisco e Joana,
Meus irmãos, Diogenes, José Carlos, Djaciane e Denise
– também à minha noiva Jaqueline –,
com muito amor e carinho dedico

Agradecimentos

- A Deus pelo dom da vida e por nos proporcionar as faculdades da consciência e do pensamento;

- A Alisson,Andre e Wallison, pela amizade e conversas divertidas ao longo de minha jornada no Mestrado;

- Ao professor Edvan Cirino da Silva, pela orientação, confiança e conselhos durante a inciação cientifica e mestrado;

- Ao Dr Sofácles Figueiredo Carreiro Soares, pela fundamental ajuda na elaboração e implementação do cógido do algortimo em Matlab;

- À familia LAQA que me acolheu de braços abertos;

- Ao professor Mario Cesar Ugulino de Araujo por me abrir as portas do LAQA;

- À turma da salinha de estudos, onde conheci pessoas incriveis e fiz amizades incriveis,essa dissertação tem a contribuição de cada um de vocês;

- Agradeço ao Instituto Nacional de Ciências e Tecnologias Analíticas Avançadas (INCTAA) pela aquisição do Software Matlab® com recursos do CNPq;

- À CAPES pela bolsa concedida;

SUMÁRIO

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

CV-Cross Validation (Validação Cruzada)

CVH- Hybrid Cross Validation (Validação Cruzada Híbrida)

CVLOO–Leave One Out Cross Validation (Validação Cruzada com uma amostra deixada de fora)

MLR– Multiple Linear Reression (Regressão Linear Multipla)

PCs–Principal Component (Componentes Principais)

PCA - Principal Component Analysis (Análise de Componentes Principais)

PCR - Principal Component Regression (Regressão de Componentes Principais)

PLS - Partial Least Square Regression (Regressão por Minimos Quadrados Parciais)

PRESS - Prediction Residual Error Sum of Squares (Somatório dos Quadrados dos Erros Residuais de Predição)

RMSECV - Raiz quadrada do Erro medio de Validação Cruzada

RMSEP - Raiz quadrada do Erro medio de Predição

RMSECV -Square Root Mean Square Cross Validation Error (Raiz quadrada do Erro Quadrático médio de Validação Cruzada).

RSCV - Cross Validation with Split of Representative Samples (Validação Cruzada com divisão de Amostras Representativas)

VLs - Variáveis Latentes

APS - Algoritmo de Projeções Sucessivas

SPXY- Sample set Partitioning based on joint x-y distances (Particionamento do conjunto de amostra com base nas distâncias x-y)

Resumo

Título: **Uma Estratégia de Validação Híbrida para Calibração Multivariada Baseada na Seleção de Amostras Fixadas pelo Algoritmo SPXY**

Autor: **João Batista de Sousa Costa**

Em análise multivariada, é usual deparar-se com o problema recorrente de particionar o conjunto de dados de modo a obter, para o subconjunto de calibração, as amostras mais representativas e que cubram a fronteira do espaço amostral de natureza multidimensional. O presente trabalho teve como objetivo desenvolver uma estratégia de validação híbrida para calibração multivarida (a exemplo da MLR-APS e do PLS group-out), a qual previne problemas de extrapolação e proporciona modelos com maior capacidade preditiva e robustez. A estratégia propostautiliza o algoritmo SPXY para selecionar as amostras de fronteira do espaço experimental para a calibração eque sejammais representativaspor explorar a estatística de X (respostas instrumentais) e Y (parâmetro de interesse). Para avaliar seu desempenho, foram empregados dois conjuntos de dados de NIR. O primeiro envolve a análise de amostras detrigo nas quaisfoi determinado o conteúdo de proteína; o segundo refere-se à determinação do teor de umidade em milho.Na validação híbridaaplicada à modelagem PLS grou-out, não foi possível avaliar a variabilidade do RMSEP em função do índice de amostras. Isso porque o mesmo não apresentava uma variabilidade significativa dos resultados para 20 execuções, impossibilitando uma melhor avaliação da estratégia proposta.Na modelagem MLR-APS, observou-se uma variabilidade em termos de RMSECV e RMSEP, tornando possível a avaliação da influência das amostras fixadas na capacidade preditiva dos modelos. Os modelos resultantes da fixação de amostras de fronteira na calibração apresentaram os maiores coeficientes de correlação, os quaisforam iguais a, respectivamente, 0,9996 e 0,9934 para o conjunto de dados de milho e de trigo. Os valores de RMSEP e RMSECV para os dois conjuntos apresentaram uma diminuição significativa. De fato, foram obtidos, repectivamente, os valores 0,194e 0,163% (m/m) para o conteúdo de proteína no trigo e 0,0121 e 0,0061% (m/m) para umidadenas amostras de milho. Um número de variáveis menor foi também obtido. A estratégiade validação híbrida é uma alternativa viável paracalibração multivariada, proporcionado modelosmais parcimoniosos e com maior robustez e capacidade preditiva.

Palavras-chave: **Validação Cruzada; Amostras fixadas; SPXY; Calibração Multivariada; PLS;APS-MLR**

Abstract

Title: **A Hybrid Validation Strategy for Multivariate Calibration Based on Fixed Sample Selection by SPXY Algorithm**

Autor: **João Batista de Sousa Costa**

In multivariate analysis, it is usual to come across the recurring problem of partitioning the dataset in order to obtain, for the calibration subset, the most representative samples that cover the boundary of the multidimensional sample space.The present work aimed to develop a hybrid validation strategy for multivariate calibration (such as MLR-APS and PLS group-out), which prevents extrapolation problems and provides models with greater predictive capacity and robustness.The proposed strategy uses the SPXY algorithm to select the experimental space boundary samples for calibration and which are more representative for exploring the statistics of X (instrumental responses) and Y (parameter of interest). To assess its performance, two sets of NIR reflectance data were used.The first involves the analysis of wheat samples in which the protein content has been determined; the second refers to the determination of the moisture content in corn. In the hybrid validation applied to the PLS grou-out modeling, it was not possible to assess the variability of the RMSEP as a function of the sample index. This is because it did not present a significant variability of results for 20 executions, making it impossible to better evaluate the proposed strategy. In the MLR-APS modeling, a variability was observed in terms of RMSECV and RMSEP, making it possible to assess the influence of fixed samples on the predictive capacity of the models.The models resulting from the fixation of frontier samples in the calibration showed the highest correlation coefficients, which were equal to, respectively, 0.9996 and 0.9934 for the corn and wheat dataset. The values of RMSEP and RMSECV for the two sets showed a significant decrease.In fact, the values of 0.194 and 0.163 % (m/m) were obtained, respectively, for the protein content in wheat and 0.0121 and 0.0061 % (m/m) for moisture in the corn samples. A smaller number of variables was also obtained. The hybrid validation strategy is a viable alternative for multivariate calibration, providing more parsimonious models with greater robustness and predictive capacity.

Keywords: **Cross validation; Fixed samples, SPXY algorithm; Multivariate calibration; PLS; APS-MLR.**

CAPÍTULO 1
Introdução

INTRODUÇÃO

1.1 Caracterizando a problemática e a proposta de solução

A calibração multivarida é um processo matemático que tem sido implementado com sucesso nas determinações quantitativas de analitos em amostras com matrizes complexas, especialmente usando técnicas espectroanalíticas[1]. Esse fato se deve, entre outras razões, ao problema recorrente de multicolinearidadeentre as variáveis (comprimentosde onda)e de sobreposição de bandas de concomitantes sobre as do(s) analito(s) quepodem ocorrer nosespectrosUV-Vis, NIR, RMN, etc[88]. Nesse cenário, a regressão linear múltipla (Multiple Linear Regression-MLR) e por mínimos quadrados parciais (Partial Least Squares-PLS) destacam-se entre as técnicasquimiométricasusadas na calibração multivariadapara superar esses problemas[14,59,62].

O uso consolidado do PLS em calibração multivariada pode ser atribuído àrobustez ealta capacidade preditiva dos modelos[5], como resultado damelhor correlação obtidaentre as variáveis latentes associadas aos escoresdos sinais e ao(s) parâmetro(s) de interesse. A regressão PLS possui a vantagem, quando comparada à técnica MLR,de permitir umamodelagemmais eficientena presença de dados ruidosos e acentuada multicolinearidade entre as variáveis[19]. Além disso, permite implementar uma calibração multivariada envolvendo matrizes (no sentido matemático do termo) com alta dimensão[19].Por outro lado, a técnica MLR tem a grande vantagem deoperar no domínio original dos dados, gerando modelosde mais fácil compreensão einterpretação química dos resultados da predição[76].O modelo MLR atrelado á seleção de variáveis se mostrou mais adequado para avaliação da estratégia proposta, tendo em vista que o mesmo apresentou uma maior sensibilidade quando comparado ao modelo PLS.

Para a construção de um modelo MLR, éfundamentalque os dados apresentem baixamulticonearidade entre as variáveis associadas aos sinais analíticos espectrométricos e que a fronteira amostral seja delimitada(específico da estratégia proposta), tendo isso em vista o algoritmo SPXY é utilizado para seleção dessas amostras(amostras mais externas). De fato, as variáveis com expressiva multicolinearidade fornecem informação redundante que dificultam o cálculo matricial naregressão e na determinação da variância dos coeficientes, prejudicando a escolha dos coeficientes de regressão adequados para os modelos MLR[76,78]. Os problemas de alta dimensionalidade e multicolinearidade podem também ser supera

dosmediante a seleção das variáveisque contenham informações úteis e não redundantes a fim de viabilizar a modelagem MLR e melhorara robustez e capacidade preditiva dos modelos[78]. Para esse propósito, pode-se recorrer aosalgoritmos de seleção de variáveis[80].

Para seleção de variáveis usandotécnicas determinísticas – que são interessantes, sobretudo, para a interpretação química dos resultados –, destaca-se o algoritmo das projeções sucessivas (APS) proposto por Araújo et al[78] para calibração MLR usando dadosUV-Vis, por ser um método que permite a obtenção de apenas um subconjunto de variáveis e que não estão atreladas a probabilidade a priori, facilitando assim a compreensão do modelo[79].Posteriormente, o APS foitambém utilizado para seleção de variáveis empregando dados espectrométricos NIR[84].Nesse trabalho, demonstrou-se que o APS proporciona as menores raizes quadradas de erros médios de previsão (RMSEP-Root Mean Square Error Prediction)[78]e menores raizes quadradas de erros medios de validação cruzada (RMSECV-Root Mean Square Error Cross validation)[80].

Em calibração multivariada via PLS, a validação cruzada (CV-Cross-validation) é o procedimento mais comumente usado para definir o número de componentes a serem definidos para o modelo[15,16,38,40].Muitas variantes de CV são utilizadas e a mais usual é a validação cruzada de exclusão única (CVLOO-Leave-one-out cross validation). Consiste na retirada deuma amostrapor vezda calibração, cujo parâmetro de interesse deve ser predito[15,16,40]. A CVLOO pode apresentar problema de extrapolação no modelo, o que acarreta um precária ou nenhuma capacidade preditiva[7, 19,41]. Na literatura não foi relatada nenhuma estratégiaquimiométrica para contornar esse problema– não apenas em calibração PLS, mas também para MLR combinada a técnicas de seleção de variáveis.

Em validação cruzada as amostras são selecionadas aleatoriamente, fato este que pode comprometer a sua significância para o conjunto de dados. O Algoritmo de validação cruzada com divisão representativa(RSCV- Representative Spltiting Cross Validation), no qual um algoritmo seleciona as amostras mais representativas.

No presente trabalho, desenvolveu-se uma nova estratégia quimiométrica – denominada aqui “validação híbrida” – para a validação de modelos em calibração PLS e MLR.A validação híbrida proposta consiste na seleção, usando o algoritmoSPXY (Sample set Partitioning based on joint x-y distances)[9], das amostras de fronteira(mais externas) do espaço experimental e as amostras restantes são selecionadas de maneira aleatória.Assim, é possível evitar o problema de extrapolação, melhorandoa capacidade preditiva e robustez dos modelos. Para esse propósito, faz-se necessário que as amostras fixadas – além de garantirem a

cobertura na fronteira do espaço experimental (amostral) explorado – sejam estatisticamente representativas para o conjunto de dados original[6,40].

1.2. Objetivos

1.2.1. Objetivo geral

O objetivo central deste trabalho foi propor uma nova estratégia de validação híbrida para calibração multivariada baseada na seleção de amostras fixadas (na fronteira do espaço amostral),empregando-se o algoritmo SPXY. Com a aplicação dessa estratégia, objetivou-se minimizar problemas de extrapolação e melhorar a capacidade preditiva e robustez dos modelos PLS e MLR-APS.

1.2.2. Objetivos específicos

- Desenvolver um algoritmo de validação híbrida que minimize problemas com extrapolação dos modelos e melhore assim sua eficiência e robustez.
- Implementar e depurar o código-fonte, em ambiente computacional do software Matlab, do programa do algoritmo desenvolvido para a estratégia proposta;
- Estudar o comportamento do RMSEP e RMSECV ao se adicionar as amostras fixadas na modelagem PLS;
- Avaliar o comportamento do RMSEP e RMSECV ao se adicionar as amostras fixadas na modelagem MLR-APS;
- Melhorar a robustez dos modelos MLR-APS, determinando o número ótimo de variáveis selecionadas em calibração MLR-APS para os modelos com e sem adição de amostras fixadas pelo SPXY;
- Aplicar a estratégia proposta para o melhoramento da capacidade preditiva e robustez dos modelos PLS e MLR-APS de dados espectométricos de milho e trigo, envolvendo matrizes com diferentes níveis de complexidade;
- Comparar o desempenho entre os modelos PLS e MLR-APS, ressaltando as caracteríticas que foram melhoradas para cada tipo particular de calibração multivariada.

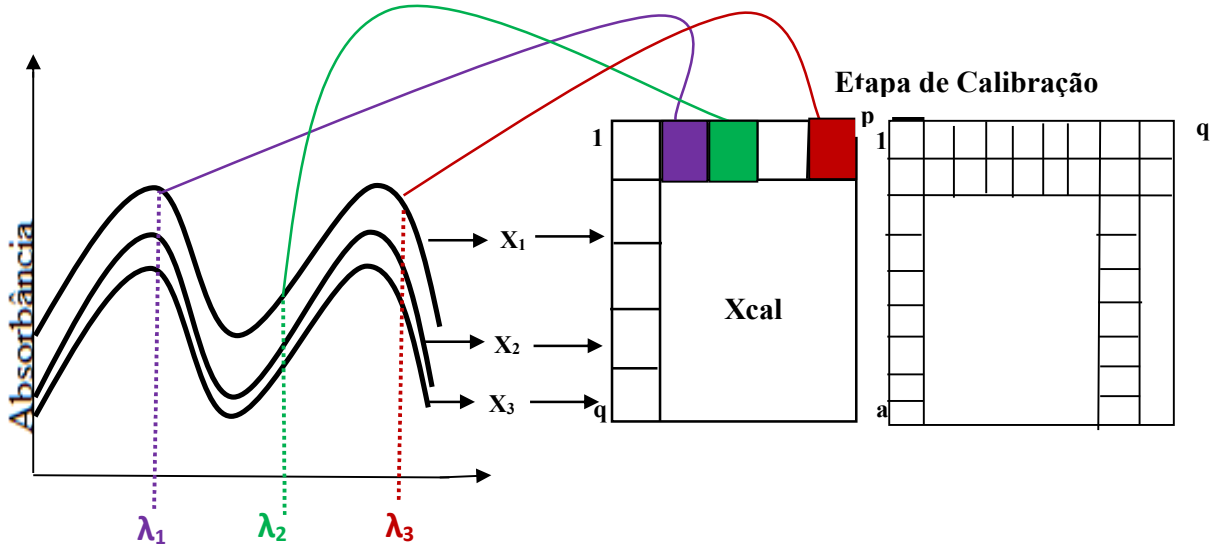
1.3Calibração Multivariada

O uso da calibração multivariada tem sido impulsionado por suas vantagens (por exemplo, calibração na presença de interferentes, redução de custos e tempo de análise) na determinação do(s) analito(s)sem a necessidade desua separação física da matriz das amostras. Nesse contexto, a obtenção de dados mutivariados em análises espectrométricas – inclusive dados de segunda ordem, a exemplo dos obtidos em espectrofluorometria[89] – tem ensejado a utilização da calibração multivariada a fim de realizar a modelagem com sucesso.

Na implementação de uma calibraçãomultivariada, várias respostas instrumentais podem ser relacionadas a uma ou mais propriedades de interesse nas amostras analisadas,empregado-se os métodos lineares e os não-lineares[23].Uma relação matemática (linear ou não) é estabelecida entre os resultados fornecidos por um método de referência – representados pela matriz **Y**– e os dados da matriz **X**obtidos nos espectros das amostras de calibração[57]. Nesse processo, a calibração possibilita a determinação de todas as propriedades modeladas,desde de que as amostras com os interferentes sejam incluídas no conjunto de calibração[12].

Para implementar uma calibração multivariada – independentemente da técnica matemática (MLR ou PLS, por exemplo) utilizada –, faz-se necessário primeiro estruturar (organizar) os dados de acordo com o esquema ilustrado na **Figura 1.1**.Os dados (espectros, por exemplo) obtidos para as amostras do conjunto de calibração comporão a matriz**X**; os sinais correspondentes às variáveis (comprimentos de onda) para cada amostra são colocados em uma linha dessa matriz. Na matriz Y, por outro lado, são lançados em uma linha o valor da(s) propriedade(s) de interesse para cada amostra. Assim, o papel da técnica de calibração usada é estabelecer um relação matemática adequada (em geral, linear) entre os dados de ambas as matrizes – diretamente no domínio original (como é o caso da MLR) ou no domínio dos dados transformados (a exemplo do PLS).

Figura 1.1. Representação Matriz de dados Instrumentais com p representando as variáveis da matriz Xcal e da Matriz de dados Padrão ou método de referênciacom q representando as variáveis da matriz Ycal.



1.3.1 Análise de componentes principais

A análise de componentes pricipais (Principal Component Analysis-PCA) é uma técnica multivariada de compressão dos dados quepermite reduzir a dimensionalidade do espaçosem a perda de informação útil. A transformação matemáticaleva a um novo sistema de eixos ortogonais – denominados de componentes principais (PCs) –, nos quais as coordenadas das amostras são os "escores" e os pesos das variáveis originais são chamados de "loadings". Assim, essa técnica permite remover as informações redundantes dos dados o que permite contornaros problemas de multicolinearidade entre as variáveis originais das respostas instrumentais[2,22,27,29].

Nos espectros para um agrupamento de *m* amostras envolvendo *n* constituintes é registrado em k comprimentos de onda, podendo assim ser descrita na forma de matriz de dados espectrais [X(m,k)][2,28]. Essas matrizes são desmembradas em uma soma dos componentes principais (PCs) como pode ser observado na **Equação 1**.

$$\mathbf{X} = \mathbf{t_1 p_1^t} + \mathbf{t_2 p_2^t} + \mathbf{t_3 p_3^t} \ldots . + \mathbf{t_n p_n^t} + \mathbf{E} \qquad (1)$$

Em termos matriciais, a Equação 1 toma a forma$\mathbf{X= T\ P^T+ E}$, sendo T uma matriz com t colunas e P^T com $\mathbf{p^t}$ linhas. Assim sendo, os vetores $\mathbf{t_n}$ que são conhecidos como scores e indicam as relações existentes entre as amostras. A maneira como as variáveis estão relacionadas entre si, são representadas pelos vetores $\mathbf{p_n}$ que são conhecidos como loadings. O **E**é a matriz de resíduos, contendo informações não modeladas na primeira PC.

Em PCA, os autovalores e autovetores são obtidos da matriz de covariância **X**que apresenta a **Equação 2**.

$$\mathrm{Cov}(\mathbf{X}) = \frac{\mathbf{X^t X}}{\mathrm{N}-1} \qquad \mathbf{(2)}$$

onde a covariância entre cada par de eixo é zero, mas os eixos principais não são correlacionados. Sendo N o número de amostras e $\mathbf{X^T X}$ é a matriz de covariância.

As linhas da matriz **P** são os autovalores de covariância em cada $\mathbf{p_i}$, tendo este relação de correspondência com uma parcela da variância, que é correlacionada com cada autovetor. Os autovetores tem autovalores relacionados a eles, e o autovetor de início é aquele que aponta para a maior diferença dos dados projetados no espaço de representação, o mesmo ocorre com todos os demais autovetores associados a demais autovalores[66].

Para realizar o cálculo dos vetores dos "loadings", é feito o produto da matriz de covariância, para os vetores $|\mathbf{p_i}$ da seguinte maneira.

$$\mathbf{Cov(X)p_i = \sigma_i p_i} \qquad \mathbf{(3)}$$

No qual, $\boldsymbol{\sigma_i}$ é o i-ésimo autovalor correlacionado com o i-ésimo autovetor **pi**. Em PCA os autovetores devem explicar a maior variabilidade possível dos dados de forma a minimizar os resíduos em cada etapa, pois cada fator é responsável pela variância dos dados[22]**.**

A maneira como as amostras estão distribuídas em um eixo diferente do original é denominado scores, que pode ser representado pela **Equação 4**.

$$\mathbf{Xp_1 = t_1} \qquad \mathbf{(4)}$$

onde**X**é a matriz de respostas inicial e $\mathbf{t_1}$é a projeção de **X**sob o autovetor inicial de $\mathbf{P_1}$.

1.3.2. Regressão por componentes principais

O PCR é baseado na modelagem de fatores, uma vez que o fator pode ser entendido como uma combinação linear das variáveis originais possibilitando assim a redução da dimensionalidade dos dados e consequentemente a construção de gráficos que possam facilitar a visualização de agrupamento de amostras e de variáveis, permitindo assim diferenciar dados de amostras com alta influência[25,29]**.** O modelo de regressão PCR pode ser compreendido por meio das **Equações5** e **6**.

$$\mathbf{X = TP^t + E} \tag{5}$$

$$\mathbf{y = T^t b + fX^t X} \tag{6}$$

Os pesos (loadings), presentes na matriz **P,**são encontrados ao escolher os primeiros autovetores de $\boldsymbol{X^t X}$**,** no qual**X**é a matriz de dados instrumentais centrados na média. A matriz **T,** que representa as pontuações é obtida projetando a matriz **X** no espaço da matriz transposta de **P**[56]**.** A matriz de $\mathbf{T^t}$é utilizada como regressor junto com o vetor **y**na equação do cálculo da distância de mahalanobis[43] como pode ser visto na **Equação 7**.

$$D = (a_i - a_j)^t (\mathrm{cov}(x)^{-1} (a_i - a_j)^{1/2} \tag{7}$$

A principal característica de um modelo PCR é o fato do mesmo considerar apenas as respostas obtidas pelo instrumento (**X**), sem considerar as informações relevantes da concentração (**y**), acarretando numa fragilidade do modelo, pois em dados onde a propriedade de interesse apresente um sinal muito pequeno e não causa forte influência nas primeiras PCs, podendo assim acarretar numa escolha de um maior de PCs para se construir o modelo[68].

1.3.3. Regressão por mínimos quadrados parciais(PLS)

Diferentemente do modelo **PCR**que não leva em consideração as estatísticas dos parâmetros de interesse (**Y)** o PLS estabelece uma relação linear entre as matrizes de scores dos dados de **X** e **Y**[22].

A abordagem PLS (Partial Least Squares) possibilita a construção de uma boa modelagem mesmo na presença de ruídos experimentais, colinearidade e não-linearidade dos dados[1]. Em uma matriz **X** com **G** colunas, que correspondem ao sinal do espectro, para cada

comprimento de onda, e **j** linhas representando cada amostra. Na matriz **Y,** tem-se **H** colunas, nas quais apresentam os parâmetros de interesse (densidade), e **n** linhas referentes a cada amostra[65].Em modelagem PLS as variáveis latentes sofrem pequenas rotações para que os scores **t**e **u** apresentem o melhor comportamento possível e que assim possam descrever as relações existentes entre **X** e **Y**[29, 55]**.**

As matrizes **X** e **Y**são divididas em uma soma de matrizes menores chamadas componentes principais ou variáveis latentes que são responsáveis por mostrar o grau de complexidade do modelo[2,67,69]**.** A decomposição dessas matrizes, propicia uma soma dessas matrizes menores geradas, conhecidas como componentes principais (PCs), podendo ser compreendida pelas **Equações 8** e **9**.

$$\mathbf{X = TP^t + E = t_1P_1^t + t_2P_2^t + t_3P_3^t \dots \dots t_nP_n^t + E} \qquad \mathbf{(8)}$$

$$\mathbf{Y = U + Q^t + F = u_1q_1^t + u_2q_2^t + u_3q_3^t \dots \dots u_nq_n^t + F} \qquad \mathbf{(9)}$$

onde as matrizes **T** e **U** são as respectivas matrizes de scores de **X** e **Y** e os loadings são representados pelas matrizes $\mathbf{P^t}$e $\mathbf{Q^t}$, a parte dos dados não modelados vai estar presente nas matrizes de resíduos **E** e **F.**

A modelagem **PLS**procura maximizar a covariância entre as variáveis das matrizes de dados reais e instrumentais[26], como pode ser visto pela **Equação 10**.

$$\mathrm{Cov}(\mathbf{\hat{Y}q_a, \hat{X}r_a}) = \mathbf{q}_a^{\mathrm{T}} \frac{\mathbf{\hat{Y}^T X}}{\mathbf{n-1}} \mathbf{r}a = \mathbf{q_a^t S_{yx}} \qquad \mathbf{(10)}$$

Após a decomposição das matrizes de dados **X** e **Y** respectivamente, as novas matrizes obtidas (scores **T** e **U**), são relacionadas por meio de um modelo linear como pode ser visualizado na Equação 11.

$$\mathbf{U = b^* T} \qquad \mathbf{(11)}$$

onde**b** é o vetor do coeficiente de regressão que estabelece a melhor relação entre os dados previstos pelo modelo e os dados obtidos por um método de referência e assim obter um modelo bem ajustado. Para uma determinação apropriada de **b,** faz-se necessário a seleção do

número adequado de VLs e a detecção e exclusão de amostras com outliers(amostras anômalas)[24].

Uma das maneiras de se avaliar a performance preditiva de um modelo PLS é usando a soma dos quadrados dos erros de predição (PRESS)[2–4], que é expresso **Equação 12**.

$$\mathbf{PRESS} = \sum_{i=1}^{m} \sum_{i=1}^{n} (C_{ij} - \hat{C}_{ij}) \quad \mathbf{(12)}$$

onde *m* são as amostras e *n* os parâmetros de interesse, C_{ij}são as concentrações obtidas por método padrão (amostras reais) e $\hat{C}_{ij}$são as previstas pelo modelo de calibração.

1.3.3.1(PLS group-out)

Na estrategia PLS-grou-out grupos de amostras são deixadas de fora durante o processo de validação cruzada, reduzindo assim o número de vezes que se recalcula o modelo[90]. Os n-grupos de amostras de calibração são utlizados para prever n-grupos de validação[90], o procedimento continua até que os grupos de amostras de calibração sejam incluidas no grupo das amostras de validação.

1.3.4 Regressão linear múltipla(MLR)

A modelagem MLR é bastante utilizada em problemas envolvendo a calibração multivariada, pois ela estabelece uma relação simples entre as variáveis do conjunto de dados de partida[75]. Nesta modelagem tem-se uma matriz **X** e um vetor **y**, onde cada variável**y** é descrita como sendo uma combinação linear das variáveis da matriz **X**[75]. A combinação linear descrita pode ser observada na **Equação 13**.

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \cdots \beta_k \mathbf{x}_k + \varepsilon \quad \mathbf{(13)}$$

na qual $\beta_0, \beta_1 e \beta_k$ são os valores obtidos pela regreção e ε é a parte do erro não modelado.

Os coeficiente $\beta_0, \beta_1 e \beta_k$podem ser obtidos de forma aproximada atraves da pseudo inversa como veremos na **Equação 14**.

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad \mathbf{(14)}$$

Os coefiecientes após a sua obtenção, são utlizados para encontrar as variáveis estimadas no pelo como pode ser visto na **Equação 15**.

$$\hat{y} = X\beta \qquad (15)$$

onde **X** e **β** são as matrizes de, respectivamente,variáveis independentes e coefiecientes de regressão.

A regressão linear multípla(MLR) atua no espaço original dos dados e por isso fornecem modelos mais inteligíveis e de facíl interpretação[76].Para uma modelagem MLRadequada,se faz necessário a escolha adequada das variáveis espectraisde modo a evitar a escolha de variáveis muito colineares no conjunto de dados que prejudiquem a modelagem[76,77].

1.3.5 Algoritmo das projeções sucessivas

Quando se tem um grande número de variáveis, algumas podem apresentar informação redundante entre si e, consequentemente, adiciona informação de pouca utilidade para a modelagem[76]. Em matrizes de dados com alta dimensão, a calibração MLR se torna inviável, em virtude da presença de variáveis colineares no conjunto de dados, sendo necessário a utilização um método de seleção de variáveis para obtenção de modelos mais robustos[76,78].

Ao longo dos anos, muitos algoritmos para seleção de variáveis foram desenvolvidos, dentre os quais cabe destacar o Algoritmo de Projeções Sucessivas[79], que busca selecionar as variáveis com menor correlação possível com as variáveis preliminarmente selecionadas. O APS utiliza manipulações simples em espaços vetoriais que diminuem a possibilidade de inserção de variáveis com informação irrelevante para a modelagem[77,78,79].

O APS funciona da seguinte maneira:

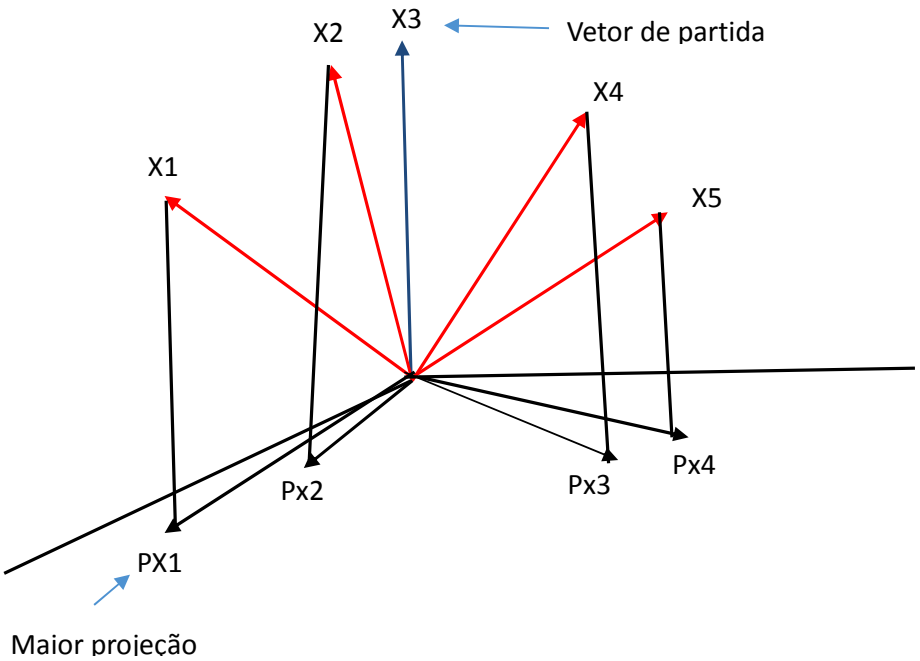
Etapa 1 *-Na primeira etapa, a matriz de dados original é utilizada para obter-se uma seríe de variáveis com a menor correlação possível;*

Etapa 2 *-Durante a segunda etapa, as cadeias de variáveis são testadas de modo a avaliar sua correlação com a propriedade de interesse em termos de menor RMSEP na modelagem MLR;*

Etapa 3 *-A terceira etapa consiste na retirada de variáveis, que não proporcionam melhora significativa do erro do RMSEP. Essa retirada é feita com base em testes estatísticos de significância tal como o teste F.*

As projeções usadas na abordagem APS podem ser melhor compreendidas com o auxílio da na **Figura 1.2** para o caso em que J=5 e M_{cal}=3, onde J é o número de variáveis de partida e M_{cal} é o número de amostras de calibração. Os valores de k representam os vetores selecionados para o cálculo das projeções ortogonais ao valor correspondnete de k.Um vetor de partida(X_3) é selecionado, e os demais vetores tem suas projeções calculadas em um espaço ortogonal ao vetor de partida. Os vetores com as maiores projeções ortogonais ao vetor de partida são selecionadas.

Figura 1.2. Representação das projeções APS para J=5, Mcal =3 e k(0)=3. Primeira interação: k(1)=1.



Na **Figura 1.2,** pode-se observar como as variáveis são escolhidas pelo algoritmo APS levando em consideração suas projeções espaciais. Inicialmente uma variável **X3**é selecionada e as demais variáveis tem suas projeções calculadas.As variáveis selecionadas posteriormente,são aquelas que apresentam a maior ortogonalidade em relação a variável de partida, fato este que diminui a correlação entre estas variáveis[7678].

O conjunto de variáveis que apresentar o menor erro médio quadratico de previsão(RMSEP) na modelagem MLR serão selecionadas[76,77].

O RMSEP para essa modelagem pode ser calculado de acordo com a **Equação 16**.

$$\mathrm{RMSEP} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_m - y_{ref(m)})^2} \quad \textbf{(16)}$$

onde M denota as amostras de calibração, y_m as amostras preditas pelo modelo e $y_{ref(m)}$ contém valores obtidos por metodo de referência.

1.3.6. Métodos de validação

O uso de algoritmos que possibilitem dividir o conjunto de dados original em subconjuntos independentes e que as amostras selecionadas sejam as mais representativas possíveis se tornou cada vez mais comum em calibração multivariada, tendo em vista que uma representação inadequada dos dados pode gerar modelos com precária capacidade preditiva. Muitos métodos de validação são usados atualmente, tais como: Validação por série de teste,Validação bootstrap e Validação cruzada[70].Cabe salientar nesses tipos de validação a forma como as amostras para construção do modelo são selecionadas. Na validação por serie de teste as amostras são selecionadas via algoritmos baseados em distância euclidiana, tais como (KS, SPXY, etc)[60]. A validação bootstrap utiliza o método de reamostragem com reposição. Para a validação cruzada as amostras são selecionadas de maneira aleatoria.

1.3.6.1 Validação por série de teste

Na validação de um modelo de calibração por série de teste, a divisão dos conjuntos de calibração e validação é realizada, utilizando-se algoritmos que selecionam as amostras de maior representatividade no conjunto amostral[71], a exemplo do KS e SPXY descritos a seguir.

1.3.6.2 Algoritmo KS

O algoritmo Kennard-Stone (KS)[8] utiliza adistância Euclidiana, $d_x(p,q)$, entre dois pontos “p eq”, correspondentes a 2 (duas) amostras no espaço multidimensionaldas variáveis, com o objetivo de determinar as amostras que seencontram mais distâncias entre si. A **Figura 1.3**ilustracomo essa distância évisualizada, geometricamente, para o caso particular de três variáveis(Var 1: x_1, Var 2: x_2 e Var 3: x_3), cujo espaço correspondente é o tridimensional. Nesse caso, a distância $d_x(p,q)$ é dada pela **Equação 17**.

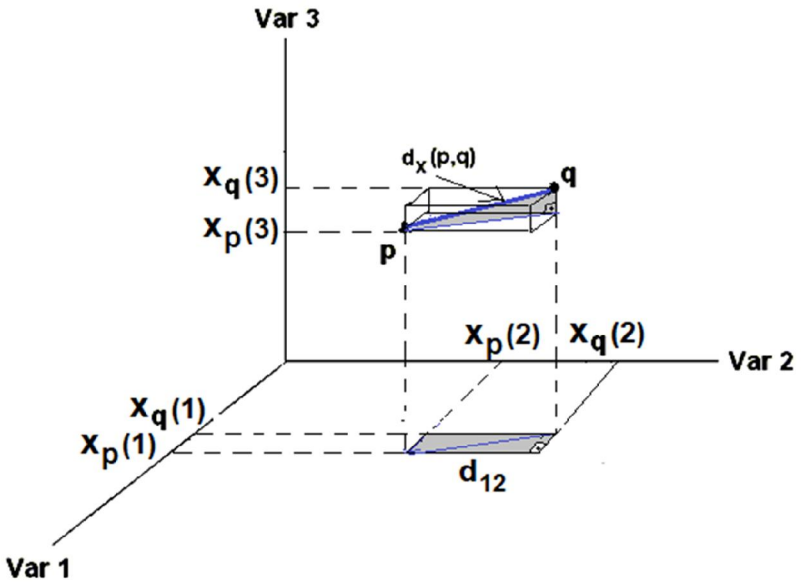
$$d_x(p,q) = \sqrt{[x_p(1) - x_q(1)]^2 + [x_p(2) - x_q(2)]^2 + [x_p(3) - x_q(3)]^2} \quad \textbf{(17)}$$

Para um espaço multidimensional com *J* dimensões, a distânica Euclidiana $d_x(p,q)$ é dada pela **Equação 18**.

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}\left[x_{p(j)} - x_{q(j)}\right]^2} \qquad p \text{ e } q \in [1;N] \qquad \textbf{(18)}$$

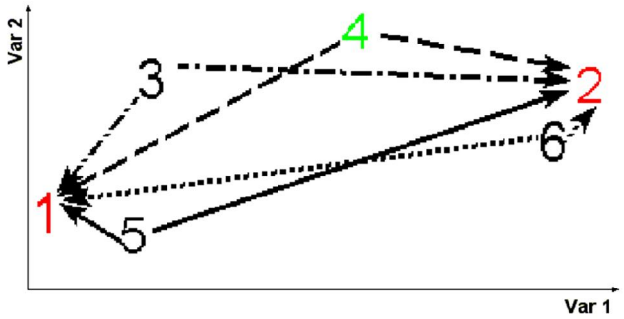
onde $x_{p(j)}$ e $x_{q(j)}$são as respostas instrumentais (sinais nos espectros) para amostras *p* e *q*nos *j*-ésimos comprimentos de onda dos espectros.

Figura 1.3. Visão geométrica da distância Euclidiana entre dois pontos p (x_{p1}, x_{p2}, x_{p3}) e q (x_{q1}, x_{q2}, x_{q3}).



Na **Figura 1.4**, ilustra-se o princípio de funcionamento do algoritmo Kennard-Stone considerando o caso de um conjunto de seis amostras que se encontram localizadas no espaço bidimensional de duas variáveis **Var 1** e **Var 2**.Primeiramente, o KS determina o par de amostras que tenham a maior distância entre sid$_x$(p,q) encontradausando a**Equação 18**para*J* = 2. Nesse exemplo,as amostras **1** e **2** são as mais distantes entre si.Em seguida, o algoritmo determina – entre as amostras remanescentes – as que se encontram mais próximas de **1** ou **2** e seleciona a mais distante. Nesse caso, a amostra **4** é selecionada nessepasso. Esse processo é repetido até que um certo número Nde amostras, usado como critério de parada, seja alcançado. Quando N = 3, o KS seleciona as amostras selecionadas: **1**, **2** e 4; para N = 4, as amostras **1**, **2**, **4** e **3** são selecionadas e assim por diante.

Figura 1.4.Ilustração da aplicação KS a um conjunto de seis amostras no espaço bidimensional das variáveis Var 1 e Var 2.



Como resultado da aplicação do KS, as primeiras amostras selecionadas – cujo número N é estabelecido *a priori* – são usadas para compor o subconjunto de calibração e as restantes do conjunto total comporão o subconjunto de validação. Com isso, assegura-se que as amostras de calibração serão sempre as mais externas e cobrirão a fronteira do espaço, conforme ilustrado na **Figura 1.5**.

Figura 1.5. Espaços cobertos pelas amostras (calibração e validação) selecionadas pelo algoritmo KS.

O algoritmo apresenta algumas limitações, por exemplo, o fato de não levar em consideração a estatística da matriz **Y**que são os parâmetros de interesse. Assim, o KS não leva em consideração a variabilidade decorrente de mudanças de composição na matriz[47] que se refletem nos parâmetros de interesse. Para levar em contaesse fato – assim como a estatística de **Y** no processo de busca das amostras mais distantes entre si –, desenvolveu-se o algoritmo SPXY descrito na **Seção1.3.6.3**.

1.3.6.3 Algoritmo SPXY

O SPXY é uma variante do algoritmo KS, mas difere no que diz respeito a inclusãoda estatística de **Y** no seuprincípio de funcionamento**.**As distâncias $d_x(p,q)$ e $d_y(p,q)$, são medidas levando-se em consideração as relações existentes em **X** e em **Y** no cálculo da distância inter-amostras.Assim,

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max p,q \in [1,N]\, d_x(p,q)} + \frac{d_y(p,q)}{\max\ p,q \in [1,N]\, d_y(p,q)} \; ; p,q \in [1,N] \qquad \textbf{(19)}$$

Visando garantir que a distribuição das amostras no espaço de **X** e **Y**tenham mesma importância, fze-se necessário dividir as distâncias $d_x(p,q)$ e $d_y(p,q)$ pelos seus valores máximos no conjunto amostral obtendo-se assim uma distância d_{xy} normalizada[9].

1.3.6.4 Validação Bootstrap

O Bootstrap e uma técnica de reamostragem, sendo assim é possível fazer inferências a cerca da distribuição das observaçãos a partir da disposição dos dados experiementais[42]**.**Ao se utilizar esse método é possível estimar o seu limite de confiança, com a intenção de fazer deduções a cerca das propriedades em análise[42,44].

A reamostragem com reposição(Bootstrap) possibilita a obtenção de distintos conjuntos de amostra[45]. Esta tecníca pode ser utilizada nos mais diversos problemas, podendo ser utilizado nas mais diversas situações, pois independem da disposição inicial da estatística das propiedades em avalição[42].

1.3.6.5 Validação cruzada

Em calibração PLS (Partial Least Squares**),** é de fundamental importância escolher um número ideal de variáveis latentes, que determina complexidade do modelo[5,58]. A escolha do número inapropriado de VLs , pode acarretar uma perda significativa na robustez do modelo e consequentemente afetar a sua capacidade preditiva[6,63]. A escolha de um número maior de variáveis latentes pode aumentar o ruído e causar problemas de sobreajuste, no qual os dados de calibração são bem ajustados, mas não apresentam desempenho satisfatorio na predição de amostras desconhecidas[83]. Um número menor de variáveis latentes escolhido, pode acarretar problemas de sub-ajustes, devido a não inclusão de informação relevante do conjunto de dados, gerando modelos de calibração mal ajustados[83]. A validação cruzada é o método mais comumente usado para definir o número de componentes escolhidas para um modelo[15,16,38,39]. Na CV os dados treinamento são divididos em subconjuntos de calibração e validação[17].

1.3.6.6 Validação cruzada de exclusão única

Em modelagem PLS a escolha do número ideal de VLs é de suma importância, tendo em vista que os modelos baseados em projeções geram resultados que não condizem com os dados observados experimentalmente e por isso carecem de ajustes para evitar o overfitting[49]. A validação cruzada é o metodo mais comum quando se pretende determinar o número ideal de VLs[64].

A CVLOO (Leave-one-out Cross validation) é uma dos métodos mais utlizados de se fazer uma validação cruzada, neste metodo uma das amostras fica de fora do conjunto de calibração para depois ser predita, o procedimento é repetido até que todas as amostras tenham ficado pelo menos uma vez fora do conjunto de calibração[46,48,49]. Após todas as amostraspassarem por esta mesma estratégia, uma raiz quadrada do erro médio de validação cruzada (RMSECV) é calculado para cada uma das VL utilizadas[49], como pode ser visualizado na **Equação 19**.

$$\mathrm{RMSECV}=\frac{\sqrt{\sum(y_p-y_e)^2}}{n} \qquad \textbf{(19)}$$

na qual y_p é o valor de concentração previsto pelo modelo, y_eo resultado esperado e n é a quantidade de amostras do conjunto de calibração.

1.3.6.7 Validação cruzada de divisão representativa

Quando se constroi e valida modelos em calibração multivariada é de fundamental importância que os conjuntos de calibração e validação sejam significativos para o conjunto de dados original[5,39,54].Em validação cruzada a divisão do conjunto de dados original é normalmente feita de maneira randômica e a determinação da quantidade de VLs geralmente varia de acordo com a abordagem CV utilizada na construção do modelo[5,35,50].

O RSCV(Representative Splitting Cross Validation) é um novo método de CV no qual a divisão dos dados em calibração e validação é feita utilizando um algortimo de seleção de amostras e não mais de maneira randômica como em CVs mais convencionais[5].

Nessa nova abordagem, o algoritmo DUPLEX é usado para dividir em partes iguais o conjunto de calibração de partida e assim fazer uma combinação dos valores de uma serie de CVs K-fold[5].Para avaliar a performance da nova estrategia os CVs fundamentados na partição

dos dados DUPLEX são utilizados para determinar o RMSE associado de RSCV como observado na **Equação 20**.

$$\text{RMSERCV} = \sqrt{\frac{1}{4 \text{ x } n} \sum_{i=1}^{4} \sum_{j=1}^{n} (y_{ij} - \hat{y}_{ij})^2} \quad \textbf{(20)}$$

sendo $\hat{y}_{ij}$ e y_{ij}os resultados previstos e de referência dos *j*-ésimos objetos da validação e *i*-ésima (i = 1, 2, 3, 4) k-fold (k = 2, 4, 8, 16) CV.

1.3.6.8 Validação Híbrida

A validação híbrida é uma estrategia de validação cruzada na qual as amostras mais externas são selecionadas e as restantes são divididas de forma randômica em conjuntos de calibração e predição. O algoritmo SPXY é utilizado para selecionar as amostras mais externas, visando a garantir que toda a variabilidade do conjunto de amostras esteja dentro dos limites de observação dos dados.

À medida que as amostras selecionadas vão sendo fixadas, as fronteiras de calibração são delimitadas e problemas com extrapolação do modelo – que ocorre quando amostras são preditas além dos limites de observação do conjunto amostral – são minimizados. A escolha de amostras mais representativas para o conjunto de dados, possibilita a construção de modelos mais robustos e com melhor capacidade preditiva.

Os resultado de RMSECV para cada processo de validação cruzada, a medida que amostras vão sendo fixadas, foi utilizado como uma das metricas para avaliação das estrategia propostas como pode ser visualizado na **Equação 21**.

$$\text{RMSECV} = \frac{\sqrt{\sum (y_p - y_e)^2}}{n - a} \quad \textbf{(21)}$$

na qual "y_p"é o valor predito pelo modelo, "y_e" o valor de referência e "n" é o número de amostras do conjunto de calibração e "a" é o número de amostras fixadas.

CAPÍTULO 2

Metodologia

2. METODOLOGIA

2.1. Conjuntos de dados

No presente trabalho, foram utilizados dois conjuntos de dados NIR(Near infrared) que são de domíniopúblico e disponíveis na Internet.O primeiro conjunto (http://www.idrc-chambersburg.org/shootout.html) consiste dos dados de 107 amostras de trigo, cujos espectros NIR foram registradosna faixa 1100-2500 nm com uma resolução de 2nm[72]. A propriedade alvo do estudo compreende o teor de proteína, que varia de 9,7 a 14,4% (w/w).A determinação do conteúdo de proteina em amostras de trigo é de fundamental importância para a indústria de produção de farinha de trigo. As proteínas presentes nesses grãos apresentam uma grande capacidade de formar glútem, que durante o processo de fermentação em pães e massas é o responsável por reter as moleculas de dioxido de carbono e favorecer o crescimento da massa.

No segundo conjunto (http://software.eigenvector.com/Data/Corn/index.htm), têm-se os dados de 80 amostras de milho, no qual seus espectros NIR foram adquiridos num intervalo de 1100-2498nm com intervalos de 2nm. O parâmetro de interesse empregados nesse estudo foi umidade. O teor de umidade em amostras de milho é um parâmetro fundamental para determinar as melhores condições de armazenagem do grão. Esse teor deve ser bem definido, pois o mesmo vai garantir a qualidade do grão durante a estocagem.

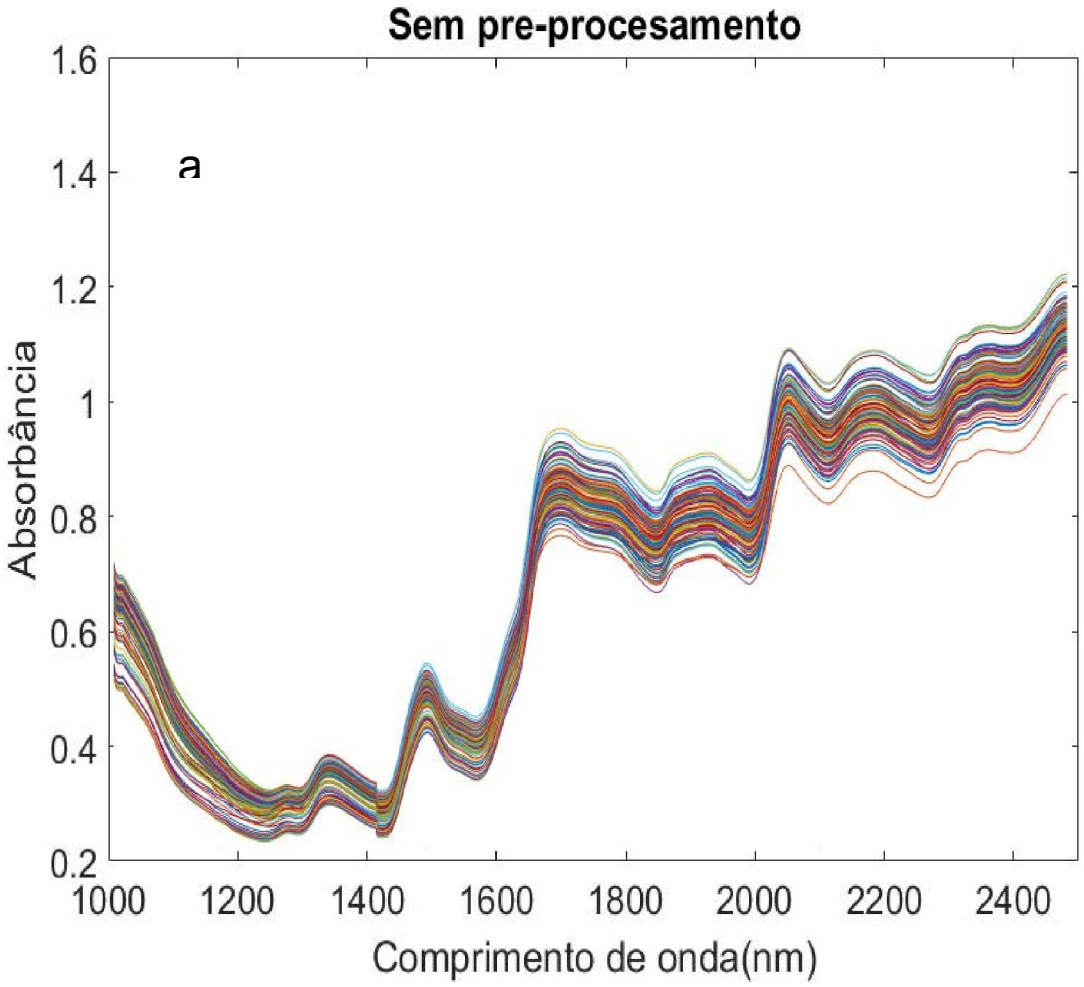
2.2. Pré-processamento

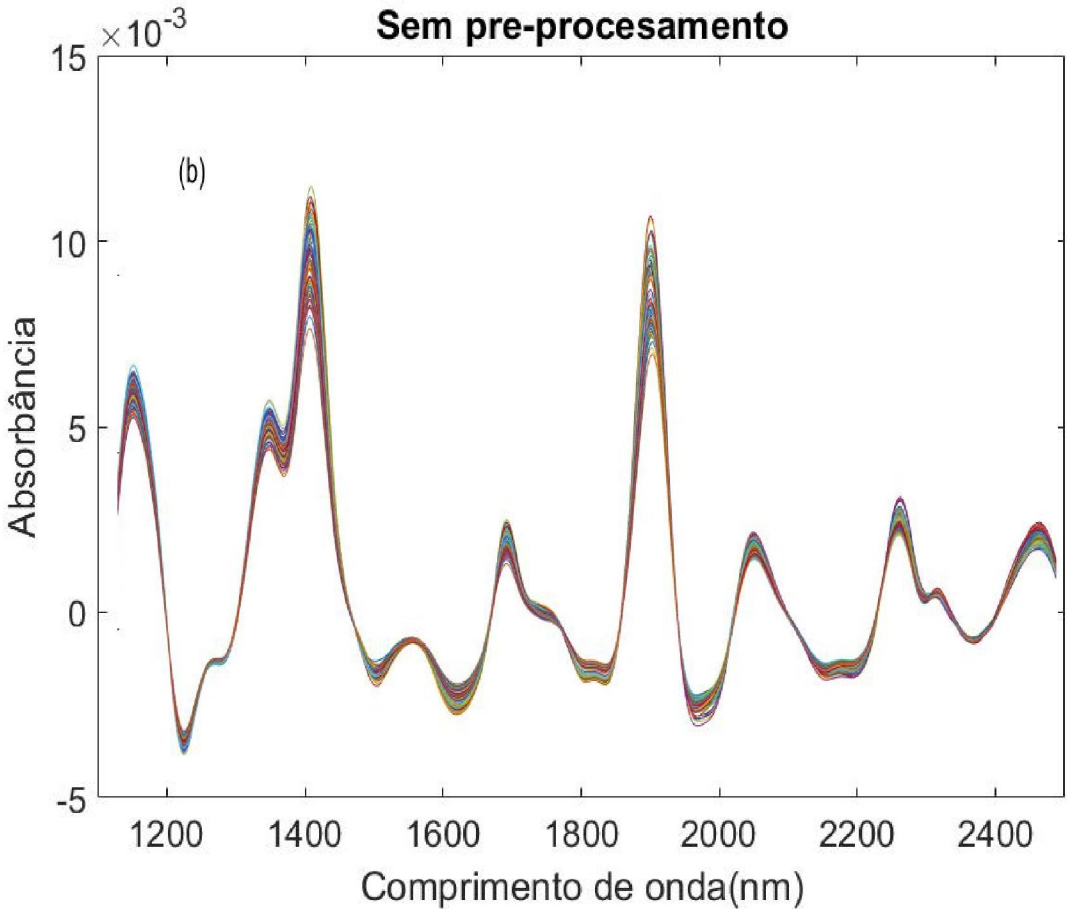
2.2.1 Dados NIR de trigo

Na analíse da **Figura 2.1**para os dados da matriz "trigo", pode-se observar que a região de 1000-1100não apresenta informação util para a modelagem(sem resposta na região para a propriedade de interesse), portanto essa região que compreende um total de 50 variáveis foram retiradas, restanto 700 variáveis para a construção do modelo. A **Figura 2.1a** os espectros brutos das 107 amostras de trigo são apresentados e pode-se observar que além de regiões não informativas, tambem observa-se pequenas mudanças na linha de base(variações sistemáticas que não apresentam relação com a propriedade em investigação, mas que podem afetar o modelo caso não sejam retiradas[73]).Em virtude do óbice apresentado, preferiu-se por utilizar os espectros derivados gerados

após a aplicação do filtro Savitsky-Golay com polinômio de 2ªordem e janela de 21 pontos[74], restando 680 variáveis para cada amostra como pode ser visualizado na **Figura 2.1b**, que serviu como base para os calculos realizados. A janela de 17 pontos[73],foi selecionada em um trabalho anterior como sendo o número idealde pontos para o conjunto de dados em análise, mas para a estratégia proposta a janelacom21 pontos foi a escolhida, pois apresenta espectros menos ruidosos e por proporcionar melhores resultados RMSECV e RMSEP durante a construção do modelo de calibração.

Figura 2.1. Espectro de 107 amostras de trigo. (a) Sem pré-processamento. (b) Derivados e Suavizados Savitsk Golay com janela de 21 pontos





O conjunto de dados foi dividido da seguinte maneira: 12 amostras foram retiradas com o auxílio do algoritmo SPXY[9](amostras que permanecerão fixas e serão adicionadas durante a etapa de calibração) em PLS e 16 para abordagem MLR-APS, e as amostras remanescentes foram divididas de forma randômica em calibração 75% (71 amostras) e predição 25% (24 amostras).

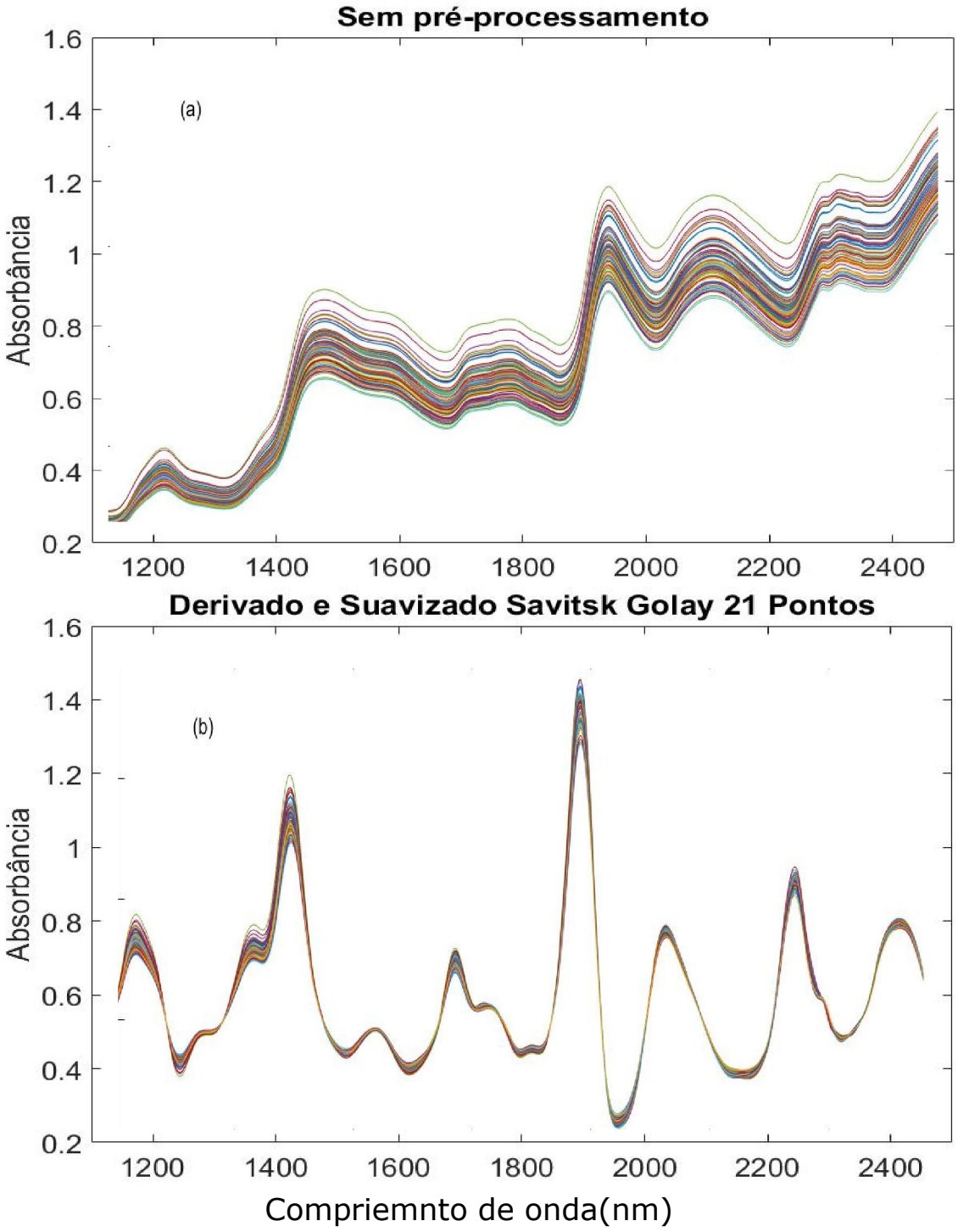
Após a divisão do conjunto de dados, implementa-sea PCA com o intuito de verificar se o SPXY selecionou as amostras de fronteira e assim minimizar problemas de extrapolação do modelo durante a etapa de calibração.

2.2.2 Dados NIR de milho

Os espectros de 80 amostras da matriz“milho” foram obtidos na faixa **1.100 – 2.498 nm** eencontram-se apresentados na **Figura 2.2**. Observa-se na **Figura 2.2a** que apresentam variação de sua linha de base,assim é necessario o pré-processamento desses dados. Diferentemente do conjunto de dados anterior, não foi necessario retirar regiões dos espectros registrados. Tendo em vista o obice encontrado, deu-se preferência a

utilização dos espectros derivados e após a derivação e suavização com filtro Savitsky-Golay e emprego do polinômio de 2ªordem com janela de 21 pontos como descrito em outros trabalhos[60,62],restaram 680 variáveis para cada amostra como pode ser observado na **Figura 2.2b**.

Figura 2.2. Espectro das 80 amostras de milho. (a) Sem pré-processamento. (b) Derivados e Suavizados Savitsk Golay com janela de 21 pontos.



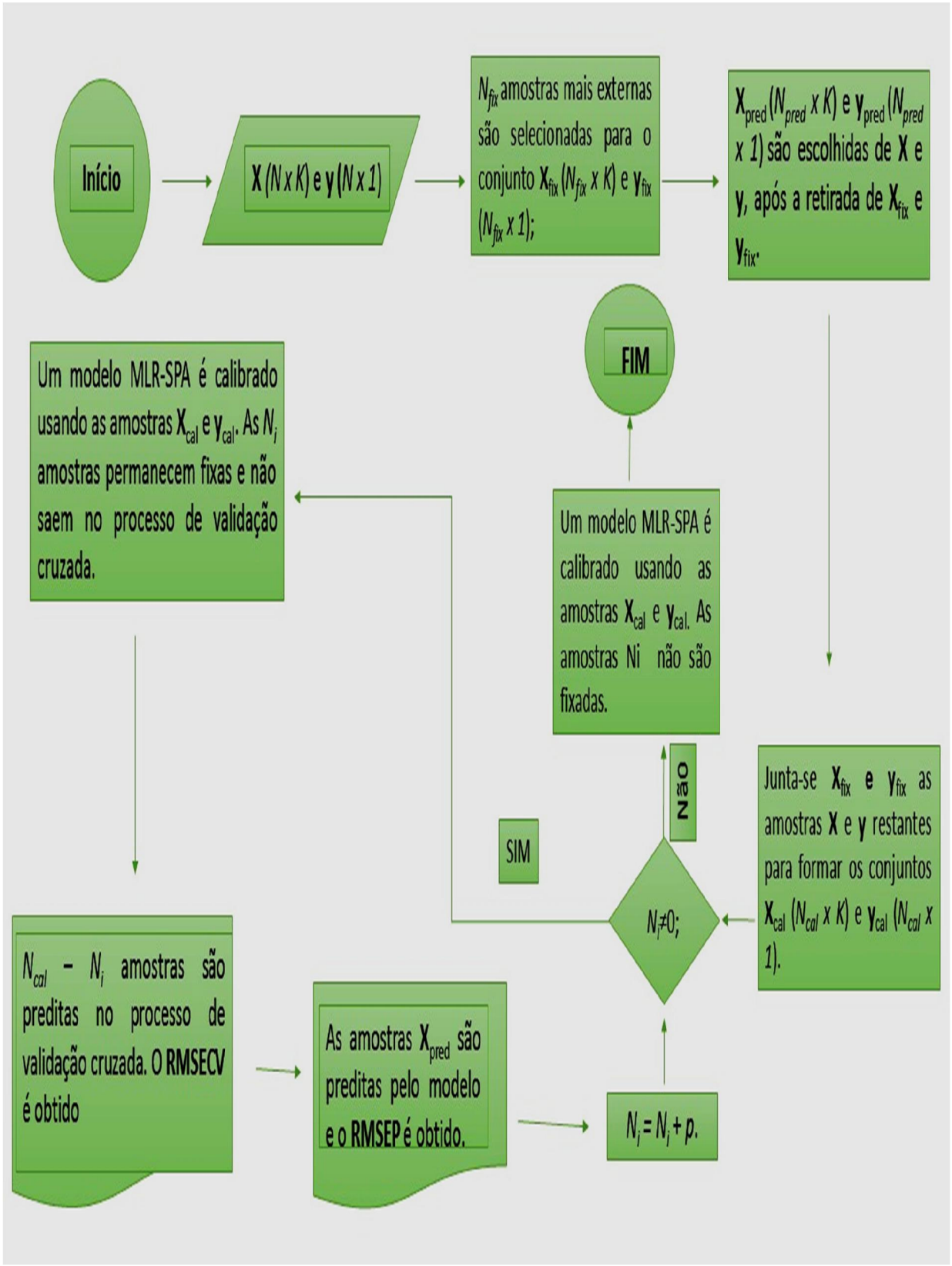
Quanto a divisão do conjunto de dados, foi adotado o mesmo procedimento dos dados anteriores sendo para calibração(51 amostras) e para predição(17 amostras) e 12 amostras selecionadas pelo SPXY.

Assim, como no conjunto de dados anterior, uma analise PCA foi realizada e foram destacadas as amostras que permanecerão fixas durante a modelagem e assim verificar se as mais externas foram realmente selecionadas.

2.3.Algoritmo de Validação Cruzada Híbrida

O algoritmo CVH foi desenvolvido com o intuito de contornar os problemas de extrapolação de modelos de calibração multivariada e, assim, melhorar a capacidade preditiva da modelagem PLS e MLR-APS.Na busca de uma melhor compreensão a cerca do algoritmo de validação cruzada híbrida, um fluxograma com as principais etapas do algoritmo foi construido,como pode ser visualizado a seguir.

Figura 2.3. **Fluxograma do Algortimo de validação cruzada híbrida.**



Inicialmente têm-se um conjunto de dados, no qual são apresentados em forma de matrizes **X** e **Y,** que são respectivamente a matriz de resposta intrumental e matriz de referência. Na etapa seguinte as amostras mais externas presentes nas matrizes **X** e **Y**são selecionadas via algoritmo SPXY e geram os conjuntos de dados$\mathbf{X_{fix}}$e $\mathbf{Y_{fix}}$**.** Tendo estas amostras mais externas sido selecionadas, as amostras de predição$\mathbf{X_{pred}}$ e $\mathbf{Y_{pred}}$são selecionadas levando-se em conta a distribuição homogênea dessas amostras. Em seguida as amostras $\mathbf{X_{fix}}$e $\mathbf{y_{fix}}$são adiconadas nas amostras restante **X** e **Y**, gerando assim os conjuntos de calibração $\mathbf{X_{cal}}$ e $\mathbf{Y_{cal}}$. Quando o número de amotras fixadas N_i é igual a zero, um modelo MLR-APS é calibrado com todas as amostras de $\mathbf{X_{cal}}$ e $\mathbf{Y_{cal}}$ e as amostras $\mathbf{N_i}$ não são fixadas. À medida que as amostras vão sendo fixadas, ou seja, $\mathbf{N_i}\neq0$, um modelo de calibração MLR-APS é obtido com as amostras $\mathbf{X_{cal}}$ e $\mathbf{Y_{cal}}$, sem a presença das amostras $\mathbf{N_i}$ que foram fixadas. Durante a etapa de validação cruzada as $\mathbf{N_i}$ amostras fixadas não são seleciondas, pois estas apenas delimitam as fronteiras do conjunto amostral. O RMSECV é obtido a partir da predição das amostras do conjunto de calibração, somente $\mathbf{N_{cal}-N_i}$ são preditas durante este processo. Após o modelo calibrado as amostras $\mathbf{X_{pred}}$ são preditas e os valores de RMSEP obtidos. Diversos modelos são gerados, levando-se em consideração a seguinte equação $\mathbf{N_i = N_i + P}$, onde P é o incremento usado para escolher as amostras que serão fixadas, esses processo se repete até que o número de incrimentos(Número de amostras fixadas por vez durante o procedimento de validação cruzada) seja igual ao número de amostras mais externas $\mathbf{X_{fix}}$ selecionadas previamente.

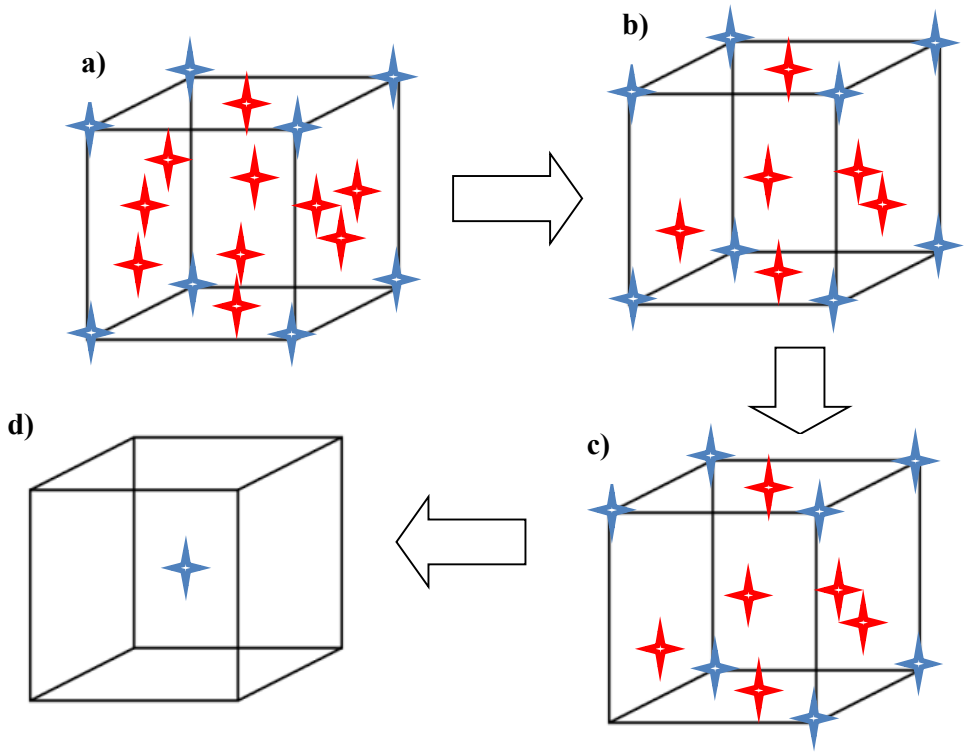
2.4. Funcionamento do Algoritmo de validação híbrida

Inicialmente as amostras mais externas(**azul**) são selecionadas pelo algoritmo SPXY, como pode ser visualizado na **Figura 2.4a**.As amostras selecionadas, representam um percentual de 15% do total das amostras existentes e delimitam as fronteiras amostrais dos dados.Após esse procedimento as amostras remanecentes(**vermelho**) são divididas de forma randômica em amostras de calibração e predição. Na **Figura 2.4b,**as amostras do conjunto de calibração(**vermelho**) e as amostras selecionadas previamente pelo algoritmo SPXY(**azul**) são unidas em um único conjunto de dados.

O conjunto de calibração incial(**vermelho** e azul) é dividido em um subconjunto de calibração, como pode ser observado na **Figura 2.4c**. Esse subconjunto de calibração é utilizado na predição da amostra do subconjunto de validação

. A amostra que ficou de fora no procedimento CVLOO, como pode ser vista na **Figura 2.4d**, é predita com base nas amostras do subconjunto de calibração. A amostra que ficou de fora do procimento de validação cruzada é mais externa.Na fase inicial do algoritmo todas as amostras de calibração(amostras restante e amostras selecionadas previamente pelo SPXY),são incluidas no procedimento de validação cruzada de exclusão e consequentemente preditas na etapa de calibração. O fato de amostras selecionadas previamente, serem usadas na etapa de validação cruzada, não impede que ocorram problemas com extrapolação, pois ao serem preditas as fronteiras amostrais perde-se a delimitação da mesma.

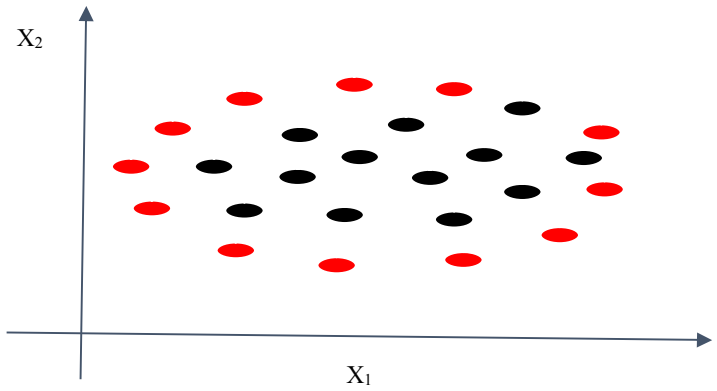
Figura 2.4.Seleção de amostras fixadaspelo SPXY e procedimento CVLOO - (a) Seleção das amostras mais externas. (b) Amostras de calibração e amostras selecionadas pelo SPXY (c) Amostras do subconjunto de calibração d)Amostra deixada de fora no processo CVLOO.



Na etapa seguinte do algoritmo, as amostras que foram selecionadas inicialmente pelo algoritmo SPXY, são divididas em grupos com as mesmas quantidades de amostras que serão fixadas por vez. A medida que cada grupo de amostras é fixado, um

procedimento CVLOO é realizado,esse procedimento é executado até que todos os grupos de amostras tenham sido fixadas. A cada grupo de amostras fixadas na etapa de calibração, tem-se uma restrição no número de amostras a serem incluidas no processo CVLOO,ou seja, o grupo de amostras fixadas,não podem ser incluídas na etapa de validação cruzada. As amostras fixadas e a delimitação das amostras a serem utilizadas no processo de validação cruzada, podem ser compreendidas de maneira mais simples na **Figura 2.5.**

Figura 2.5. Amostras selecionadas pelo SPXY–Vermelho(amostras mais externas selecionadas previamente)**Preto**(amostras que serão selecionadas no procedimento de validaçao cruzada).



As amostras mais externas que foram selecionadas previamente,passam a ser fixadas (**vermelho**) na etapa de calibração. As amostras mais internas (**preto),** são as amostras a serem incluidas na validação cruzada,sendo estas delimitadas, problemas com predição de amostras fora dos limites de observação amostral são reduzidos.Quando as amostras fixadas(**vermelho**) são incluídas no conjunto de calibração, se torna possível avaliar sua influência na modelagem de calibração, umas vez que é possível verificar a variabilidade do RMSECV(Root-Mean-Square Error of Cross-Validation) e RMSEP (Root Mean Square Error of Prediction) à medida que mais amostras são fixadas.

As amostras de predição foram selecionadas tendo como parâmetro sua distribição no espaço amostral(amostras mais externas, amostras mais internas e amostras homogeneamente distribuídas), as amostras que fornecessem os melhores resultados em termos de menores raizes quadratícas de erros médios de previsão e de validação cruzada (RMSEP e RMSECV) respectivamente. A escolha das amostras de

predição em face da sua distribuição espacial, foi utilizada na modelagem PLS-grou-out e MLR-APS.

O algoritmo foi,integralmente, implementado (código-fonte no **APÊNDICE**) e executado em ambienteMatlab® 2010b (Mathworks) e opré-processamento dos dados foi realizado, exclusivamente, usando o software Unscrambler® 9.7 (CAMO AS).

2.5. Uso da Espectrometria NIR para avaliação da estratégia proposta

A espectrometria do infravermelho próximo(NIR-Near InfraRed espectrometry) é uma técnica espectrométrica que utiliza a REM (Radiação EletroMagnética), cujos comprimentos de onda encontam-se na faixa de 750 a 2500 nm[80]. A interação dessa radiação com a matéria (absorção ou emissão de fótons) promove apenas mudanças em estados vibracionaisdas moléculas[81,82]. Os espectros NIR consistem de bandas largas resultantes de transições vibracionas – envolvendo, geralmente, ligações do tipo C-H, N-H, O-H e S-H – associadas aos sobretons e bandas de combinação de transições fundamentais que ocorrem no MIR (Mid InfraRed)[1,3,86].

Em determinações espectroanalíticas usando a espectrometria NIR, é comum a obtenção de espectros que acarretam dadosmultivariados complexos,dificultando a extração das informações mais relevantes para as análises químicas. Ademais, a complexidade dos espectros se torna maior quando os sinais do analito são sobrepostos pelas bandas de, sobretudo, concomitante(s) em matrizes complexas.Devido às caracteríticas dos espectros NIR, a calibração univariada geralmente não permite obter modelos com acurácia satisfatória[80-81] mesmo na determinação de um único componente (analito ou parâmetro de interesse) em uma matriz menos complexa[3] . Para superar esse problema, recorre-se usualmente à modelagem empregando técnicas quimiométricas de calibração multivarida

2.5. Espectroscopia NIR na determinção de conteúdo de Proteina em trigo e de umidade em milho.

Na espectroscopia NIR a quantificação desses parâmetros nas amostras em estudo são feitas sem separação previa do analito da matriz e isso requer a utilização de metódos de calibração multivariada, tendo em vista que o sinal do analito não é evidenciado de forma isolada no espectro NIR[81]**.** O fato de necessitar de escasso ou

nenhum tratamento das amostras, proporciona análises mais rapidas e com geração de residuos minimizada, contribuindo assim para o meio ambiente[81].

A determinação de conteúdo de proteína em amostras de trigo e úmidade em amostras milho foram feitas empregando a espectroscopia NIR, por ser uma técnica não-invasiva, não-destrutiva e por fornecer bons resultados semdanificar a amostra para estudos posteriores[3,84]. O uso no NIR na determinação desses parâmetros, propicia a inserção de algoritmos de seleção de variáveis, tendo em vista que essas análises espectroscópicas fornecem uma grande quantidade de informações e faz-se necessario selecionar as variáveis relevantes na construção do modelo[84,87].

CAPÍTULO 3

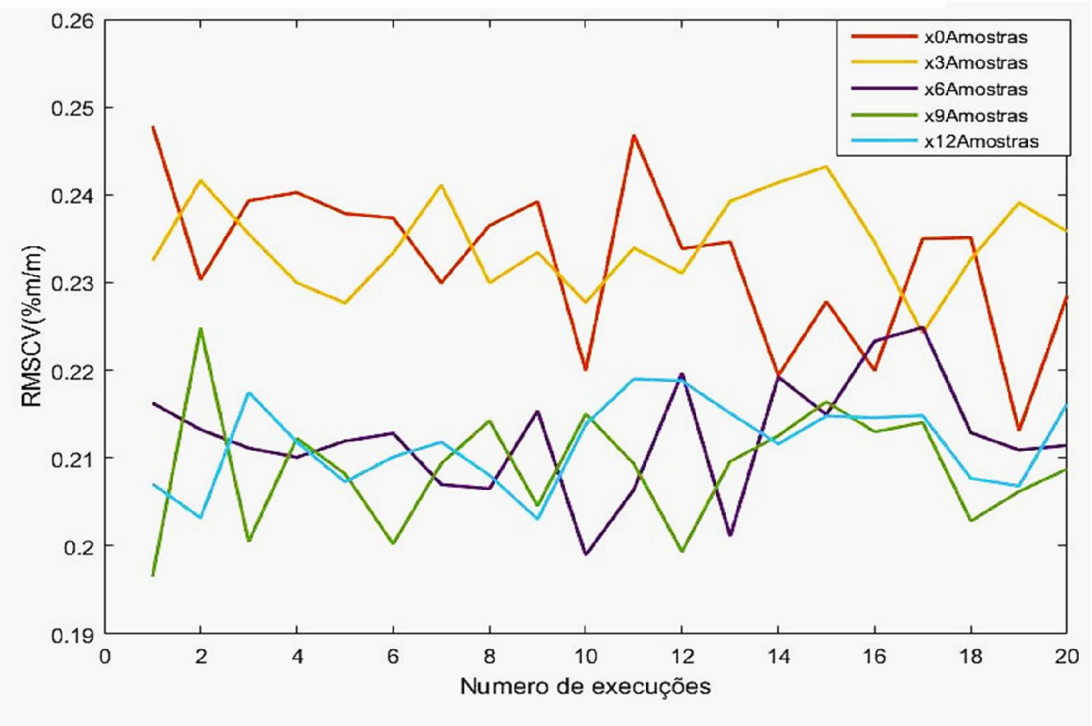
Resultados e Discussão

3.1. PLS group-out com seleção de amostras fixadas

Inicialmente, vários testes foram realizados com o intuito de verificar quais amostras de validação (mais internas, externas e distribuídas homogeneamente) proporcionariam uma diminuição do RMSECV e do RMSEP. Os resultados expressos foram obtidos depois de vinte execuções do algoritmo e avaliados seus respectivos valores de RMSECV e RMSEP. Foram utilizadas 20 execuções,com o intuito de avaliar a previsibilidade do modelo quanto a fixação das amostras mais externas e tambem avaliou-se o RMSECV médio para cada modelo obtido.Na determinação das amostras a serem retiradas na modelagem PLS group-out, diferentes tipos de amostras de validação foram explorados e o modelo com menores valores de RMSECV foi escolhido.

No conjunto de dados das amostras de trigo, as amostras de predição mais internas proporcionaram uma diminuição significativa em termos de RMSECV. Após o diagnóstico das amostras de predição que mais se adequassem à modelagem,gerou-se o gráfico da **Figura 3.1** no qual se avaliou a variabilidade do RMSECV em função do número deamostras fixas adicionadas.

Figura 3.1. RMSECV amostras trigo em função do número de amostras.fixadas.



Na **Figura 3.1** pode-se observar a diminuição do RMSECV a medida que as amostras que permanecerão fixas são adicionadas, a variabilidade desses valores

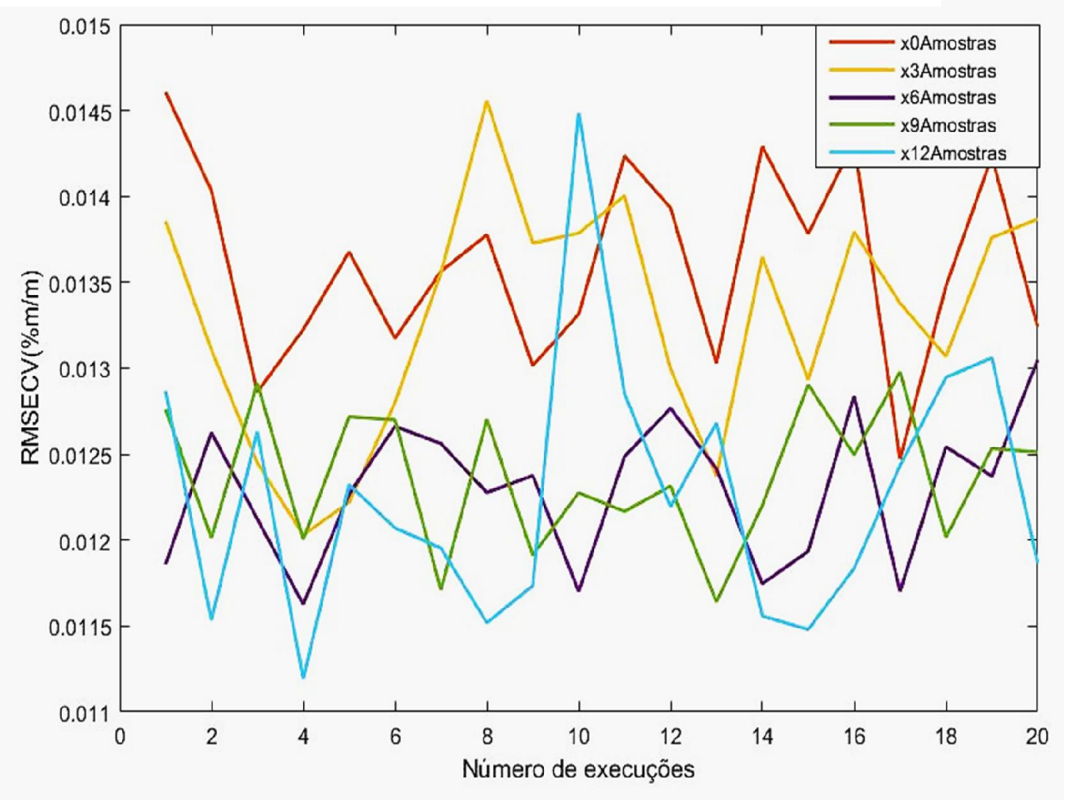
tambem é reduzida, sendo esta mais perceptível ao comparar-se o modelo com zero amostras fixadas(**vermelho**) com os modelos com seis, nove e doze amostras fixadas(**roxo,verde,azul**). Ao comparar-se o modelo com nenhuma amostras fixada(vermelho) com os demais modelos no qual as amostras foram fixadas, é perceptível que a adição das mesmas favorece uma diminuição significativa do RMSECV e de sua variabilidade, demonstrando que ao se adicionar amostras que permanecerão fixas no conjunto de calibração pode-se melhorar a robustez da modelagem. O modelo com zero amostras fixadas(**vermelho**), apresentou um RMSECV medio de 0,235(%m/m), enquanto que os modelos com três(**laranja**),seis(**roxo**),nove(**verde**) e doze(**azul**) apresentam os respectivos valores de RMSECV 0,234(%m/m), 0,212(%m/m), 0,210(%m/m) e 0,212(%m/m).Pode-se observar que o modelo com nove amostras fixadas(**verde**), apresenta um menor RMSECV médio, evidenciando assim que esse número de amostras é o mais adequado para a construção do modelo de calibração. A medida que aumenta o número de amostras fixadas, as fronteiras de calibração vão sendo melhor delimitadas, diminuindo assim a ocorrência de predições além dos limites de observação, acarretando assim uma diminuição no RMSECV(raiz quadrada do erro quadratico médio de validação cruzada).

Na análise do conjunto de dados de milho, vinte execuções tambem foram realizadas com o intuito de avaliar a obviedade dos modelos quando amostras são fixadas e assimobteveum RMSECV médio. A escolha das amostras de predição mais externas forneceram os menores valores e a menor variavibilídade dentre os modelos estudados em termos de RMSECV. Pode-se observar uma diminuição da variabilidade do RMSECV quando amostras são fixadas no conjunto de calibração. Na **Figura 3.2** é possivel observar a medida que mais amostras fixas vão sendo adicionadas no conjunto de calibração,uma queda acentuada do RMSECV e uma baixa alterabilida do mesmo ao longo das vinte execuções.Os modelos com zero(**vermelho**) e com três (**laranja**) amostras fixadas apresentaram as maiores mutalibilidades de RMSECV, mesmo assim é possivel observar que o modelo no qual amostras foram fixadas, exibe um perfil com menor varibilidade dos dados.

Dentre os modelos estudados, aquele com menor alterabilidade em termos de RMSECVfoi selecionado. Quando doze amostras (**azul claro**) foram fixadas no conjunto de validação, pode-se notar uma queda em termos de variação do RMSECV, exceto na decima execução, onde possivelmente as amostras mais externas não eram de

fronteira, acarretando assim um aumento na sua variabilidade.Os modelos obtidos apresentaram uma pequena diferença em termos de RMSECV, levando-se em conta aqueles com e sem a fixação de amostras na etapa de calibração. O modelo com zero amostras fixadas(**vermelho**) apresentou um RMSECV de 0,0140(%m/m), já os modelos com três(**laranja**) 0,0135(%m/m),seis(roxo) 0,0124(%m/m),nove(**verde**) 0,0124(%m/m) e doze(**azul**) 0,0121(%m/m).Em face do exposto, é perceptível que os modelos com amostras fixadas no conjunto de calibração proporcionam em média menores valores de RMSECV que os sem nenhuma amostras fixada, demonstrando assim que a inclusão destas pode proporcinar uma melhora significativa no modelo de calibração.

Figura 3.2. RMSECV amostras milho em função do número de amostras fixas



Com a abordagem PLS adotada para esses conjuntos de dados, não foi possivel avaliar a variabilidade do RMSEP em função do índice de amostras, pois os mesmos não apresentavam variabilidade significativa dos resultados para as vinte execuções, impossibilitando assim uma melhor avaliação para a modelagem proposta.

3.2. MLR-APS Fixação de amostras

3.2.1 Dados NIR de amostras de trigo

À medida que amostras selcionadas, previamente, são fixadas no conjunto de calibração proporcionam uma diminuição em termos de RMSEP e RMSECV em modelagem MLR-APS. Na **Tabela 3.1** é possivel observar que o modelo no qual quatro amostras são fixadas apresentam os menores valores de raiz quadrada do erro medio de previsão(RMSEP) e de raiz quadrática do erro quadrático médio de validação cruzada (RMSECV), mesmo com um número menor de amostras que participam do procedimento de validação cruzada, demonstrando que as amostras mais externas ao serem fixadas proporcinam a aquisição de melhores modelos de calibração.

Apesar de ter um número menor de amostras, que participam do procedimento CVLOO, os modelos que possuem amostras fixadas na etapa de calibração apresentam desempenho ligeiramente melhor que o modelo sem que amostras sejam fixadas na etapa de calibração. Ao comparar-se o modelo com o número minímo de amostras fixadas e o modelo com o maxímo de amostras fixadas, pode-se observar que o número de amostras que participam do procedimento validação cruzada diminui significativamente de 77 para 61 amostras e pode-se observar seus respectivos valores de RMSECV e RMSEP que são 0,261%(m/m) e 0,197%(m/m), para o número minímo de amostras fixadas e 0,231%(m/m) e 0,164%(m/m) com o maxímo de amostras fixadas.

Tabela 3.1. Valores de RMSEP e RMSECV em relação ao número de amostras de trigo fixadas.

Nº de Amostras Fixadas	RMSEP %(m/m)	Nº de Variáveis selecionadas	RMSECV%(m/m)
0	0,261	24	0,197 (77)
4	0,194	19	0,163 (73)
8	0,266	18	0,164 (69)
12	0,219	19	0,177 (65)
16	0,231	20	0,164 (61)

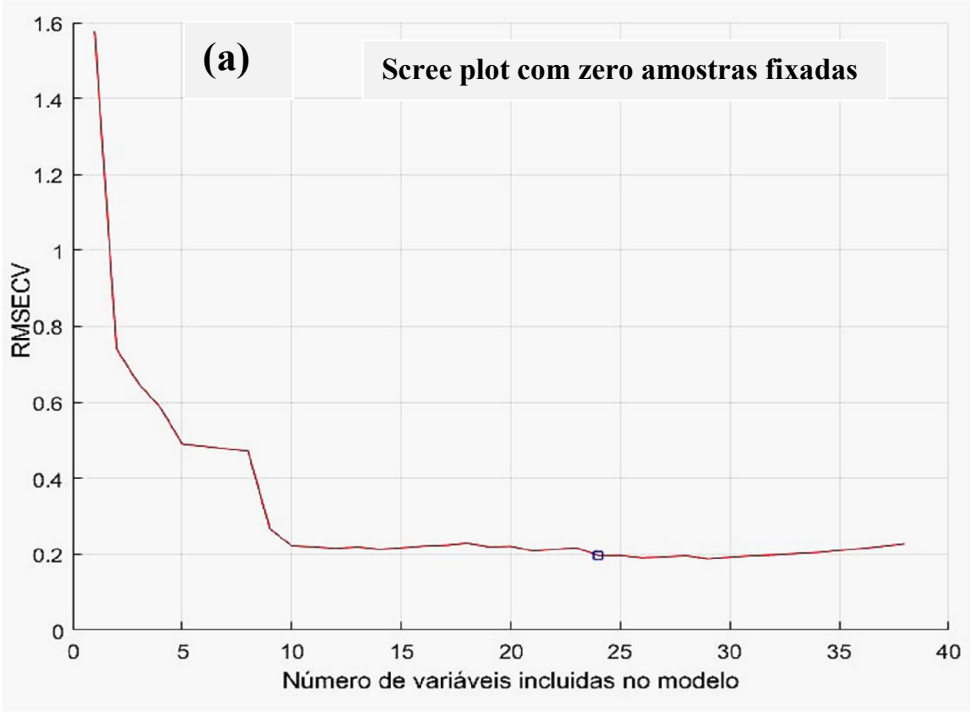
Os valores de RMSEP para os modelos com amostras fixadas diminuem sinificativamente(com excesssão do modelo com 8 amostras fixadas)quando

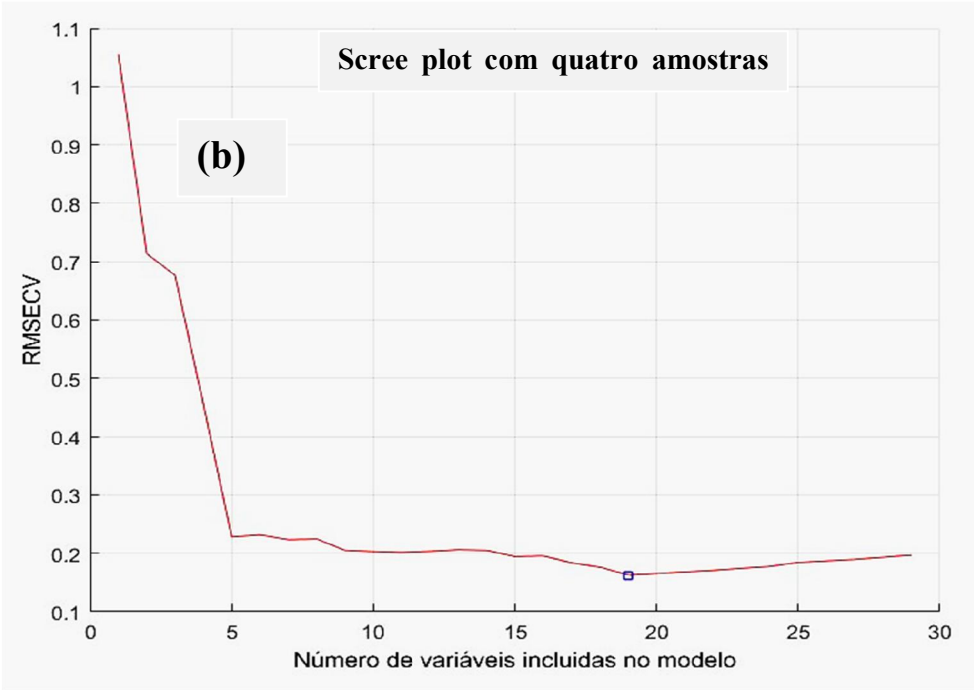
comparados com o modelo sem amostras fixadas. O aumento em termos de RMSEP em alguns modelos com amostras fixadasocorreu em virtude da fixação de amostras mais internas,pois nem todas as amostras selecionadas pelo SPXYsão necessariamente as mais externas(de fronteira), tendo em vista que nesse algoritmo a seleção dessas amostras leva-se em conta tambem a correlação entre as matrizes X e Y. Amedida que amostras são fixadas na etapa de calibração, um número menor de variáveissão incluidas no modelo de calibração, demonstrando bom condicionamento dos dados.O modelo com zero amostras fixadas apresentou um baixo desempenho em termos de RMSEP, RMSECV e númerode variáveis selecionadas para a construção do modelo, que são respectivamente 0,261% (m/m),0,197% (m/m) e 24 variáveis selecionadas. Já o melhor modelo foi obtido quando quatro amostras foramfixadas 0,194% (m/m), 0,163% (m/m) e 19, que foram seus respectivos valores de RMSEP, RMSE e número de variáveis selecionadas. Quando essas amostras são fixadas na etapa de calibração, em modelagem APS-MLR demonstrou grande potencial, pois uma vez que as amostras de fronteira são fixadas os problemas com extrapolação de modelo são minimizados e consequentemente ocorre uma melhoria na capacidade preditiva do mesmo.

3.2.2 Análise das figuras scree plot

Nas **Figuras 3.3a e 3.3b** é possível observar que quando quatro amostras fixas são adicionadas, o modelo apresenta baixos valores de RMSECV com apenas cinco variáveis selecionadas na modelagem MLR-APS, enquanto o modelo sem nenhuma amostra fixa adicionada apresenta diminuição considerável do RMSECV a partir de oito variáveis selecionadas.O número de variáveis ideal para o modelo com zero amostras adicionadas foi 24 variáveis selecionadas e para o modelo com quatro amostras adiciondas foi 19 variáveis, demonstrando assim que a adição de amostras fixadas na etapa de calibração proporciona uma simplificação para a modelagem APL-MLR.

Figura 3.3. (a) Scree plot com zero amostras fixadas.e (b) Scree plot com quatro





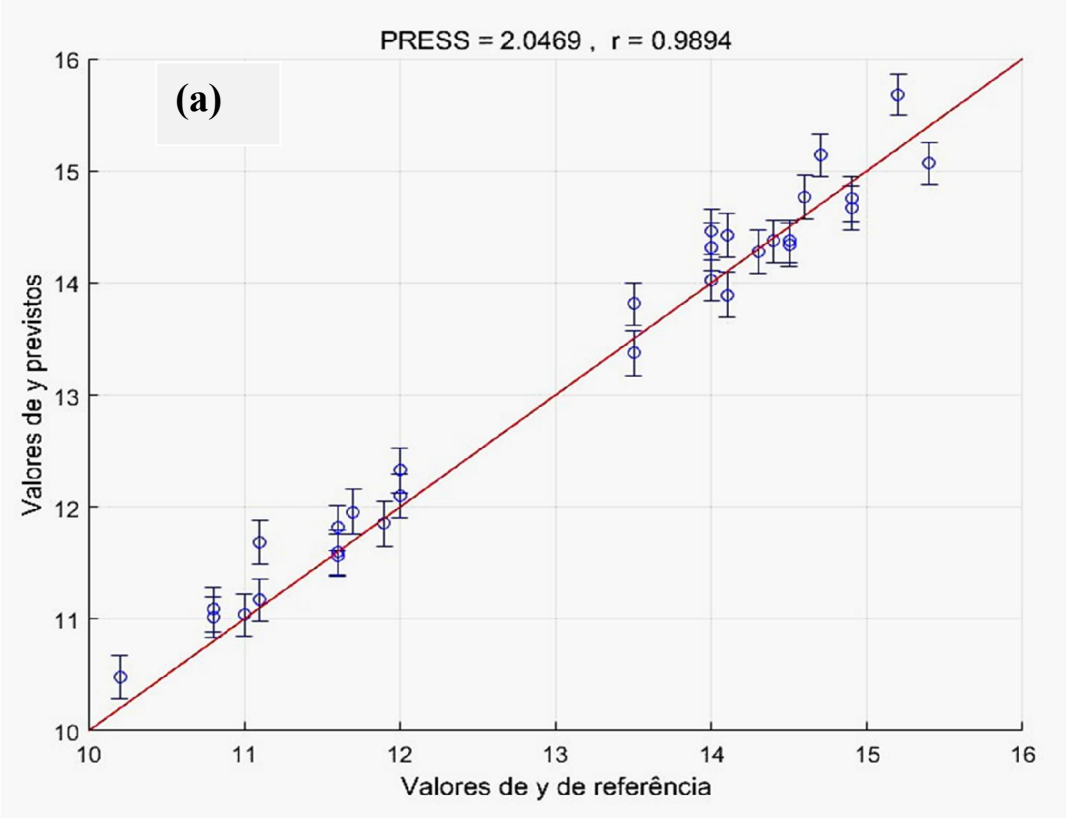
Vale ressaltar também que, com quatro amostras fixadas,observa-se uma diminuição significativa em termos RMSECV com apenas 5 variáveis selecionadas, enquanto que o modelo com zero amostras fixadas esse efeito pode ser observado

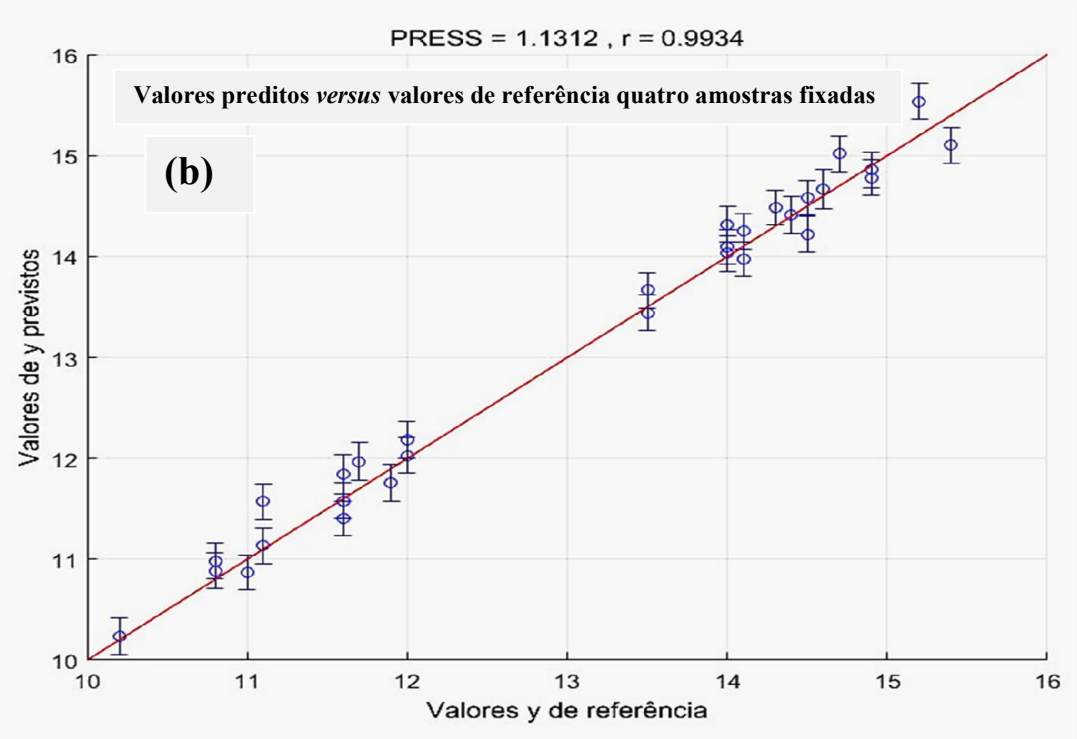
quando 8 variáveis são selecionadas, demosntrando assim que o modelo com quatro amostras fixadas geram modelos mais simples

3.2.3. Valores preditos *versus* valores de referência

Nas **Figuras 3.4a e 3.4b** pode-se observar que o modelo com quatro amostras fixadas estão ligeiramente mais ajustados a reta, quando compararado ao modelo com nenhuma amostra fixada. Os coeficientes de correlação para os modelos com zero e quatro amostras foramrepectivamente 0,9894 e 0,9934, demonstrando que ao se adicionar as amostras fixas na calibração proporcina uma melhoria no modelo.Os modelo com zero e quatro amostras fixadas apresentam bom condicionamento dos dados, visto que não há tendência quanto a disposição dos mesmos. Porém o modelo com 4 amostras fixadas apresenta um melhor coeficiente correlação dos dados sendo este mais indicado para a construção de um bom modelo de calibração.

Figura 3.4. (a) Valores preditos *versus* valores de referência zero amostras fixadas e (b) Valores preditos *versus* valores de referência quatro amostras fixadas.





Quando quatro amostras são adicionadas é possivel perceber uma redução do somatório dos quadrados dos erros risuduais de predição (PRESS) em aproximadamente 50% em relação ao modelo com zero amostras adicionadas.Essa redução indica que o modelo com quatro amostras fixadas é mais adequado para a construção de um bom modelo de calibração MLR-APS.

3.2.4 Dados NIR de amostras de milho

Na determinação do teor de umidade em amostras de milho, o modelo no qual amostras foram fixadas apresentaram melhores desempenhos em termos de RMSEP e RMSECV, como pode ser evidenciado na **Tabela 3.2**. Dentre todos os modelos, aquele com três amostras fixadas apresentou o pior desempenho, tendo selecionado um número maior de variáveis incluidas no modelo, que aumenta a complexidade do mesmo e apresentou os maiores valores de RMSECV e RMSEP que são respectivamente 0,0085%(m/m) e 0.0155%(m/m). Essas três amostras fixadas não eram amostras de fronteira e portanto o espaço amostral não foi delimitado com precisão, afetando assim a capacidade preditiva do modelo.

Tabela 3.2. Valores de RMSEP e RMSECV em relação ao número de amostras de milho fixadas.

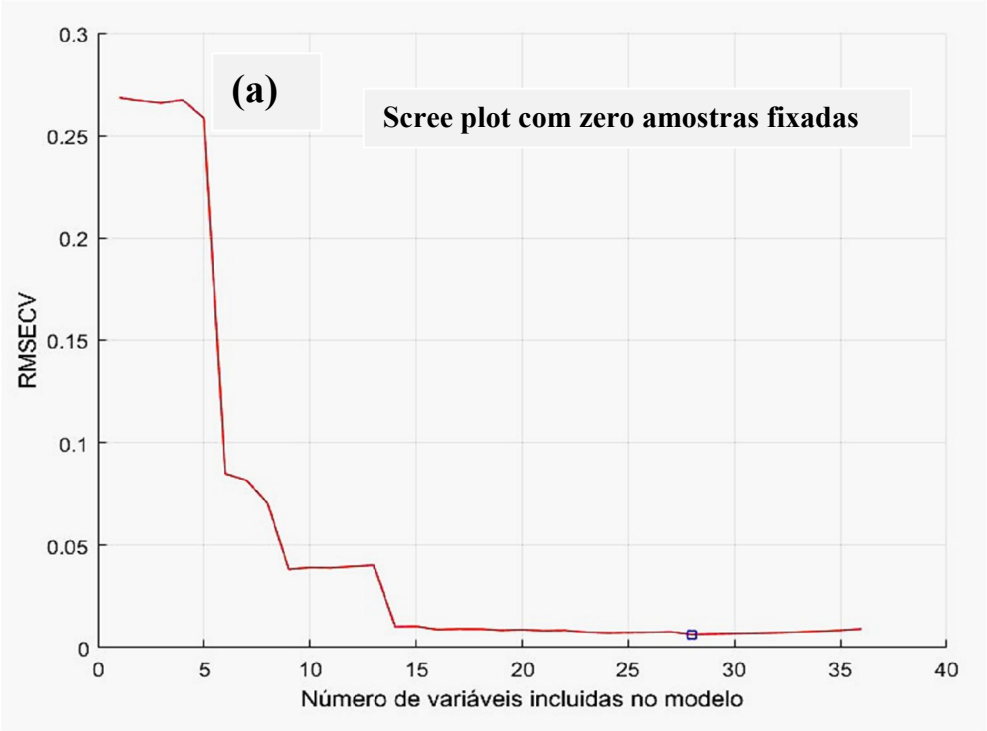
Nº de Amostras Fixadas	RMSEP %(m/m)	Nº de Variáveis	RMSECV %(m/m)
0	0.0142	28	0.0065 (60)
3	0.0155	30	0.0085 (57)
6	0.0115	24	0.0063 (54)
9	0.0125	21	0.0087 (51)
12	0.0114	21	0.0061 (48)

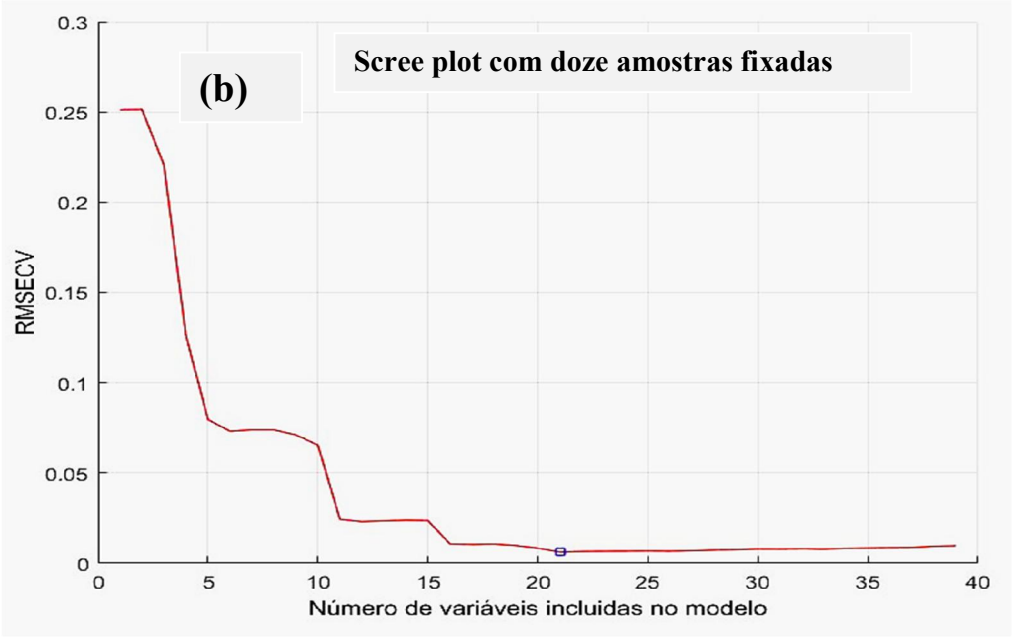
O modelo com zero amostras fixas adicionadas apresentou valores de RMSEP e RMSECV bem maiores que os modelos com seis, nove e doze amostras, mesmo selecionando um número maior de variáveis, demonstrando que os modelos sem amostras fixadas são menos parcimoniosos.Quando doze amostras fixas são adicionadas na etapa de calibração apenas 21 variáveis foram selecionadas e os menores erros de predição(RMSEP) e erros de validação cruzada(RMSECV) foram obtidos 0,0114%(m/m) e 0,0061%(m/m). No modelo sem amostras fixadas, 28 variáveis foram selecionadas e mesmo assim o modelo apresentou altos valores de RMSEP e RMSECV por essa ordem 0,0142% (m/m) e 0,0065%(m/m). A fixação de amostras mais externas na etapa de calibração, além de minimizar problemas com extrapolação do modelo, tambem permite a obtenção de modelos mais simples e parcimoniosos.

3.2.5 Análise das figuras Scree Plot

Os gráficos de“scree plot, obtidos para o pior e o melhor modelo, são apresentandos, respectivamente, nas **Figuras 3.5a e 3.5b**. Pode-se observar a diminuição do RMSECV a medida que mais variáveis são selecionadas.

Figura 3.5. (a) Scree plot com zero amostras fixadas.b) Scree plot com doze amostras fixadas.





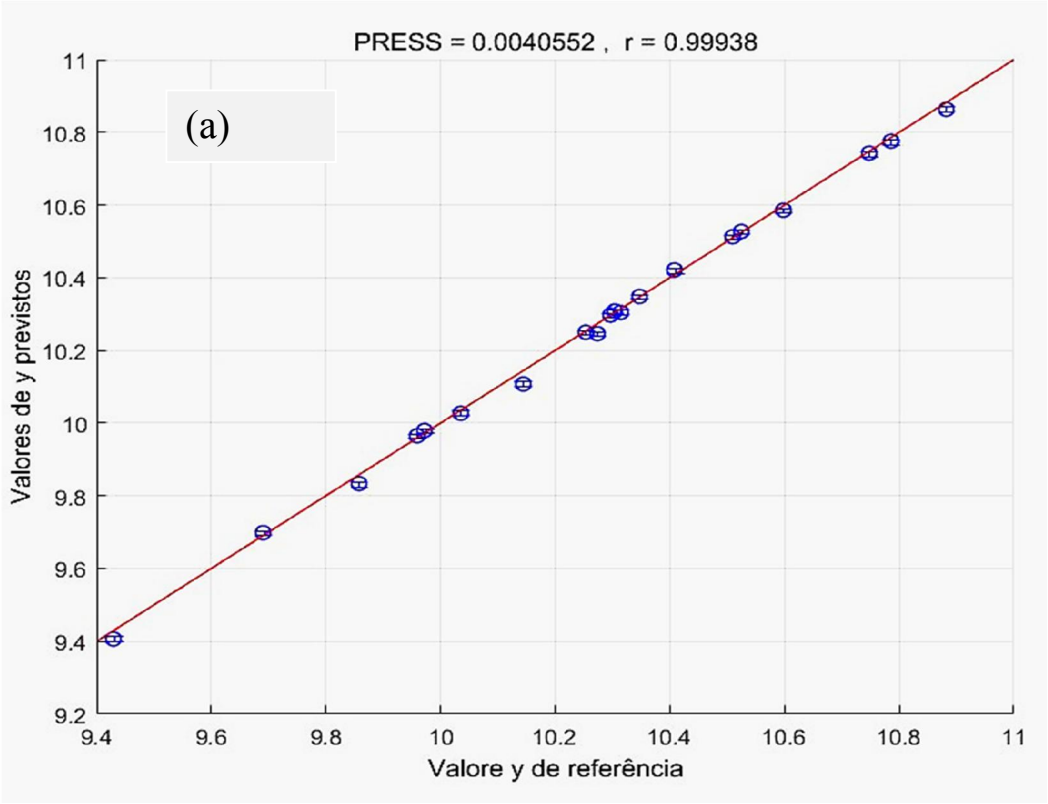
Na **Figura 3.5b**, pode-se observar que três variáveis são incluidas no modelo, proporcionam um decréscimo do RMSECV, já na figura 3.5a, um número maior de

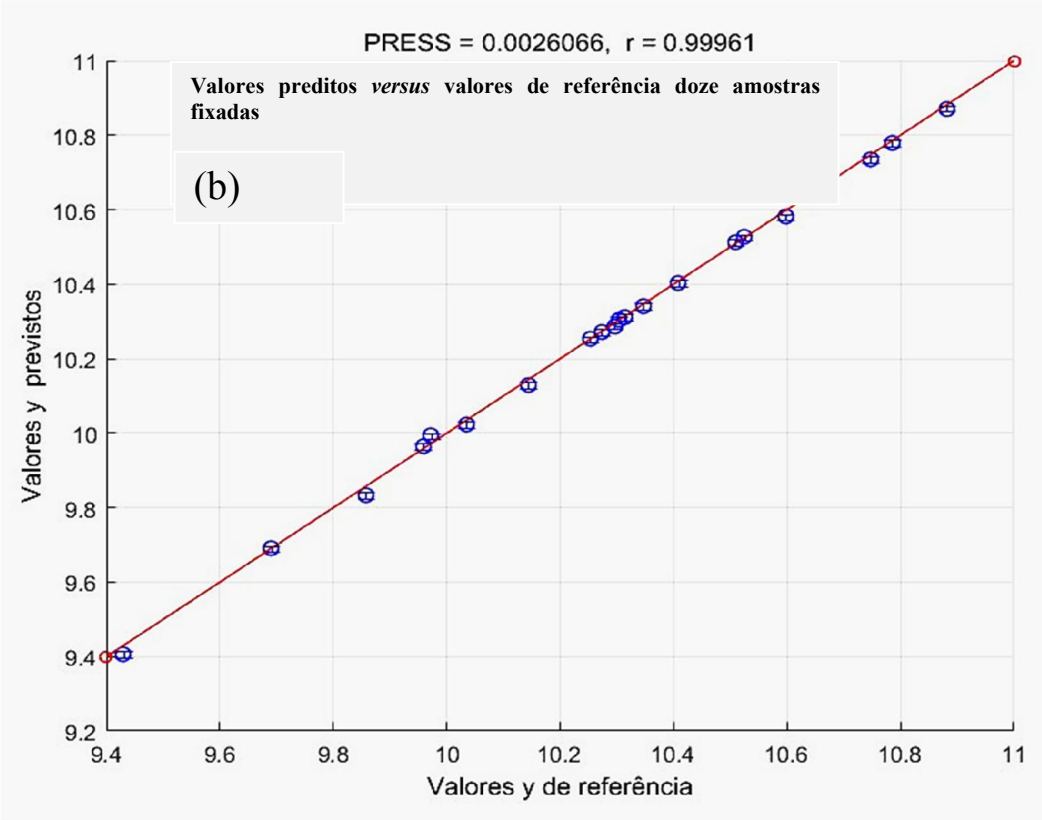
variávies devem ser incluidas no modelo para proporcinar uma diminuição do RMSECV.O modelo com doze amostras fixadas selecionou 21 variáveis como sendo o número ideial para a construção de um bom modelo, enquanto que aquele sem a presença de amostras fixadas, incluiu 28 variáveis para construção de um modelo bem ajustado.

3.2.6. Valores Preditos *versus* valores de referência

O modelo com doze amostras fixadas apresenta um coefiencente de correlação ligeiramente maior que o obtido sem nenhuma amostra fixada sendo estes iguais a, repectivamente, 0,99961 e 0,99938. Nas **Figuras 3.6a e 3.6b** é percebivel que não ocorreu diferenças significativas de como as amostras estão dispostas na reta de ajuste para ambos modelos.

Figura 3.6. (a) Valores preditos *versus* valores de referência com zero amostras fixadas e (b) Valores preditos *versus* Valores de referência com dezesseis amostras fixadas.





A diferença mais significativa entre os resultados expostos, ocorre ao comparar-se os resultados em termos de valores de PRESS. O modelo com doze amostras fixadase o modelo sem amostras fixadas, percebe-se uma redução significativa 0,00261%(m/m) e 0,0040%(m/m) respectivamente, que demonstra que a modelagem com doze amostras fixadas é o mais adequado para construção de modelos mais parcimonisos e de melhor capacidade preditiva em modelos MLR-APS.

CAPÍTULO 4

Conclusões

4.1 CONCLUSÕES

No presente trabalho, propôs-se uma nova estratégia de validação híbrida com seleção de amostras fixadas pelo algoritmo SPXY em modelagens PLS group-out e MLR-APS para minimizar problemas de extrapolação de modelos de calibração multivariada. A estratégia proposta foi avaliada, inicialmente, em modelagem PLS group-out usando dois conjuntos de dados. O primeiro compreende 107 amostras de trigo, no qual o parâmetro de interesse era o conteúdo de proteína sendo os espectros NIR 680 variáveis (comprimentos de onda) após realizado o pré-processamento. O segundo conjunto envolveu 80 amostras de milho, cujo parâmetro alvo no estudo foi o teor de umidade, sendo os sinais medidos também em 680 varáveis (após pré-processamento).

Para ambos os conjuntos de dados, a estratégia proposta foi avaliada em termos de RMSECV e RMSEP em função do número de amostras fixadas. Durante a realização desses estudos percebeu-se que, apesar de obter bons resultados em termos de RMSECV, o mesmo não ocorreu com os valores de RMSEP. Em virtude dos resultados insatisfatórios em termos de RMSEP para o modelo PLS para uma única execução do algoritmo, foram realizadas vinte execuções do algorimo para os dois conjuntos de dados; avaliou-se então seus respectivos RMSEP médios. De fato, essa última métrica apresentava variabilidade muito pequena à medida que mais amostras fixadas foram adicionadas ao longo das vinte execuções, impossibilitando uma avaliação adequada da validação híbrida em calibração PLS.

Em face desse óbice, optou-se por utilizar essa validação híbrida em modelos de calibração multivariada baseados em seleção de variáveis. A modelagem MLR-APS foi utlizada para os mesmos conjuntos de dados. Observou-se que, diferentemente da modelagem PLS, os resultados em termos de RMSEP mostraram uma certa variabildade à medida que mais amostras fixadas são adicionadas na etapa de calibração.

Para o conjunto de dados de trigo, houve uma redução do RMSEP de, aproximadamente, 25% quando se compara o modelo, obtido com base em quatro amostras fixadas, com o obtido sem nenjuma amostra fixada. Além disso, o modelo baseado em quatro amostras fixadas proporcionou uma redução no RMSECV em cerca de 17% e com um número de variáveis selecionadas bem menor.

No conjunto de dados do milho, quando se compara o modelo com doze amostras fixadas e o modelo sem amostras fixadas, é possível observar uma diminuição

do número de variáveis selecionadas, bem como uma diminuição dos respectivos valores de RMSEP e RMSECV para, respectivamente, cerca de 20% e 6%.

A abordagem proposta mostrou-se bem eficiente em modelagem MLR-APS, demonstrando que, quando as amostras mais externas são fixadas, os problemas de extrapolação na modelagem são minimizados e proporcionam uma melhoria na capacidade preditiva do modelo de calibração. A inserção dessas amostras faz com que um menor número de variáveis sejam selecionadas, proporcionando modelos mais parcimoniosos e de fácil interpretação. Como as amostras mais externas são fixadas, os modelos tornam-se menos susceptíveis a variações que prejudicam sua robustez. Ademais, o uso das amostras de fronteira pode minimizar (ou até prevenir) problemas de extrapolação da região envolvida na modelagem de calibração.

4.2. Perspectivas Futuras

- Re-aplicar a validação híbrida, com seleção de amostras fixadas pelo algoritmo SPXY, em modelagem PLS para investigar a razão da baixa variabilidade do RMSE à medida que mais amostras fixadas são adicionadas.

- Aplicar a estratégia de validação híbrida a novos conjuntos de dados.

- Adaptar a estratégia proposta à modelagem multivariada envolvendo problemas de classificação.

Referências

1. Ferreira,M.M.C;Antunes,A.M;Melgo.M.S,&Volpe,P.L.O.(1999).Q**uimiométria I:Calibração Multivariada, um tutorial.**Química Nova, Vol. 22, No. 5, 724-731.
2. Blanco.M; Coello.J, Iturriaga.H, Maspoch. S,& Bertran.E.(1995). **Simultaneous Determination of Rubber Additives by FT-IR Spectrphotometry with Multivariate Calibration**.Applied Spectroscopy,Vol 49,N°.6,747-753.
3. Bokobza, L. (1998). **Near Infrared Spectroscopy**. Journal of Near Infrared Spectroscopy, 6(1), 3–17.
4. Walmsley, A.(1997). **Improved variable selection procedure for multivariate linear regression**. Analytica Chimica Acta, 354(1-3), 225–232.
5. Xu, L., Fu, H.Y., Goodarzi, M., Cai, C.-B., Yin, Q.-B., Wu, Y., She,Y.B. (2018). **Stochastic cross validation.** Chemometrics and Intelligent Laboratory Systems, 175, 74–81.
6. Wakeling, I. N., & Morris, J. J. (1993**). A test of significance for partial least squares regression**. Journal of Chemometrics, 7(4), 291–304.
7. Efron, B. (1986). **How Biased is the Apparent Error Rate of a Prediction Rule?** Journal of the American Statistical Association, 81(394), 461–470.
8. Kenard,R.W;Stones,L.A(1969).**Computer Aided Design of Experiments**.Technometrics,Vol. 11.No.1,137-148.
9. GALVAO, R., ARAUJO, M., JOSE, G., PONTES, M., SILVA, E., &SALDANHA, T. (2005). **A method for calibration and validation subset partitioning**. Talanta, 67(4), 736–740.
10. Höskuldsson, A. (1996). **Dimension of linear models**. Chemometrics and Intelligent Laboratory Systems, 32(1), 37–55.
11. Varma, S., & Simon, R. (2006**). Bias in error estimation when using cross-validation for model selection BMC**. Bioinformatics, 7(1), 91.
12. Braga,J.W.B,Poppi,R.J.(2004).**Validação De Modelos De Calibração Multivariada: Uma Aplicação Na Determinação De Pureza Polimórfica De Carbamazepina Por Espectroscopia No Infravermelho Próximo**.Quim. Nova, Vol. 27, No. 6, 1004-1011.
13. Zhang, P.(1993).**Model Selection Via Multifold Cross Validation, the** Annals of Statistics,Vol. 21, No. 1 . 299-313.

14. Brereton, R.G.(2000).**Introduction to multivariate calibration in analytical chemistry**. Analyst, 2000, **125**, 2125–2154.
15. Xu, Q.-S., Liang, Y.-Z., & Du, Y.-P. (2004). **Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration**. Journal of Chemometrics, 18(2), 112–120.
16. Du, Y. P., Kasemsumran, S., Maruo, K., Nakagawa, T., & Ozaki, Y. (2006). **Ascertainment of the number of samples in the validation set in Monte Carlo crossvalidation and the selection of model dimension with Monte Carlo crossvalidation**. Chemometrics and Intelligent Laboratory Systems, 82(1-2), 83–89.
17. Arlot, S., &Celisse, A. (2010). **A survey of cross-validation procedures for model selection**. Statistics Surveys, 4(0), 40–79.
18. Martens,H.A,Dardene.P.(1998**).Validation and verification of regression in small data sets,** Chemometrics and intelligent laboratory systems,44,99-121.
19. Wold,S;Sjöström.M;Eriksson.Lennart.(2001).**PLS-regression:abasictool of chemometrics**; Chemometrics and inteligent laboratory systems, 58,109-130.
20. Mao, X. D., Sun, L. J., Hao, G., Xu, L. L., &Hui, G. Y. (2013). **Optimization of Wheat Protein Near Infrared Calibration Model Based on SPXY**. Advanced Materials Research, 803, 122–126.
21. Valderrama, P., Braga, J. W. B., &Poppi, R. J. (2009). **Estado da arte de figuras de mérito em calibração multivariada**. Química Nova, 32(5), 1278–1287.
22. Kowalski,B,R. **Chemometrics Mathematics and Statistics in Chemistry.** Proceedings of the NATO Advanced Study Institute on Chemometrics–Mathematics and Statistics in Chemistry Cosenza, Italy September 12-23, 1983.
23. Cerqueira,E.O;Andrade,J.C;Poppi,R.J. (2001). **Redes neurais e suas aplicações em calibração multivariada.** Quim. Nova, Vol. 24, No. 6, 864-873, 2001.
24. Nagata, N; Bueno, M.I.M.S; Peralta, P.G.Z. **Métodos matemáticos para correção de interferências espectrais e efeitos Inter-elementos na análise quantitativa por fluorescência de raios-x.**Quim. Nova, Vol. 24, No. 4, 531-539, 2001.
25. Beebe, K. R., & Kowalski, B. R. (1987). **An introduction to multivariate calibration and analysis**. Analytical Chemistry, 59(17), 1007A–1017A.

26. Gemperline, P.**Practical guide To chemometrics**, 2° ed, Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, 2006.
27. Jolliffe, I. T. (1993). **Principal component analysis: A beginner's guide - II. Pitfalls, myths and extensions**. Weather, 48(8), 246–253.
28. Martínez,A.M; Kak,A.C. **PCA versus LDA**.IEEE Transactions on pattern analysis and machine intelligence, VOL. 23, NO. 2, FEBRUARY 2001.
29. MESSERSCHMIDT, I. **Análise quantitative por espectroscopia no infravermelho médio empregando técnicas de reflectância e calibração multivariada.**Tese de Doutorado, IQ-UNICAMP, 1999.
30. Jolliffe, I. T. (1982). **A Note on the Use of Principal Components in Regression**. Applied Statistics, 31(3), 300.
31. Marana, Aparecido Nilceu; Ribeiro, Icaro; Chiachia, Giovani. **Reconhecimento de faces sob diferentes condições de iluminação utilizando PCA e a transformada census**. Learning and Nonlinear Models, v. 9, n. 2, p. 138-144, 2011.
32. E.Z. Tenorio, C.E. Thomaz. **Análise multilinear discriminante de formas frontais de imagens 2d de face** Proceedings of the X Simpósio Brasileiro de Automação Inteligente, SBAI (2011), pp. 266-271.
33. Picard, R. R., & Cook, R. D. (1984). **Cross-Validation of Regression Models**. Journal of the American Statistical Association, 79(387), 575–583.
34. Shao J. **Linear model selection by crossvalidation**.J. Amer. Statist. Assoc. 1993; 88: 486–494.
35. A.M. Molinaro, R. Simon, R.M. Pfeiffer**, Prediction error estimation: a comparis on of Resampling methods**, Bio informatics 21 (2005) 3301–3307.
36. A.A. Gowen, G. Downey, C. Esquerre, C.P. O'Donnell, **Preventing over-fitting in PLS Calibration models of near-infrared (NIR) spectroscopy data using regression coefficients,** J. Chemometrics. 25 (2011) 375–381.
37. H.van der Voet;**Comparing the predictive accuracy of models using a simple randomization test,** Chemometrics and Intelligent Laboratory Systems 25 (1994) 313-323.
38. Xu, Q.-S., & Liang, Y.-Z.. **Monte Carlo cross validation**. Chemometrics and Intelligent Laboratory Systems, 56(2001), 1–11.

39. David.W.O.(1988). **Selection of optimal regression models via Cross-validation.**JOURNAL OF CHEMOMETRICS. VOL. 2. 39-48.
40. D.K. Pedersen, H. Martens, J.P. Nielsen, S.B. Engelsen. **Near-Infrared Absorption and Scattering Separated by Extended Inverted Signal Correction (EISC): Analysis of Near-Infrared Transmittance Spectra of Single Wheat Seeds**. Appl. Spectrosc. 56 (2002) 1206–1214.
41. Barros Neto, B. de, Pimentel, M. F., & Araújo, M. C. U. (2002). **Recomendações para calibração em química analítica: parte I. Fundamentos e calibração com um componente (calibração univariada).** Química Nova, 25(5), 856–865.
42. Hongyu, K., Sandanielo, V. L. M., Junior, G. J. O. **Análise de Componentes Principais: resumo teórico, aplicação e interpretação**.E&S - Engineering and Science, (2016),5:1.
43. Brown, S. D. (1995**). Chemical Systems under Indirect Observation: Latent Properties and Chemometrics**. Applied Spectroscopy, 49(12), 14A–31A.
44. Efron, B. (1979). **Bootstrap Methods: Another Look at the Jackknife.**The Annals of Statistics, 7(1), 1–26.
45. Pires,B.O.L. **Métodos de aprendizagem de máquina em química analítica: Floresta Randômica aplicada na avaliação de petróleo.**2019.141f. Tese(Doutorado em Química)-Universidade Federal do Espirito Santo,Vitoria.
46. Correr,C.J;Cordeiro,J,Gasparetto,J;Peralta.Zamora,Cco&Pontarolo,R.(2005).**Determinação de Ácido Kógico em Produtos Farmacêuticos por Espectroscopia Uv-vis e Processo de Calibração Multivariada.**Acta Farm. Bonaerense **24** (3): 416-20
47. Souza,A.M;Breitkreitz,M.C;Roberto,P.F;Rodrigues,J.J.R&Poppi,R.J.(2013).**Experimento Didático De Quimiometria Para Calibração Multivariada Na Determinação De Paracetamol Em Comprimidos Comerciais Utilizando Espectroscopia No Infravermelho Próximo: Um Tutorial, Parte II.**Quim. Nova, Vol. 36, No. 7, 1057-1065.
48. Garcia, M. H. F., Farias, S. B., & Ferreira, B. G. (2004). **Determinação quantitativa da concentração de silicone em antiespumantes por espectroscopia FT-IR / ATR e calibração multivariada.** Polímeros, 14(5), 322–325.

49. Martins, J. P. A., & Ferreira, M. M. C. (2013). **QSAR modeling: um novo pacote computacional open source para gerar e validar modelos QSAR.** Química Nova, 36(4), 554–560.
50. Gowen, A. A., Downey, G., Esquerre, C., & O'Donnell, C. P. (2010). **Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients**. Journal of Chemometrics, 25(7), 375–381.
51. Girard, D.A. (1989**). A fast 'Monte-Carlo cross-validation' procedure for large least squares problems with noisy data**. Numerische Mathematik, 56(1), 1–23.
52. Little, M. A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D. C., & Kording, K. P. (2017). **Using and understanding cross-validation strategies**. Perspectives on Saeb et al. GigaScience, 6(5).
53. Xu, Y., & Goodacre, R. (2018). On Splitting Training and Validation Set: **A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning.** Journal of Analysis and Testing, 2(3), 249–262.
54. Chen, D., Cai, W., & Shao, X. (2006). **An adaptive strategy for selecting representative calibration samples in the continuous wavelet domain for near-infrared spectral analysis.** Analytical and Bioanalytical Chemistry, 387(3), 1041–1048.
55. Ferreira,M.M.C.Quimiometria**:conceitos,métodoseaplicações.**Campinas,SP:Editora da Unicamp,2015,493 p.ISBN:978-85-268-1471-4.
56. Næs, T. (1989). **Leverage and influence measures for principal component regression**. Chemometrics and Intelligent Laboratory Systems, 5(2), 155–168.
57. Galvão, R. K. H., Araújo, M. C. U., Martins, M. do N., José, G. E., Pontes, M. J. C., Silva, E. C., & Saldanha, T. C. B. (2006). **An application of subagging for the improvement of prediction accuracy of multivariate calibration models**. Chemometrics and Intelligent Laboratory Systems, 81(1), 60–67.
58. Dias, C.T., and W.J. Krzanowski. 2003. **Model selection and cross validation in additive main effects and multiplicative interaction model**s. Crop Sci. 43:865–873.
59. Galvão, R. K. H., Araújo, M. C. U., Fragoso, W. D., Silva, E. C., José, G. E., Soares, S. F. C., et al. (2008**). A variable elimination method to improve the**

parsimony of MLR models using the successive projections algorithm. Chemometrics and Intelligent Laboratory Systems, 92(1), 83–91.

60. Daszykowski, M., Walczak, B., & Massart, D. L. (2002). **Representative subset selection.** Analytica Chimica Acta, 468(1), 91–103.
61. Denham, M. C. (2000). **Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction**. Journal of Chemometrics, 14(4), 351–361.
62. Faber, N. M., & Rajkó, R. (2007**). How to avoid over-fitting in multivariate calibration—The conventional validation approach and an alternative**. Analytica Chimica Acta, 595(1-2), 98–106. doi:10.1016/j.aca.2007.05.030
63. Galvão, R. K. H., Araújo, M. C. U., Silva, E. C., José, G. E., Soares, S. F. C., & Paiva, H. M. (2007). **Cross-validation for the selection of spectral variables using the successive projections algorithm.** Journal of the Brazilian Chemical Society, 18(8), 1580–1584.
64. Rajer-Kanduč, K., Zupan, J., & Majcen, N. (2003). **Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment**. Chemometrics and Intelligent Laboratory Systems, 65(2), 221
65. Kowalski, B. R., & Seasholtz, M. B. (1991). **Recent developments in multivariate calibration.** Journal of Chemometrics, 5(3), 129–145.
66. Joe Qin, S., Li, W., & Henry Yue, H. (1999). **Recursive PCA for Adaptive Process Monitoring.** IFAC Proceedings Volumes, 32(2), 6686–6691.
67. Messick, N. J., Kalivas, J. H., & Lang, P. M. (1997). **Selecting Factors for Partial Least Squares**. Microchemical Journal, 55(2), 200–207
68. Naes, T., & Mevik, B.-H. (2001**). Understanding the collinearity problem in regression and discriminant analysis**. Journal of Chemometrics, 15(4), 413–426
69. Mevik BH, Wehrens R (2007). "**The pls Package: Principal Component and Partial Least Squares Regression in R**." Journal of Statistical Software, 18(2).
70. Seipel, H. A., & Kalivas, J. H. (2004). **Effective rank for multivariate calibration methods**. Journal of Chemometrics, 18(6), 306–311
71. Zhu, X.R., Shan, Y., Li, G.Y., Huang, A.M., and Zhang, Z.Y. (2009) **Prediction of wood property in Chinese Fir based on visible/near-infrared**

spectroscopy and least square-support vector machine. Spectrochim. Acta A, 74 (2): 344–348.

72. De Araújo Gomes, A., Galvão, R. K. H., de Araújo, M. C. U., Véras, G., & da Silva, E. C. (2013). **The successive projections algorithm for interval selection in PLS**. Microchemical Journal, 110, 202–208.
73. K. Beebe, R. Pell. M. B. Seasholtz, John Wiley & Sons (1998)CHEMOMETRICS A Practical Guide.
74. M. Forina, S. Lanteri, M. Casale, M.C. Cerrato Oliveros.**Stepwise orthogonalization of predictors in classification and regression techniques.**Chemometrics and Intelligent Laboratory Systems 87 (2007) 252–261.
75. Silva, W. R., Paula, L. C. M., Soares, A. S., Coelho, C. J.**Algoritmo do Morcego para Seleção de Variáveis em Calibração Multivariada.** ARITHMOS, Goiânia, v. 1, n.1, p. 13-17, jan./jun. 2019.
76. Antônio,C.C.J.**Algoritmo Evolutivo Multi-Objetivode Tabelas para Seleção de Variáveis em Calibração Multivariada.**Dissertação de Mestrado, II-UFG,Goiânia2014.
77. Adonias, H,D.F.**Desenvolvimento de técnicas quimiométricas de compressão de dados e de redução de ruído instrumental aplicadas a óleo diesel e madeira de eucalipto usando espectroscopia NIR".**Tese de Doutorado,IQ-Unicamp, Campinas,2007
78. Araujo, A.G.**Algoritmo de Projeções Sucessivas Aplicado à Seleção de Variáveis em Regressão PLS.** Dissertação de Mestrado, DQ-UFPB, João Pessoa, 2012.
79. Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). **The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems, 57(2), 65–73.**
80. A.M. DeSouza, M.C. Breitkreitz, P.R. Filgueiras, J.J.R. Rohwedder, R.J. Poppi. **Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II.**Quimica Nova, 36 (2013), pp. 1057-1065

81. Valderrama, Patrícia. **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, Campinas: UNICAMP, 2005.

82. FRANCO,Camila Manara**. Determinação de hidrocarbonetos majoritarios presentes no gás natural utilizando espectroscopia no infravermelho proxímo e calibração multivariada.**2008.110p.dissertação(mestrado- Universidade Estadual de Campinas, Instituto de Química, Campinas,SP.

83. Almeida,P.B.A;2013.**Uso da espectroscopia NIR e Calibração Multivariada para prospecção de olaginosas quanto as suas caracteristicas de óleo e proteína**.Dissertação(mestrado)- Universidade Estadual da Paraíba/Embrapa Algodão,Campina Grande, PB 48p.

84. Tristão, J.C.R. **Emprego de espectrocopia no infravemelho e metodos quimiometricos para a identificação e quantificação de petroleos a partir de misturas de frações de diesel.**2009.122p.Tese(doutorado- Universidade Federal do Espirito Santo, Vitoria, ES.

85. Pasquini, C. (2018). **Near infrared spectroscopy: A mature analytical technique with new perspectives – A review**. Analytica Chimica Acta, 1026, 8–36.

86. Pedro, A.M.K(2004). **Determinação simultânea e não-destrutiva de sólidos totais e solúveis, licopeno e beta-caroteno em produtos de tomate por espectroscopia no infravermelho próximo utilizando calibração multivariada**. Dissertação (mestrado em físico-química)- Universidade de Campinas, Campinas, São Paulo.

87. PAULA, L. C. M. D. **Paralelização de algoritmos aps e firefly para seleção de variáveis em problemas de calibração multivariada**. Master's thesis, Universidade Federal de GOiás - Instituto de Informática, 2014.

88. Meira M, Quintella CM, Ferrer TM, Silva HRG, Guimarães AK, Santos MA, et al. 472 **Identificação de adulteração de biocombustível por adição de óleo residual ao 473 diesel por espectrofluorimetria total 3D e análise das componentes principais**. 474 Química Nova 2011;34:621–4.

APÊNDICE

Nesse apêndice, é apresentado o código-fonte do programa – escrito e executável somente no ambiente computacional Matlab – do algoritmo desenvolvido para a estratégia proposta no presente trabalho.

Apêndice A - Validação cruzada com conjunto separado fixo

```
function[e,yhat]=validation_fixed_one(Xcal,ycal,Xfix,yfix,v
ar_sel)


%[yhat,e]=validation_fixed_one(Xcal,ycal,Xfix,yfix,var_sel)
--> Validation with a separate set fixed
% [yhat,e] = validation_fixed_one(Xcal,ycal,[],[],var_sel)
--> Cross-validation leave one out
%
Nfix = size(Xfix,1); % número de amostras que serão
mantidas no conjunto em cal
N = size(Xcal,1);
g = floor(0.1*N);   % número de amostras que saem no
processo cross_validation
yhat = zeros(N,1); % Setting the proper dimensions of yhat
fori = 1:N
% Removing the ith object from the calibration set cal =
[[1:i-1] [i+1:N]];

X = [Xcal(cal,var_sel); Xfix(:,var_sel)];
y = [ycal(cal); yfix];
xtest = Xcal(i,var_sel);
X_ones = [ones(N+Nfix-1,1) X];
    b = X_ones\y; % MLR with offset term (b0)
yhat(i) = [1 xtest]*b; % Prediction for the ith object

end
e = ycal - yhat; % Cross-validation error
```

Apêndice A - Métricas de Validação

```
function[PRESS,RMSEP,SDV,BIAS,r]=validation_metrics_v2p1(Xc
al,ycal,Xval,yval,var_se)

%[PRESSV,RMSEPV,SDV,BIASV,rV]=validation_metrics_v2p1(Xcal,
ycal,Xval,yval,var_sel) --> Validation with a separate set
%[PRESSV,RMSEPV,SDV,BIASV,rV]=validation_metrics_v2p1(Xcal,
ycal,Xval,[],var_sel)  -->Validation with a separate set
with no reference values for y
```

```
%[PRESSCV,RMSEPCV,SDCV,BIASCV,rCV]=validation_metrics_v2p1(
Xcal,ycal,[],[],var_sel) --> Cross-validation
%
% Record of modifications:
% 08 Sept 2011 (v2):
% Allows for the use of yval = [ ] (no reference values)
% get(h,'XLim') modified to get(h,'xlim');
% 09 Sept 2011 (v2p1):
%    Use    of    prediction_uncertainty.m    instead    of
statistical_prediction_error.m

if size(Xval,1) > 0 % Validation with a separate set
    y = yval;
else % Cross-validation
    y = ycal;
end

[yhat,e] = validation_v2(Xcal,ycal,Xval,yval,var_sel);

if (~isempty(e))
    PRESS = e'*e;
    N = length(e);
    RMSEP = sqrt(PRESS/N);
    BIAS = mean(e);
ec = e - BIAS; % Mean-centered error values
SDV = sqrt(ec'*ec/(N - 1));
yhat_as = (yhat - mean(yhat))/std(yhat); % Autoscaling
y_as = (y - mean(y))/std(y); % Autoscaling
r = (yhat_as'*y_as)/(N-1);
else
    PRESS = [];
    RMSEP = [];
    SDV = [];
    BIAS = [];
    r = [];
end

% Prediction uncertainty (one-sigma)
pred_unc = prediction_uncertainty(Xcal,ycal,Xval,var_sel)

if (~isempty(e))
    % Plot of Predicted vs Reference values
figure, hold on, grid
errorbar(y,yhat,pred_unc,'o')
xlabel('Reference y value'),ylabel('Predicted y value')
h = gca; XLim = get(h,'xlim');
h = line(XLim,Xlim);
title(['PRESS   =   '   num2str(PRESS)   ',   RMSEP   =   '
num2str(RMSEP)  ',  SDV  =  '  num2str(SDV)  ',  BIAS  =  '
num2str(BIAS) ', r = ' num2str(r)])
else
```

```
% Plot of predicted values with +/- one-sigma bars
figure, hold on, grid
errorbar([1:length(yhat)],yhat,pred_unc,'o')
xlabel('Sample'),ylabel('Predicted y value')
end
```

Apêndice A - Cálculo SPA para os conjuntos de dados

```
clear,clc,close all

Nfix = 16; %MilhoNfix = 12; TrigoNfix = 16.
g = 4; % Milho g = 3; Trigo g = 4;
%txp = 0.30; % taxa usada no conjunto milho (corresponde a
20 amostras de predição).
txp = 0.33; % % taxa usada no conjunto Trigo (corresponde a
30 amostras de predição).

Nmax = 59; % Milho Nmax = 39; Trigo Nmax = 59.

load('D:\Trabalho\Sófacles\Orientações\João\Dados\X.mat') %
local onde está os dados trigo
XNIR = X(:,351:1050);


%load('D:\Trabalho\Sófacles\Orientações\João\Dados\Milho\mi
lho1.MAT')

% 701 = 'Moistrure';
% 702 = 'Oil';
% 703 = 'Protein';
% 704 = 'Starch';


%XNIR = AllSamples_AllVariables(:,1:700);
%Y = AllSamples_AllVariables(:,701); % Umidade


Xder = derivadaSG(XNIR,1,2,21); % preprocessamento milho

Ntot = size(Xder,1);
[mfix,dminmax] = ksxy(Xder,Y,Nfix); % Milho 12 amostras com
o ksxy
mdif = setdiff(1:Ntot,mfix);

% Separando as amostras fixas
Xfix = [Xder(mfix,:)];
Yfix = [Y(mfix,:)];

% amostras restantes
Xrest = [Xder(mdif,:)];
```

```
Yrest = [Y(mdif,:)];
Nrest = size(Xrest,1);


[mtot,dminmax] = ksxy(Xrest,Yrest,Nrest); % Ordenando as
amostrasrestantes
NP = round(txp*Nrest);

mpred = mtot([round(2:(1/txp):(NP*(1/txp)))]);
m = setdiff(1:Nrest,mpred);

Xpred = Xrest(mpred,:);
Ypred = Yrest(mpred,:);
Xcal = Xrest(m,:);
Ycal = Yrest(m,:);

cont = 1;
fori = 0:g:Nfix
 [var_sel_fix{cont},var_sel_phase2_fix{cont}]              =
spa_fix_one([Xcal;                Xfix((i+1):Nfix,:)],[Ycal;
Yfix((i+1):Nfix,:)],Xfix(1:i,:),Yfix(1:i,:),1,39,0);      %
trigo

[PRESS_fix(cont),RMSEP_fix(cont),SDV_fix(cont),BIAS_fix(con
t),r_fix(cont)]       =       validation_metrics_v2p1([Xcal;
Xfix],[Ycal; Yfix],Xpred,Ypred,var_sel_fix{cont});
cont = cont+1;
end
```

Apêndice A- Validação SPA com amostras fixas e figuras de mérito

```
function              [var_sel,var_sel_phase2]              =
spa_fix_one(Xcal,ycal,Xfix,yfix,m_min,m_max,autoscaling)

%                  [var_sel,var_sel_phase2]                 =
spa_fix_one(Xcal,ycal,Xfix,yfix,m_min,m_max,autoscaling) --
> Validation with fixed sample
%
% If m_min = [], the default value m_min = 1 is employed
% If m_max = [], the default values m_max = min(N-1,K)
(validation with a separate set)
% or min(N-2,K) (cross-validation) are employed.
% autoscaling --> 1 (yes) or 0 (no)

autoscaling
if ((autoscaling ~= 1) & (autoscaling ~= 0))
error('Please choose whether or not to use autoscaling.')
end
```

```
N = size(Xcal,1); % Number of calibration objects
K = size(Xcal,2); % Total number of variables

if length(m_min) == 0, m_min = 1; end
if length(m_max) == 0,
if size(Xfix,1) > 0
m_max = min(N-1,K);
else
m_max = min(N-2,K);
end
end

ifm_max> min(N-1,K)
error('m_max is too large !');
end

% Phase 1: Projection operations for the selection of
candidate subsets

% The projections are applied to the columns of Xcal after
% mean-centering and (optional) autoscaling

ifautoscaling == 1
normalization_factor = std(Xcal);
else
normalization_factor = ones(1,K);
end

for k = 1:K
    x = Xcal(:,k);
Xcaln(:,k) = (x - mean(x)) / normalization_factor(k);
end

SEL = zeros(m_max,K);

h = waitbar(0,'Phase 1 (Projections). Please wait...');
loopStart = now;
tic
for k = 1:K
SEL(:,k) = projections_qr(Xcaln,k,m_max);
loopEnd = loopStart + (now-loopStart)*K/k;
waitbar(k/K,h,['Phase 1 ETC: ' datestr(loopEnd)]);
end
toc
close(h);

disp('Phase 1 (projections) completed !')

% Phase 2: Evaluation of the candidate subsets according to
the PRESS criterion
```

```
PRESS = Inf*ones(m_max,K);
h = waitbar(0,'Phase 2 (Evaluation of variable subsets).
Please wait...');
warning off MATLAB:singularMatrix
warning off MATLAB:nearlySingularMatrix
loopStart = now;
tic
for k = 1:K
for m = m_min:m_max
var_sel = SEL(1:m,k);
        [e,yhat]                                                =
validation_fixed_one(Xcal,ycal,Xfix,yfix,var_sel);
PRESS(m,k) = e'*e;
end
loopEnd = loopStart + (now-loopStart)*K/k;
waitbar(k/K,h,['Phase 2 ETC: ' datestr(loopEnd)]);
end
close(h);
warning on MATLAB:singularMatrix
warning on MATLAB:nearlySingularMatrix

[PRESSmin,m_sel] = min(PRESS);
[dummy,k_sel] = min(PRESSmin);

var_sel_phase2 = SEL(1:m_sel(k_sel),k_sel);
toc
disp('Phase 2 (evaluation of variable subsets) completed
!')

% Phase 3: Final elimination of variables
tic
% Step 3.1: Calculation of the relevance index
Xcal2 = [ones(N,1) Xcal(:,var_sel_phase2)];
b = Xcal2\ycal; % MLR with intercept term
std_deviation = std(Xcal2);
relev = abs(b.*std_deviation');
relev = relev(2:end); % The intercept term is always
included
% Sorts the selected variables in decreasing order of
"relevance"
[dummy,index_increasing_relev] = sort(relev); % Increasing
order
index_decreasing_relev = index_increasing_relev(end:-1:1);
% Decreasing order

% Step 3.2: Calculation of PRESS values
fori = 1:length(var_sel_phase2)
    [e,yhat]                                                    =
validation_fixed_one(Xcal,ycal,Xfix,yfix,var_sel_phase2(ind
ex_decreasing_relev(1:i)) );
```

```
PRESS_scree(i) = e'*e;
end
RMSEP_scree = sqrt(PRESS_scree/length(e));
figure, grid, hold on
plot(RMSEP_scree)
xlabel('Number of variables included in the
model'),ylabel('RMSE')

% Step 3.3: F-test criterion
PRESS_scree_min = min(PRESS_scree);
alpha = 0.25;
dof = length(e); % Number of degrees of freedom
fcrit = finv(1-alpha,dof,dof); % Critical F-value
PRESS_crit = PRESS_scree_min*fcrit;
% Finds the minimum number of variables for which
PRESS_scree
% is not significantly larger than PRESS_scree_min
i_crit = min(find(PRESS_scree<PRESS_crit));
i_crit = max(m_min,i_crit); % The number of selected
variables must be at least m_min

var_sel = var_sel_phase2(index_decreasing_relev(1:i_crit)
);
title(['Final number of selected variables: '
num2str(length(var_sel)) ' (RMSE = '
num2str(RMSEP_scree(i_crit)) ')'])
toc
% Indicates the selected point on the scree plot
plot(i_crit,RMSEP_scree(i_crit),'s')

disp('Phase 3 (final elimination of variables) completed
!')
% Presents the selected variables
% in the first object of the calibration set
figure,plot(Xcal(1,:));hold,grid
plot(var_sel,Xcal(1,var_sel),'s')
legend('First calibration object','Selected variables')
xlabel('Variable index')
```