UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

LUCIANA DOS SANTOS CANOVA

UMA VERSÃO APRIMORADA DO ALGORITMO DE PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM REGRESSÃO LINEAR MÚLTIPLA

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Química

Porto Alegre, outubro/2023.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

LUCIANA DOS SANTOS CANOVA

UMA VERSÃO APRIMORADA DO ALGORITMO DE PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM REGRESSÃO LINEAR MÚLTIPLA

Tese apresentada como requisito parcial para a
obtenção do grau de Mestre em Química

Prof. Dr. Adriano de Araújo Gomes
Orientador

Porto Alegre, outubro/2023.

CIP - Catalogação na Publicação

Canova, Luciana dos Santos
UMA VERSÃO APRIMORADA DO ALGORITMO DE PROJEÇÕES SUCESSIVAS PARA SELEÇÃO DE VARIÁVEIS EM REGRESSÃO LINEAR MÚLTIPLA / Luciana dos Santos Canova. -- 2023.
90 f.
Orientador: Adriano de Araújo Gomes.

Dissertação (Mestrado) -- Universidade Federal do Rio Grande do Sul, Instituto de Química, Programa de Pós-Graduação em Química, Porto Alegre, BR-RS, 2023.

1. Seleção de Variáveis. 2. Algoritmo das Projeções Sucessivas. 3. Regressão Linear Múltipla. 4. Regressão por Mínimos Quadrados Parciais. 5. Espectrometria no Infravermelho próximo. I. Gomes, Adriano de Araújo, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

A presente dissertação foi realizada inteiramente pelo autor, exceto as colaborações as quais serão devidamente citadas nos agradecimentos, no período entre Agosto de 2020 e Outubro de 2023, no Instituto de Química da Universidade Federal do Rio Grande do Sul, sob orientação do Professor Doutor Adriano de Araújo Gomes. A dissertação foi julgada adequada para a obtenção do título de Mestre em Química pela seguinte banca examinadora:

Comissão Examinadora:

Prof. Dr. Paulo Henrique G.s Dias Diniz
Universidade Federal do Oeste da Bahia (UFOB)
(Membro Externo)

Prof. Dr. Marco Flôres Ferrão
Universidade Federal do Rio Grande do Sul (UFRGS)
(Membro Interno)

Prof. Dr. Vladimir Lavayen
Universidade Federal do Rio Grande do Sul (UFRGS)
(Membro Interno)

Prof. Dr. Adriano de Araújo Gomes
Universidade Federal do Rio Grande do Sul (UFRGS)
Prof. Dr. Orientador

Luciana dos Santos Canova
Universidade Federal do Rio Grande do Sul (UFRGS)
Aluno

Dedico este trabalho às pessoas mais importantes da minha vida e que sempre estiveram ao meu lado, me apoiando em todos os momentos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me abençoar com o dom da vida e muita saúde para toda a minha família nestes tempos difíceis de pandemia que estamos enfrentando.

Agradeço aos meus pais Altair e Maribel, por todo apoio e incentivo ao longo de toda a minha vida e por serem os melhores exemplos de vida, de luta, de companheirismo e de amor. À minha avó Terezinha, com sua admirável lucidez aos 92 anos, és fonte de inspiração na minha vida.

Ao meu esposo Felipe, pela compreensão e incentivo. Obrigada por estar sempre ao meu lado e me oferecer todo o suporte necessário para que eu pudesse seguir na condução da minha formação acadêmica da forma mais leve possível.

Aos meus irmãos, Rafael, Thaís e Daniela, por serem presentes na minha vida e por me proporcionarem momentos de alegria com meus amados sobrinhos. Em especial a eles: Arthur, Matheus, Sebastian, Letícia, Melissa, Diana, Isadora e Otto por serem uma válvula de escape nos finais de semana, após uma semana estressante e cansativa.

Ao professor Dr. Adriano de Araújo Gomes, pela orientação e apoio no desenvolvimento deste trabalho. Você é motivo de orgulho para a educação, com toda a sua dedicação e inclusão de atividades inovadoras nas aulas, bem como a disponibilidade de atender aos seus alunos da melhor forma possível.

Aos amigos e amigas que me acompanham desde o início, aos que estão comigo até hoje e aos que por um motivo ou outro estão hoje distantes, por sempre me ensinarem a lidar com as situações adversas e me apoiarem em todos os momentos.

À BAT Brazil, em especial ao time do laboratório de Quimiometria e Quimiosensorial, pela oportunidade diária de desenvolvimento profissional e pessoal. Obrigada por toda a ajuda e compreensão durante este período.

Por último, mas não menos importante, à minha pequena Giovana, que chegou na minha vida junto à finalização deste trabalho, trazendo novos desafios ao que antes já parecia complicado, mas que, ao mesmo tempo, com sua doçura e amor trouxe um novo propósito à esta entrega. Te amo, filha.

PRODUÇÃO CIENTÍFICA GERADA POR ESTE TRABALHO

ARTIGO COMPLETO PUBLICADO EM PERIÓDICO

1. Canova, L. S.; Vallese, F. D.; Pistonesi, M. F.; Gomes, A. A. An improved successive projections algorithm version to variable selection in multiple linear regression. Analytica Chimica Acta, 1274, 341560, 2023. https://doi.org/10.1016/j.aca.2023.341560.

RESUMO

O Algoritmo de Projeções Sucessivas (SPA), também conhecido em inglês como Successive Projection Algorithm, foi desenvolvido com o propósito de selecionar um subconjunto de variáveis informativas e minimamente redundantes para a construção de modelos de regressões lineares múltiplas (do inglês, Multiple Linear Regression - MLR). Esse método visa minimizar o impacto da multicolinearidade, que é comumente presente em dados instrumentais, ao mesmo tempo em que alcançar uma melhor acurácia na predição. A combinação do SPA com o MLR, como uma abordagem de seleção variável/calibração multivariada, resultou no método SPA-MLR, o qual tem sido relatado na literatura como capaz de produzir modelos com boa capacidade de predição em comparação com os modelos convencionais de "espectro completo" via mínimos quadrados parciais (PLS), em alguns casos. Neste trabalho, é proposta a adição de uma etapa de filtro (*f*) à versão atual do algoritmo SPA, a fim de reduzir o número de variáveis não informativas antes da fase de projeção. Essa adição auxilia o algoritmo na seleção das melhores variáveis nas etapas subsequentes. O algoritmo resultante, denominado de *f*SPA-MLR, é avaliado em dois estudos de caso que envolvem dados de espectroscopia no infravermelho próximo: (i) quantificação do ingrediente farmacêutico ativo (IFA) em comprimidos e (ii) quantificação de biodiesel em misturas diesel/biodiesel. Comparados com o método PLS, os modelos *f*SPA-MLR demonstram desempenho semelhante ou superior. Além disso, os modelos *f*SPA-MLR superam o SPA-MLR original tanto na validação cruzada quanto na predição externa. Independentemente do algoritmo de pré-processamento testado, incluindo primeira derivada da função Savitzky-Golay (SG) e a função Standard Normal Variate (SNV), ou mesmo em dados de espectros brutos, os modelos *f*SPA-MLR oferecem resultados superiores.

Palavras-chave: Seleção de Variáveis, Algoritmo das Projeções Sucessivas, Regressão Linear Múltipla, Regressão por Mínimos Quadrados Parciais, Espectrometria no Infravermelho próximo.

ABSTRACT

The Successive Projection Algorithm (SPA), also known in Portuguese as APS, was initially developed with the purpose of selecting a subset of informative and minimally redundant variables for the construction of multiple linear regression models (MLR). This method aims to minimize the impact of multicollinearity, which is commonly present in instrumental data, while achieving better forecast accuracy. The combination of SPA with MLR, as a variable selection/multivariate calibration approach, resulted in SPA-MLR method, which has been reported in literature as capable of producing models with good predictive ability compared to conventional models of "full spectrum" via Partial Least Squares (PLS) in some cases. In this work, it is proposed to add a filter step (*f*) to the current version of SPA algorithm, to reduce the number of non-informative variables before projection phase. This addition assists the algorithm in selecting the best variables in subsequent steps. The resulting algorithm, called *f*SPA-MLR, is evaluated in two case studies involving near-infrared spectroscopy data: (i) quantification of the active pharmaceutical ingredient (IFA), also known in English as API, in tablets and (ii) quantification of biodiesel in diesel/biodiesel blends. Compared with the PLS method, *f*SPA-MLR models demonstrate similar or superior performance. Furthermore, *f*SPA-MLR models outperform the original SPA-MLR in both cross-validation and external prediction. Regardless of the tested pre-processing algorithm, including Savitzky-Golay (SG) First Derivative and Standard Normal Variate (SNV), or even on raw spectral data, *f*SPA-MLR models deliver superior results.

Keywords: Variable selection, Successive projections algorithm, Multilinear regression, Partial least squares, NIR spectrometry.

LISTA DE FIGURAS

Figura 1. Relação entre duas variáveis, onde as coordenadas do ponto referem-se aos valores de cada uma das variáveis relacionadas. 21
Figura 2. Simulação das etapas do algoritmo de recozimento simulado. 30
Figura 3. Etapa 1 AG – Codificação de variáveis, utilizando respostas binárias. 31
Figura 4. Etapa 2 AG – Iniciação da população. 32
Figura 5. Etapa 4 AG – Reprodução da população através dos mecanismos de (a) seleção e cópia e (b) cruzamento. 33
Figura 6. Etapa 5 AG – Mutação. 33
Figura 7. Espectro com modelo iPLS aplicado, relacionando os intervalos com o RMSECV. 35
Figura 8. Exemplo de Eliminação de Variáveis não-informativas. 37
Figura 9. (a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processamento pela função SNV das 155 amostras de comprimidos farmacêuticos do conjunto de calibração. 47
Figura 10. Valores de J calculados para (a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processamento pela função SNV das 155 amostras de comprimidos farmacêuticos do conjunto de calibração. 49
Figura 11. (a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de ingredientes ativos, do conjunto de calibração (círculos azuis indicam as variáveis selecionadas em SPA-MLR e os quadrados verdes indicam as variáveis selecionadas em fSPA-MLR). 51
Figura 12. Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 53
Figura 13. Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros pré-processados com 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 53
Figura 14. Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS..... 54

Figura 15. Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de quantidade de ingrediente ativo. 55
Figura 16. Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS para o parâmetro de quantidade de ingrediente ativo. 55
Figura 17. Erros residuais dos modelos para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS para o parâmetro de quantidade de ingrediente ativo. 56
Figura 18. (a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de peso dos comprimidos, do conjunto de calibração (círculos azuis indicam as variáves selecionadas em SPA-MLR e os quadrados verdes indicam as variáveis selecionadas em fSPA-MLR). 57
Figura 19. Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 59
Figura 20. Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros pré-processados com 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 60
Figura 21. Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS..... 60
Figura 22. Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de peso dos comprimidos. 61
Figura 23. Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de peso do comprimido..... 62
Figura 24. Erros residuais dos modelos para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de peso dos comprimidos..... 62
Figura 25. (a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de dureza do

comprimido, do conjunto de calibração (círculos azuis indicam as variáves selecionadas em SPA-MLR e os quadrados verdes indicam as variáveis selecionadas em fSPA-MLR). 64
Figura 26. Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 66
Figura 27. Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros pré-processados com 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 67
Figura 28. Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS..... 67
Figura 29. Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de dureza dos comprimidos. 68
Figura 30. Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de dureza do comprimido..... 69
Figura 31. Erros residuais dos modelos para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, para o parâmetro de dureza dos comprimidos. . 69
Figura 32. (a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da função SG e (c) com pré-processamento na função SNV das 40 amostras de mistura diesel/biodiesel do conjunto de calibração. 71
Figura 33. Valor de J calculado para (a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da função SG e (c) com pré-processamento na função SNV das 40 amostras de mistura diesel/biodiesel do conjunto de calibração. 72
Figura 34. Estrutura Molecular do Diesel e Biodiesel [70]..... 73
Figura 35. (a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da funçãoSG e (c) com pré-processamento na função SNV das variáveis selecionadas para construir o modelo MLR a partir do conjunto de calibração (círculos azuis indicam as variáveis selecionadas em SPA-MLR; quadrados verdes indicam as variáveis selecionadas em fSPA-MLR)..... 74
Figura 36. Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 77

Figura 37. Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros pré-processados com 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 77
Figura 38. Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS. 78
Figura 39. Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel. 79
Figura 40. Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel. 79
Figura 41. Erros residuais dos modelos para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel. 80

LISTA DE TABELAS

Tabela 1. Aplicações do SPA como método de seleção de variáveis..... 41
Tabela 2. Resultados de modelos de predição de quantidade de ingrediente ativo em amostras de comprimidos farmacêuticos..... 52
Tabela 3. Resultados de modelos de predição do peso de cada comprimido em amostras de comprimidos farmacêuticos..... 58
Tabela 4. Resultados de modelos de predição da dureza de cada comprimido em amostras de comprimidos farmacêuticos..... 65
Tabela 5. Resultados dos modelos de predição da porcentagem de biodiesel no diesel (% m/m) em amostras de mistura diesel/biodiesel..... 76

SÍMBOLOS E ABREVIATURAS

AG: Algoritmos Genéticos (Genetic Algorithm - GA)

SPA: Algoritmo de Projeções Sucessivas

CF: Colônia de Formigas (Ant Colony - AC)

IDRC: International Diffuse Reflectance Conference

IFA: Ingrediente Farmacêutico Ativo (Active Pharmaceutical Ingredient – API)

iPLS: Mínimos Quadrados Parciais de Intervalo (Interval Partial Least Squares)

Jcost: Função de custo

MLR: Regressão Linear Múltipla (Multiple Linear Regression)

NIR: Espectroscopia no Infravermelho Próximo (Near-Infrared Spectroscopy)

PCR: Regressão por Componentes Principais (Principal Component Regression)

PLS: Mínimos Quadrados Parciais (Partial Least Squares)

RMN: Ressonância Magnética Nuclear

RMSECV: Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada (Root-Mean-Square Error of Cross-Validation)

RMSEP: Raiz Quadrada do Erro Quadrático Médio de Predição (Root-Mean-Square Error of Prediction)

RMSEV: Raiz Quadrada do Erro Quadrático Médio de Validação (Root-Mean-Square Error of Validation)

RS: Recozimento Simulado (Simulated Annealing - SA)

SG: Primeira derivada Savitzky-Golay

SNV: Standard Normal Variate

SPA: Successive Projections Algorithm

ST-PLS: Soft-threshold PLS

UV: Ultravioleta

UVE: Eliminação de Variável não informativa (Uninformative Variable Elimination)

VIF: Fator de Inflação da Variância (Variance Inflation Factor)

SUMÁRIO

1 Introdução 17

2 Objetivos 19

2.1 Objetivo Geral 19

2.2 Objetivos Específicos 19

3 Revisão Bibliográfica 20

3.1 Notação 20

3.2 Análises Multivariadas 20

3.2.1 Modelo de Regressão Linear Múltipla (MLR) 22

3.2.2 Modelo dos Mínimos Quadrados Parciais (PLS) 25

3.3 Algoritmos de Seleção de Variáveis 27

3.3.1 Recozimento Simulado (RS) 29

3.3.2 Algoritmos Genéticos 31

3.3.3 Mínimos Quadrados Parciais por Intervalo (iPLS) 34

3.3.4 Eliminação de Variáveis não-informativas 35

3.3.5 Algoritmo das Projeções Sucessivas (SPA) 37

4 Experimental 42

4.1 Modelo *f*SPA-MLR proposto 42

4.2 Procedimento quimiométrico e software 44

4.3 Estudos de caso 45

4.3.1 Estudo de caso I 45

4.3.2 Estudo de caso II 45

5 Resultados e Discussão 46

5.1 Estudo de caso I 46

5.1.1 Modelagem para predição de quantidade de ingrediente ativo 50

5.1.2 Modelagem para predição do peso do comprimido 56

5.1.3 Modelagem para predição da dureza do comprimido 63

5.2 Estudo de caso II 70

6 Conclusão 81

7 **Referências** 82

1 Introdução

A aplicação de técnicas quimiométricas e/ou inteligência artificial para tratamento de dados químicos tem despertado grande interesse nas indústrias, devido a sua ampla aplicabilidade na classificação de produtos ou predição de propriedades químicas, facilitando a tomada de decisão em diversas aplicações [1-2]. As informações sobre amostras obtidas por meio de instrumentação analítica podem ser modeladas para construir modelos matemáticos capazes de predizer um parâmetro de interesse ou uma classe à qual a amostra pertence. Porém, a construção desses modelos pode ser fortemente afetada pelo grande número de variáveis em análise, principalmente pelo fato de muitas delas apresentarem informações redundantes, não informativas ou com presença excessiva de ruído instrumental [1, 3-5].

O elevado número de variáveis associadas a informações redundantes pode reduzir a precisão dos modelos e produzir resultados pouco confiáveis, tornando o modelo insatisfatório para o objetivo proposto [1, 6-7]. A criação de modelos mais robustos, precisos e interpretáveis pode ser obtida selecionando um número reduzido de variáveis informativas [4, 7-9] como alternativa aos modelos baseados em estrutura latente. A redução da dimensionalidade, portanto, pode ser obtida identificando-se um subconjunto de variáveis que são mais significativas para a predição do proposto [10-11].

Ao longo dos anos, vários métodos foram criados para auxiliar nos processos de seleção de variáveis relevantes, como seleção à frente (*forward selection*) [12], eliminação reversa (*backward elimination*) [12-13], regressão passo a passo (*stepwise)* [12, 14], recozimento simulado (RS) [12, 15-16], algoritmos genéticos (AG) [16-21], mínimos quadrados parciais de intervalo (iPLS, *interval partial least squares)* [12, 22-23], eliminação de variável não informativa (UVE, *uninformative variable elimination*) [12, 15, 24-25], colônia de formigas (CF) [26-27], *Jack-Knife* [28], busca Tabu [29], estratégias que usam pesos de carregamento [7, 30-31] e algoritmo de projeções sucessivas (SPA) [12, 14, 23, 25, 32-33].

O algoritmo de projeções sucessivas, SPA (do inglês Successive Projections Algorithm), foi inicialmente projetado para melhorar o condicionamento da regressão linear múltipla (MLR, *multiple linear regression*), minimizando os efeitos da colinearidade no conjunto de dados de calibração [33-34] e reduzindo a

dimensionalidade. O SPA-MLR compreende três fases. A primeira fase corresponde ao estágio de projeções sucessivas que permite a criação de cadeias de variáveis minimamente redundantes. A etapa de projeção é do tipo à frente (*forward*), começando com a variável $\mathbf{x}_j$ e incluindo uma nova a cada iteração. O critério de projeção máxima garante subconjuntos de variáveis usadas para o propósito do cálculo da matriz inversa, mas até agora, nenhuma relação com a variável dependente **y** foi considerada [5, 35-36].

A fase 2 corresponde a um ciclo embutido que avalia as cadeias de variáveis geradas na fase 1 com base em uma função de custo (*Jcost*) em geral RMSEV ou RMSECV, que são descritas como Raiz Quadrada do Erro Quadrático Médio de Validação e de Validação Cruzada, respectivamente. Finalmente, o melhor subconjunto de variáveis que minimizam a função de custo é indicado como o subconjunto das variáveis selecionadas. A fase 3 visa remover quaisquer variáveis restantes no subconjunto selecionado com base na relevância dos coeficientes de regressão.

As cadeias de variáveis da fase 1 são criadas e avaliadas por modelos MLR; o número máximo de variáveis contidas em cada fragmento é limitado ao número de amostras no conjunto de calibração. Quando conjuntos de dados de alta dimensão são processados, a fragilidade dessa abordagem fica exposta, pois muitas variáveis informativas não têm a possibilidade de serem incluídas nas cadeias. Em aplicações recorrentes na literatura, os autores usam etapas manuais anteriores para excluir regiões do sinal do instrumento para que o SPA-MLR possa lançar resultados compatíveis com mínimos quadrados parciais (PLS, *partial least squares*) [5, 37-39]. No entanto, um algoritmo de seleção de variáveis deve apresentar mecanismos automáticos para selecionar o melhor subconjunto de variáveis possível, com robustez suficiente para descartar variáveis ruidosas e/ou não informativas e manter apenas informações úteis.

2 Objetivos

2.1 Objetivo Geral

O presente trabalho tem como objetivo principal a inclusão de uma etapa adicional ao SPA, denominada etapa de filtro (*f*SPA) a ser aplicada antes da fase de projeção (fase 1), também podendo ser denominada fase 0. Esta estratégia visa excluir previamente variáveis que não possuem correlação com a variável dependente **y**, bem como variáveis redundantes e ruidosas. Isso permitirá que apenas as variáveis mais promissoras sejam incluídas nas cadeias geradas na fase 1.

2.2 Objetivos Específicos

- Desenvolvimento do algoritmo para etapa adicional ao SPA, denominada etapa de filtro (*f*SPA);
- Aplicação do algoritmo desenvolvido em dois estudos de caso envolvendo a análise de comprimidos farmacêuticos e amostras de mistura de diesel/biodiesel por espectroscopia de infravermelho próximo (NIR), a fim de esclarecer e discutir os benefícios da incorporação da etapa de filtro proposta ao SPA;
- Realizar análise comparativa dos resultados obtidos nos estudos de casos utilizando SPA antes e após a inclusão da etapa do filtro, aplicados a um método de regressão múltipla linear (MLR), bem como à resultados obtidos por mínimos quadrados parciais (PLS).

3 Revisão Bibliográfica

3.1 Notação

A seguir, matrizes em letras maiúsculas em negrito, vetores em letras minúsculas em negrito e escalares em caracteres itálicos. O T sobrescrito indica uma transposição de um vetor ou matriz.

3.2 Análises Multivariadas

Considerando que concentração não é uma propriedade mensurável diretamente, a prática analítica consiste na aquisição de um sinal instrumental que possui relação, de preferência linear, com a concentração. Subsequentemente se estabelece uma relação matemática empírica entre sinal e concentração de amostras e padrões cujas concentrações são previamente conhecidas, denominada de calibração, sendo a calibração univariada, ou curva analítica, a mais comum, bem estabelecida e recorrente. Contudo, as técnicas instrumentais como as cromatográficas e espectroscópicas, por exemplo, são capazes de fornecer dados multivariados, o que significa que a análise de uma única amostra gera múltiplas informações, algumas úteis, já outras nem tanto. Os conjuntos de dados multivariados são comuns, mas nem sempre são analisados de maneira multivariada, onde o principal objetivo é relacionar as amostras e variáveis (ou componentes) de forma a identificar similaridades e diferenças dentro do conjunto de dados [3, 40-41].

Conforme descrito anteriormente, em técnicas instrumentais que geram conjuntos de dados multivariados, é importante entender como funciona a relação entre as variáveis geradas. Analisando-se graficamente, conforme mostrado na **Figura 1**, é possível observar a relação entre duas variáveis medidas. As coordenadas do ponto (x_i,y_i) mostrado na figura indicam os valores medidos para cada uma das variáveis x e y. O vetor que parte da origem até o ponto (x_i,y_i) é chamado de vetor dos dados. Sendo assim, os objetos que apresentarem vetores de dados semelhantes, apresentam propriedades semelhantes e

ficarão próximas umas das outras no espaço definido pelas variáveis, formando um agrupamento [42].

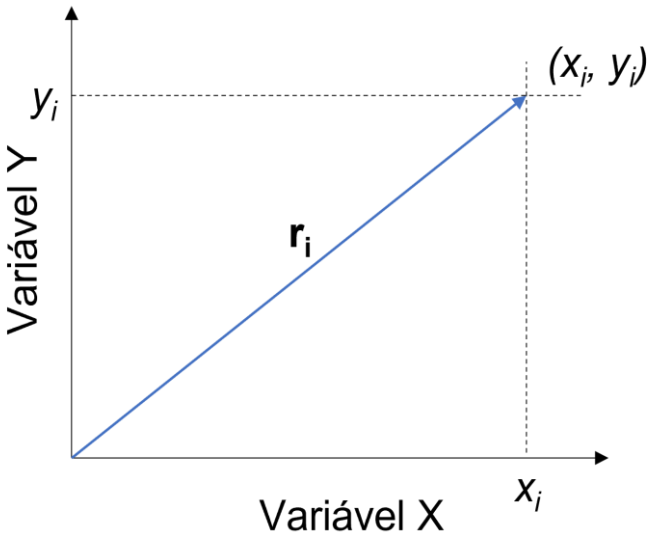

Figura 1. *Relação entre duas variáveis, onde as coordenadas do ponto referem-se aos valores de cada uma das variáveis relacionadas.*

Quando partimos para análises de três variáveis, uma representação gráfica é mais difícil, e não mais possível para quatro ou mais variáveis. Desta forma, análises computacionais são extremamente importantes para definir padrões, similaridades e diferenças entre os conjuntos de dados, através do uso de álgebra matricial [42].

A regressão linear múltipla é amplamente aplicada para resolver vários tipos de problemas em uma ou poucas análises de componentes; no entanto, em muitos casos, o envolvimento de múltiplas variáveis de interação dos analitos entre si, leva a erros de quantificação. Portanto, nesses casos, métodos de regressão com viés podem fornecer melhores resultados. Esses métodos são comumente conhecidos como métodos de calibração multivariada [3, 8].

As técnicas de regressão multivariada são amplamente utilizadas para calibração química multivariada (determinação de compostos químicos a partir de propriedades físicas medidas, por exemplo), calibração física multivariada (índice de octanagem e

viscosidade, por exemplo), calibração sensorial multivariada (notas sensoriais do painel através de características físicas e químicas medidas nas amostras), e para estudo da relação entre propriedade (tempo de retenção, coeficiente de partição, atividade biológica) e estrutura molecular [31].

Métodos de calibração multivariada, como regressão de mínimos quadrados parciais (do inglês, Partial Least Squares - PLS) e regressão de componente principal (do inglês, Principal Component Regression - PCR), são comumente aplicados ao predizer um ou vários parâmetros de um conjunto de dados multivariados. Esses métodos podem manipular conjuntos de dados mesmo quando o número de variáveis é muito maior que o número de amostras. No entanto, em algumas situações pode ser vantajoso reduzir o número de variáveis para, entre outras, obter (a) melhoria das predições do modelo, (b) uma melhor interpretação ou (c) menores custos de medição [7]. Entre os procedimentos regressão mais usada atualmente está Regressão Linear Múltipla (MLR), Regressão de Componentes Principais (PCR) e Regressão de Mínimos Quadrados Parciais (PLS). Enquanto os métodos lineares assumem que as relações entre as variáveis independentes e dependentes são de natureza linear, eles são capazes de lidar com relações não lineares [43]. Neste trabalho iremos abordar de maneira mais aprofundada sobre os modelos de Regressão Linear Múltipla (MLR) e de Regressão de Mínimos Quadrados Parciais (PLS), que serão aplicados aos estudos de caso avaliados.

3.2.1 Modelo de Regressão Linear Múltipla (MLR)

Modelos de regressão linear correspondem ao estabelecimento de uma relação linear, nos coeficientes, entre um ou mais parâmetros *y* a serem preditos, chamados de variáveis dependentes ou respostas, e um ou mais *x*, chamados de variáveis independentes ou preditoras. O número de variáveis selecionadas pode distinguir os modelos de regressão linear em três casos diferentes: regressão linear simples, regressão linear múltipla e regressão linear múltipla multivariada. Na regressão linear simples a relação é feita a partir de uma variável dependente (y) e uma variável independente (x), por exemplo, a predição da concentração de composto em relação à absorbância do analito, em um único comprimento de onda. A regressão linear múltipla considera a relação linear entre múltiplas variáveis independentes (x) e uma variável dependente (y). A regressão

linear múltipla tem uma lógica semelhante à regressão linear simples, mas com duas ou mais variáveis independentes. A regressão linear múltipla multivariada usa várias variáveis dependentes (*y*) e independentes (*x*) ao mesmo tempo [37, 38].

Em modelos de regressão linear múltipla (MLR), as características das amostras são medidas com o objetivo de estabelecer uma função linear entre várias (*m*) variáveis independentes x_j (j = 1-m) e qualquer variável dependente, mais um erro, que para um modelo ajustado devidamente deve ser aleatório (ε), como mostrado nas equações abaixo (**Eq. 1a, 1b e 1c**) [37, 39].

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_mx_m + \varepsilon \quad (1a)$$

$$y = \sum_{j=1}^{m} b_jx_j + \varepsilon \quad (1b)$$

$$y = x^Tb + \varepsilon \quad (1c)$$

Na **Eq. 1a** *y* é a variável dependente, b_j são as sensibilidades e ε o erro aleatório. Na **Eq. 1c** *y* é um escalar, **b** é um vetor coluna e **x** um vetor linha. As dependências multilineares para uma amostra foram descritas pelas equações **Eq. 1(a-c)**. Para n amostras, o $\mathbf{y_i}$ (i = 1 – n) pode ser descrito como um vetor coluna (n × 1), **b** como um vetor coluna (m × 1), os vetores $\mathbf{x_j^T}$ formam as linhas da matriz **X** (n × m) e **ε** o erro aleatório (n × 1):

$$y = Xb + \varepsilon \quad (2)$$

Considerando a **Eq. 2** para n amostras, é possível distinguir três casos diferentes em relação ao número de amostras (n) e ao número de variáveis independentes (m) [39].

(I) m > n: Neste caso há mais variáveis do que amostras, ocorre uma situação indesejada porque *b* tem um número infinito de soluções, todas se encaixando na equação.

(II) m = n: Quando o número de variáveis é igual ao número de amostras, gera uma solução única para *b* se **X** tem posto completo. Este caso não é comum em situações práticas, mas fornece um vetor residual de erro aleatório (**ε**) como um vetor de zeros.

(III) m < n: Existem mais amostras do que variáveis neste caso, embora não permita uma solução exata para *b*, uma solução pode ser obtida minimizando o comprimento do vetor residual. O método mais popular para fazer isso é chamado de método dos mínimos quadrados.

Executando o método dos mínimos quadrados para encontrar a solução exata para *b* quando m < n, deve-se usar a **Eq. 3**, onde *b* representa o valor esperado para o vetor **b** minimizando o comprimento do vetor residual [37].

$$b = (X^T X)^{-1} X^T y \qquad (3)$$

Para que apenas uma solução **b** seja obtida com sucesso em um método MLR, não basta que o número de variáveis (m) seja menor que o número de amostras (n), a correlação entre essas variáveis precisa ser baixa. A existência de colinearidade entre as variáveis, o que significa que o ângulo entre elas é pequeno, e têm praticamente a mesma direção, dificulta a inversão da matriz ($\mathbf{X}^T\mathbf{X}$), tornando o problema instável do ponto de vista matemático. O determinante da matriz ($\mathbf{X}^T\mathbf{X}$) é próximo de zero, causando um mau condicionamento, e a matriz inversa não é confiável [44].

O caminho a seguir é fazer uma seleção das variáveis, para superar a singularidade nos dados. A seleção é baseada na capacidade preditiva das variáveis. Os três modos de seleção mais comuns são: para frente (*forward*), para trás (*backward*) e passo a passo

(*stepwise*). Uma fraqueza dessas abordagens é que elas descrevem apenas a capacidade de modelar os dados de treinamento, em vez da capacidade de predizer novas amostras. A falta de capacidade de predição pode ser a principal falha ao usar MLR [7,45].

Atualmente, o uso de métodos de calibração multivariada, como MLR, tem aumentado e requer a determinação simultânea de vários analitos, principalmente em análises espectrofotométricas. A aplicação da MLR, apesar de fácil e simples de interpretar, geralmente requer seleção de variáveis espectrais para construção de modelos bem ajustados, pois é mais dependente de uma boa escolha de variáveis. Ao longo deste trabalho, serão abordadas algumas abordagens de seleção de variáveis como Recozimento Simulado (RS), Algoritmos Genéticos (AG), Mínimos Quadrados Parciais por Intervalo (iPLS), Eliminação de Variáveis não-informativas (UVE) e Algoritmo das Projeções Sucessivas (SPA).

3.2.2 Modelo dos Mínimos Quadrados Parciais (PLS)

Modelos de Regressão dos Mínimos Quadrados Parciais (PLS) são usados principalmente na análise de dados altamente colineares e ruidosos. De forma geral, os dados possuem muitas variáveis *x* (independentes) e deseja-se modelar simultaneamente muitas variáveis do tipo *y* (dependentes). Em instrumentos de medição modernos, como espectrofotômetros e cromatógrafos, podemos observar a capacidade de medida de uma infinidade de variáveis do tipo *x* altamente correlacionadas, e como consequência obtêm-se dados ruidosos e incompletos. Desta forma, o PLS demonstra-se extremamente útil neste contexto, já que permite analisar estas complexidades através da manipulação de uma grande quantidade de variáveis *x* e *y* de maneira mais racional [3].

Neste tipo de modelagem a variância de *x* e *y* é descrita através do ajuste da matriz de dados **X** e **Y**, baseado em centralização na média ou dimensionamento. Como o PLS tem como principal objetivo predizer as variáveis *y* através de *x*, é necessário que se obtenha dados com máxima covariância onde as matrizes **X** e **Y** são decompostas em variáveis escores de **X** (**t**) e escores de **Y** (**u**). Os escores de **X** estimam a combinação linear da variável x_k com o coeficiente de peso (**W***):

$$T = X \times W^* \quad (4)$$

No entanto, o peso W pode ser transformado em **W*** que está diretamente relacionado a **X**, através da equação:

$$W^* = W(P^T W)^{-1} \quad (5)$$

Nos modelos PLS primeiramente se faz uma relação externa que irá descrever as matrizes **X** e **Y** de maneira individual. Sendo definida pelas seguintes equações:

$$X = T \times P^T + E \quad (6)$$

$$Y = U \times C^T + F \quad (7)$$

onde $\mathbf{P^T}$ é a matriz de carregamentos (*loadings)* do espaço X, $\mathbf{C^T}$ é a matriz de carregamentos (*loadings)* do espaço Y e **E** e **F** são as matrizes residuais dos espaços X e Y, respectivamente.

Posteriormente se faz a relação interna a fim de relacionas as duas matrizes, já que os escores de **X** (T) também são bons preditores para variáveis **Y**, ou seja, correlacionados de acordo com a seguinte equação:

$$Y = T \times C^T + G \quad (8)$$

Pela combinação das **Eq. 4** e **Eq. 8** pode-se escrever:

$$Y = XW^*C^T + G = XB + G \quad (9a)$$

Onde:

$$B = W^{*}C^{T} \quad (9b)$$

B representa o coeficiente do PLS e **G** é a matriz residual, ou seja, a porção não explicada pelo modelo. A aplicabilidade do modelo pode ser determinada pelo valor residual dele, sendo um modelo ruim aquele que apresenta grande valor residual.

A predição de variáveis *y* de novas amostras é determinada através da **Eq. 9a**. Adicionando o valor de **W*** obtido através da **Eq. 5** na **Eq. 9b**:

$$B = W(P^{T}W)^{-1}C^{T} \quad (10)$$

Em um modelo PLS, quando o primeiro componente for calculado, então um outro pode ser calculado com base nas matrizes residuais. O número de componentes PLS significativos em um modelo de calibração pode ser decidido por meio de validação cruzada [24-27]. A principal limitação deste método é a preparação da calibração, bem como a predição do conjunto e o emprego da decisão humana para selecionar o número de fatores.

3.3 Algoritmos de Seleção de Variáveis

Dentro do campo de aplicação da quimiometria, as técnicas multivariadas e de inteligência artificial aplicadas à dados químicos têm interessado muito às indústrias, devido à sua grande aplicabilidade na classificação de produtos ou predição de propriedades químicas, facilitando a tomada de decisão em diversos campos de aplicação [1-2]. As informações a respeito das amostras obtidas através de instrumentação analítica, podem ser modeladas a fim de construir modelos matemáticos capazes de predizer um parâmetro de interesse ou uma classe à qual a amostra pertence. Porém, a construção destes modelos pode ser fortemente prejudicada pelo grande número de variáveis sendo

analisadas, principalmente pelo fato de muitas delas apresentarem informações redundantes ou pela presença excessiva de ruído instrumental [1, 4-5, 46].

O elevado número de variáveis associado às informações redundantes pode reduzir a acurácia dos modelos e produzir resultados que não são confiáveis, fazendo com que o modelo não seja satisfatório à finalidade proposta [1, 6-7, 46-48]. A criação de modelos mais robustos, precisos e interpretáveis pode ser obtida através da seleção de um número reduzido de variáveis informativas [4, 7-9]. Esta redução de dimensionalidade pode ser obtida através da identificação de um subconjunto de variáveis que sejam mais significativas à predição do parâmetro proposto ou através de combinação de algumas das variáveis originais [10-11].

Os algoritmos de seleção de variáveis podem atuar como: a) filtros de pré-processamento; b) externos ao modelo ou; c) internos ao modelo. Nos algoritmos de filtro (a) é estabelecido um limiar de corte em um dado modelo previamente ajustado, onde as variáveis são selecionadas por estarem acima ou abaixo deste valor limite (Ex.: *Jack-Knife* e estratégias que usam ponderação dos pesos). No caso dos algoritmos externos ao modelo (b), atribui-se um valor associado a uma dada função de custo, através da análise individual de cada subconjunto de variáveis gerado (Ex.: AG, SPA, CF e seleção de intervalos). Já nos algoritmos internos ao modelo (c), a seleção de variáveis ocorre simultaneamente ao processo de modelagem (Ex.: *soft-threshold-PLS* (ST-PLS)) [28, 49].

Além da classificação por forma de atuação, os algoritmos de seleção de variáveis podem também ser subdivididos de acordo com a característica da variável de entrada: randômicos ou determinísticos e; de acordo com a forma de seleção dos subconjuntos de variáveis: variáveis individuais discretas, intervalo ou combinações de intervalos contínuos. Para os algoritmos nos quais alguma variável de entrada seja aleatória, o subconjunto de variáveis selecionado está associado a um grau de probabilidade (estocástico), por isso, são denominados como randômicos (Ex.: métodos que simulam processos naturais como AG, CF, busca tabu). Já nos casos em que não existe grau de probabilidade nas variáveis de entrada, é possível obter uma única solução para o subconjunto de variáveis e, por isso, são denominados métodos determinísticos (Ex.: SPA). [49]

Em se tratando da forma de seleção dos subconjuntos de variáveis, as variáveis individuais discretas, compostas por conjuntos descontínuos, são muito empregadas como uma solução para os problemas de multicolinearidade em métodos de regressão multivariada. No caso dos algoritmos de seleção por intervalos ou conjuntos de intervalos, são mais indicados para métodos em que se utiliza rotações ortogonais prévias ou em variáveis contínuas (espectros, voltamogramas e cromatogramas) [49].

Ao longo dos anos, diversos métodos foram criados para auxiliar nos processos de seleção de variáveis relevantes [7, 31], sendo alguns eles: recozimento simulado (RS) [12, 15-16], algoritmos genéticos (AG) [16-21], mínimos quadrados parciais de intervalo (iPLS, *interval partial least squares)* [12, 22-23], eliminação de variável não informativa (UVE, *uninformative variable elimination*) [12, 15, 24-25] e algoritmo de projeções sucessivas (SPA) [12, 14, 23, 25, 32-33]. Cada um dos métodos de seleção descritos aqui, serão descritos brevemente nos subcapítulos a seguir.

3.3.1 Recozimento Simulado (RS)

O algoritmo de otimização por Recozimento Simulado tem sido utilizado mais recentemente devido ao grande interesse na aplicação de métodos matemáticos que mimetizam processos naturais [42]. Tem sido amplamente utilizado em seleção de comprimentos de onda para análise de múltiplos componentes usando espectroscopia de UV-visível e infravermelho próximo, além da utilização no refinamento de estruturas moleculares determinadas por espectroscopia de ressonância magnética nuclear (RMN) e cristalografia de raios X [50].

É um método de busca iterativo, baseado no processo de recozimento de sólidos, mais especificamente de metais [51]. Este algoritmo visa encontrar uma solução ótima para problemas de otimização combinatória, explorando novas áreas do espaço para solução de problemas de forma interativa [52]. A habilidade de exploração destas áreas em um pequeno espaço de tempo e com pouco esforço marca a performance desta técnica [53]. Este algoritmo criado foi analogamente relacionado com a otimização combinatória por Kirkpatrick et al. (1983) e aperfeiçoada por Cerny (1985). O termo recozimento refere-se a um processo térmico de transformação de um material liquefeito em um sólido, ou seja, o processo tem início em altas temperaturas, seguido pela redução lenta e

gradativa da temperatura, fazendo com que o ponto de solidificação seja atingido, chegando a um estado de mínima energia [54-55].

Teoricamente, o sistema no qual o processo está ocorrendo deve estar em equilíbrio ao longo do tempo, porém, na prática ocorrem aumentos de energia de curta duração, que são provocados por alguns processos aleatórios [50, 52]. No processo natural a temperatura é ocasionalmente aceita como uma probabilidade controlada que gera movimentos ascendentes, ou seja, à medida que a temperatura diminui, a probabilidade de aceitação destes movimentos também reduz. Com base neste processo, em alta temperatura o processo de busca se torna aleatório, enquanto em baixas temperaturas a busca se torna quase exagerada [51, 54-55].

A fim de simular este processo natural, o método matemático é iniciado através da identificação, aleatória ou por experiência, de valores iniciais para os níveis dos k fatores. Na sequência, ocorre a inclusão de funções de perturbação, através da adição de um vetor aleatório obtido usando os k números aleatórios, nos valores iniciais obtidos, gerando um novo conjunto de condições experimentais. Se o novo conjunto de respostas (R2) for melhor que o conjunto de respostas iniciais (R1), a etapa de adição aleatória é repetida até a obtenção de um mínimo global, que representa o estado de menor energia, conforme ilustrado na **Figura 2** [50-51].

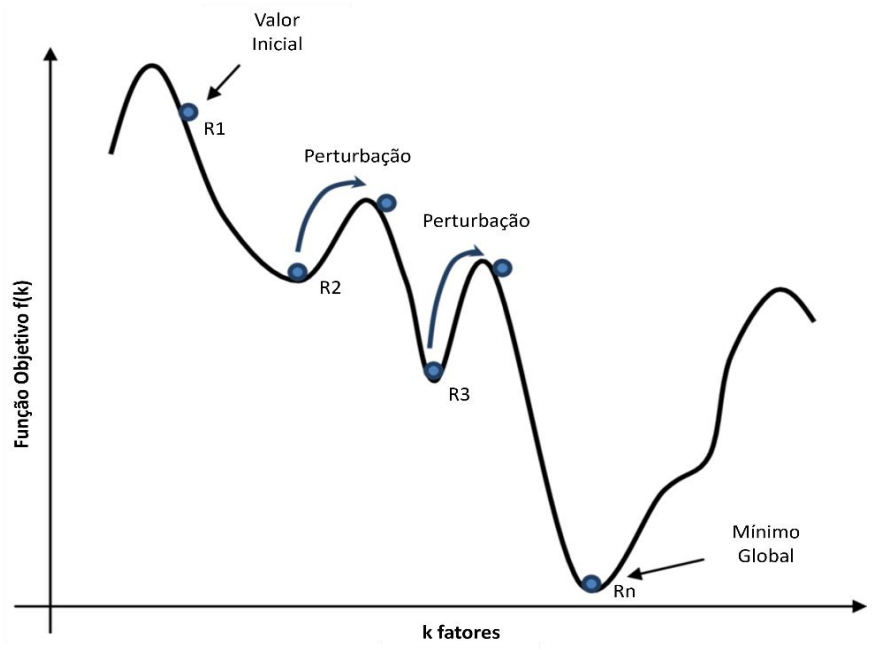

***Figura 2**. Simulação das etapas do algoritmo de recozimento simulado.*

O recozimento simulado tem como vantagem o fato de que a cada iteração uma única solução é encontrada, desta forma, os processamentos de avaliação das soluções não impactam muito negativamente na eficiência do algoritmo. Um inconveniente deste tipo de busca é que ele utiliza poucas informações do problema, como a variação da função objetivo, o que a torna pouco inteligente [52].

3.3.2 Algoritmos Genéticos

Assim como o Recozimento Simulado que foi introduzido a fim de mimetizar processos naturais, o Algoritmo Genético (AG) também tem o objetivo de simular o processo evolutivo de uma espécie viva. O método foi proposto por John H. Holland nos anos 70 como uma abordagem de otimização, seguindo as regras clássicas de Charles R. Darwin sobre evolução natural e utilizando passos aleatórios de forma a convergir a uma solução ótimo não-aleatória [21]. A metodologia de AG está baseada em cinco etapas principais. A primeira etapa consiste em codificar as variáveis onde cada uma irá corresponder a um gene e cada condição experimental corresponderá a um cromossomo (sequência de genes). Dentre as diversas maneiras de codificar os valores das variáveis, a mais usual é a utilização de um código binário, atribuindo-se um valor 0 ou 1 à variável que representa um gene não incluído ou incluído no modelo, respectivamente [21]. Conforme mostrado na **Figura 3**.

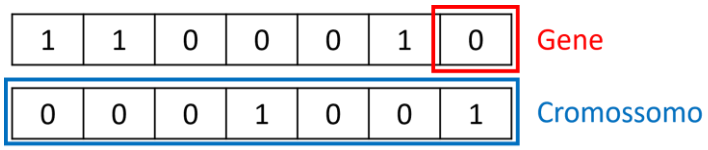

Figura 3. *Etapa 1 AG – Codificação de variáveis, utilizando respostas binárias.*

A segunda etapa é a iniciação da população (**Figura 4**) na qual a população original é composta por um número N de cromossomos e, depois é decidida a ordem dos genes nestes cromossomos, através da sequência de valores binários 0 e 1, de forma totalmente aleatória [21]. A terceira etapa realiza a avaliação da resposta associada às condições experimentais para cada cromossomo, onde uma resposta nula é atribuída se a condição

experimental estiver fora do domínio experimental ou corresponder a um experimento impossível de realizar [21].

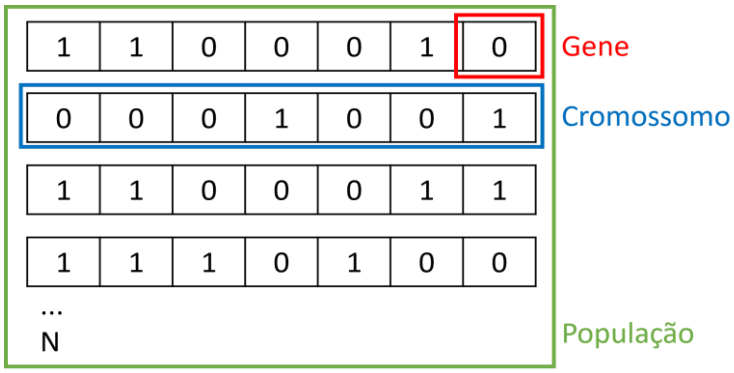

***Figura 4**. Etapa 2 AG – Iniciação da população.*

Na reprodução, quarta etapa, uma nova população de cromossomos N é criada, podendo ser considerada como a próxima geração, através de mecanismos de seleção e cópia ou cruzamento. No mecanismo de seleção e cópia (**Figura 5a**) obtém-se de forma geral uma nova população na qual os melhores cromossomos são copiados com maior frequência, levando a uma resposta média melhor. Já no mecanismo de cruzamento (**Figura 5b**) ocorre a exploração de novas condições experimentais através da mistura de variáveis já testadas, porém em combinações diferentes [21].

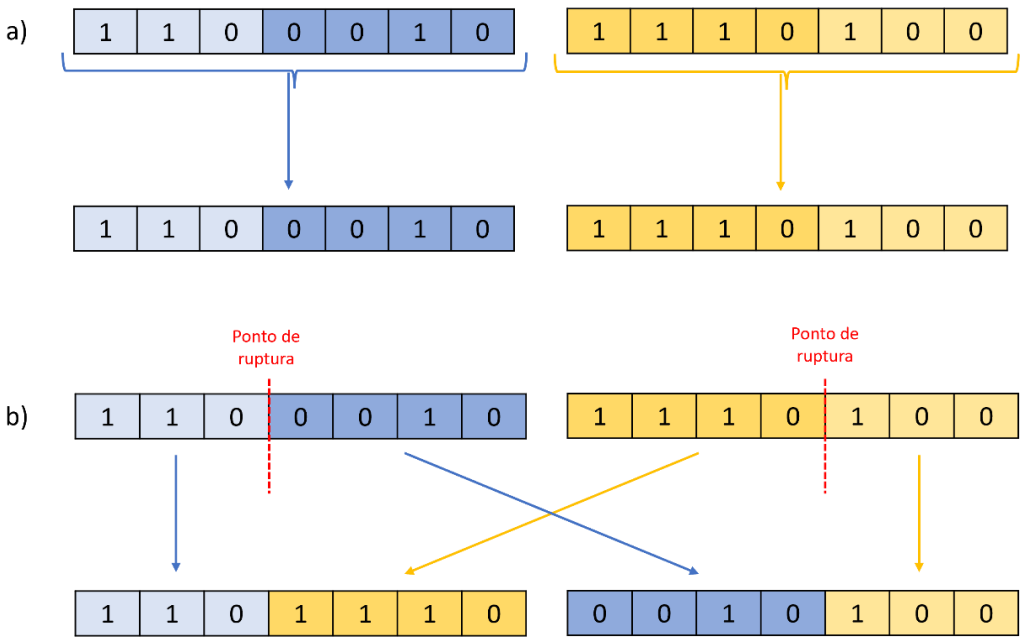

Figura 5. *Etapa 4 AG – Reprodução da população através dos mecanismos de (a) seleção e cópia e (b) cruzamento.*

Na quinta e última etapa, chamada de mutação, ocorre o sorteio dos genes dos cromossomos que devem ser afetados por uma mutação e a inversão do valor binário 0 ou 1 é realizada, conforme mostrado na **Figura 6**. Esta operação impede que o algoritmo fique preso a mínimos locais, fazendo com que se mova a novas regiões do domínio experimental [21].

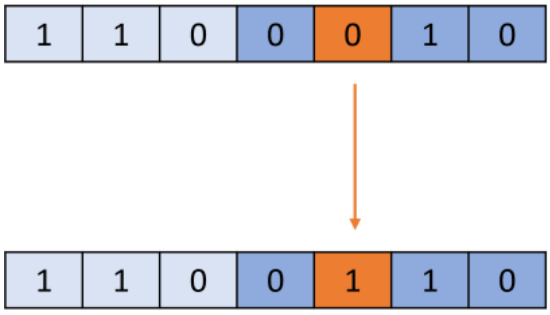

Figura 6. *Etapa 5 AG – Mutação.*

As etapas 3, 4 e 5 se alternam até que um critério de término seja alcançado, que pode ser baseado em um número máximo de gerações, no tempo total permitido para a elaboração ou na falta de melhoria na resposta [21, 56].

3.3.3 Mínimos Quadrados Parciais por Intervalo (iPLS)

O algoritmo dos Mínimos Quadrados Parciais por intervalo (iPLS) foi introduzido nos anos 2000 com o intuito de selecionar comprimentos de onda de um espectro, projetando e dividindo o mesmo em subintervalos de igual distância, para posteriormente aplicar um modelo PLS para cada um dos intervalos definidos. A seleção de subintervalos propicia modelos mais estáveis e de fácil interpretação, onde a colinearidade não seja tão importante. Os subintervalos podem ser otimizados através da inclusão ou eliminação de novas variáveis, sendo selecionado como melhor subintervalo aquele em que o erro de predição é menor [12, 28].

A **Figura 7** exemplifica um espectro subdividido em 10 intervalos de igual distância, sendo as barras uma estimativa dos valores de RMSECV para cada um destes intervalos. A linha tracejada na imagem representa o RMSECV do modelo global, ou seja, os intervalos que apresentam RMSECV menor do que o do modelo global indicam as melhores regiões para compor o modelo PLS final [57]. Este tipo de seleção é muito útil na melhoria do desempenho dos métodos de calibração, já que minimiza os erros de predição, além de excluir ruídos não significantes.

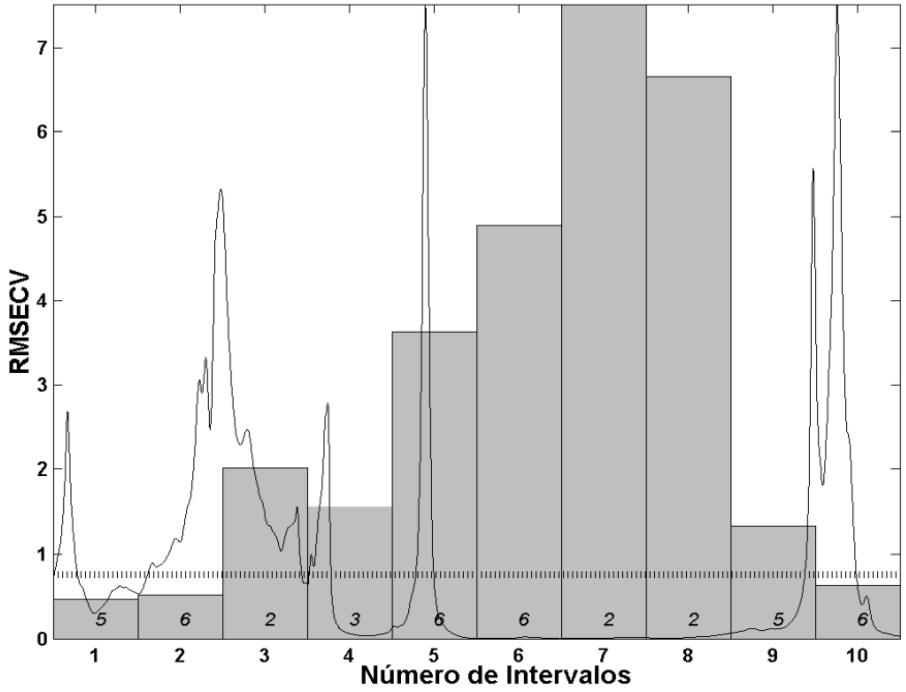

Figura 7. *Espectro com modelo iPLS aplicado, relacionando os intervalos com o RMSECV.*

3.3.4 Eliminação de Variáveis não-informativas

O método de Eliminação de Variáveis não-informativas consiste em adicionar variáveis artificiais a fim de se obter um modelo PLS ou PCR com conjunto de variáveis contendo as variáveis experimentais e um ruído adicional (variáveis artificiais). Neste método proposto por Centner e colaboradores (1996), as variáveis experimentais menos importantes do que as variáveis artificiais adicionadas são eliminadas, através da utilização de um critério baseado no coeficiente *b* [24].

Com base em uma avaliação de dados simulados, os autores concluíram que este tipo de eliminação pode melhorar a capacidade preditiva do modelo (RMSEP conhecido como Raiz Quadrada do Erro Quadrático Médio de Predição), afinal, a qualidade de um modelo de calibração multivariada está relacionada a alguns fatores como a qualidade das variáveis. Desta forma, com a eliminação de variáveis ruidosas ou aleatórias, que levam a uma maior variância no modelo pelo erro adicionado por elas (menor precisão), pode se observar um incremento na qualidade do modelo [24].

Neste método, parte-se da matriz **X** (n × p) contendo as variáveis experimentais e uma matriz **R** (n × p) de variáveis artificiais é gerada através da multiplicação por uma

constante, desta forma ambas as matrizes possuem p variáveis. A matriz **R** é então incluída na matriz **X**, obtendo-se uma matriz **XR** (n × 2p). Um modelo PLS é calculado para a matriz **XR** utilizando-se o procedimento leave-one-out, obtendo-se 2p coeficientes de regressão que representa o vetor **b** em uma matriz **B** (n × 2p). A partir daí, é avaliado o critério de confiabilidade (*c*), baseado em uma analogia ao modelo MLR, onde a **Eq. 11** a seguir é aplicada aos dados centrados. Desta forma, a inserção da j-ésima variável (para j = 1, ..., p) está baseada na razão entre o coeficiente de regressão b_j e seu desvio padrão s(b_j):

$$c_j = \frac{b_j}{s(b_j)} \qquad (11)$$

Para aplicação desta forma de seleção de variáveis em modelos PLS, é necessário que b_j seja estimado como uma média e s(b_j) como um desvio padrão. Na etapa seguinte determina-se o maior valor absoluto de c_j para valores de j maiores que p (região das variáveis artificiais), ou seja, o max (c_{art}). As variáveis de **X** consideradas como não informativas, ou seja, a serem eliminadas são $|c_j| < |\max(c_{art})|$ (para j = 1,...,p), conforme exemplo ilustrativo da **Figura 8** [24].

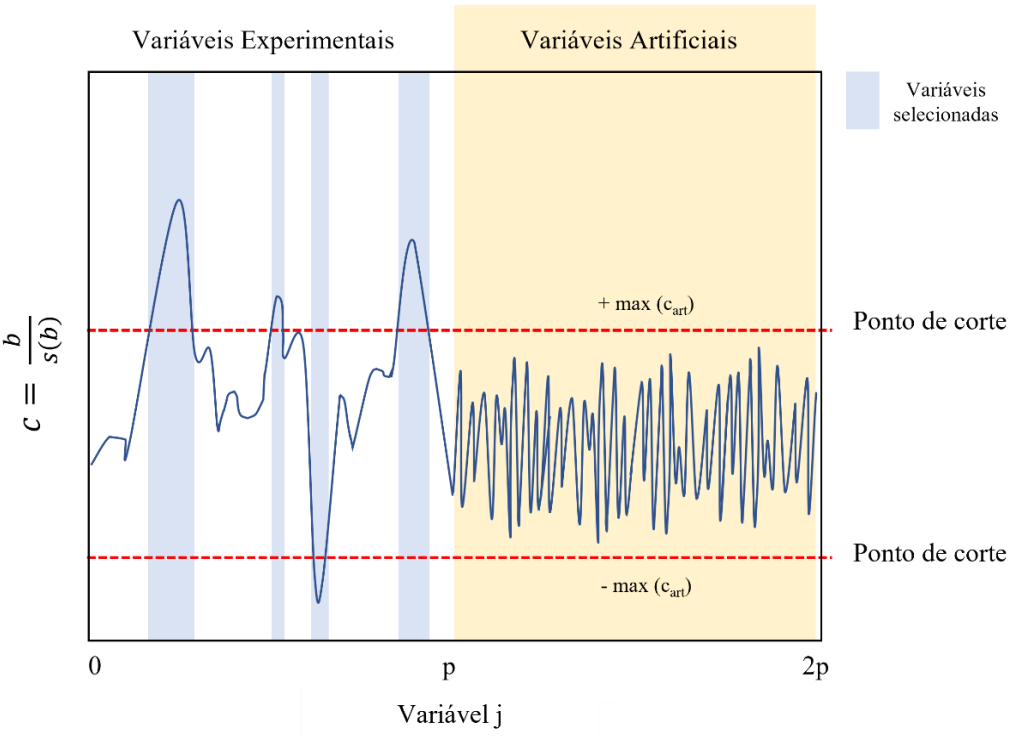

***Figura 8**. Exemplo de Eliminação de Variáveis não-informativas.*

3.3.5 Algoritmo das Projeções Sucessivas (SPA)

O algoritmo de projeções sucessivas (SPA) apresenta a vantagem de encontrar um pequeno conjunto representativo de variáveis com um nível mínimo de colinearidade [7, 28]. O SPA é um método iterativo de seleção direta que opera na matriz de resposta instrumental, selecionando variáveis minimamente redundantes, para resolver problemas de colinearidade. A seleção de variáveis começa com uma variável, então incorpora uma nova a cada iteração, até que um número especificado N de variável seja alcançado [7, 58]. A combinação de SPA como seleção de variáveis para métodos MLR é chamada de SPA-MLR e pode ser dividida em três fases. Na primeira fase, Fase I, as respostas instrumentais do conjunto de calibração são dispostas em uma matriz $\mathbf{X}_{cal}$ de dimensões centrada na média ($N_{cal} \times K$) de modo que a k^{th} variável x_k esteja associada ao k^{th} vetor coluna $\mathbf{x}_k \in \Re^{N_{cal}}$. Cada resposta instrumental (comprimentos de onda) corresponde a um vetor coluna submetido a uma sequência de operações de projeção, resultando em K cadeias de M variáveis, onde $M = \min(N_{cal} - 1, K)$ é o número máximo de variáveis que

podem ser incluídas em um modelo MLR com dados centrados na média. Cada cadeia é inicializada com uma variável x_k e incrementada com as variáveis que apresentam a menor colinearidade com as anteriores, conforme procedimento descrito em seis passos abaixo (**Eq. 12 a Eq. 15**) [1, 46, 59].

Passo 1: Inicialização

$$\mathbf{z}^1 = \mathbf{x}_k \qquad \text{(vetor que define as operações iniciais de projeção)} \tag{12}$$

$$\mathbf{x}_j^1 = \mathbf{x}_j, \qquad j = 1, \ldots, K \tag{13}$$

$$\text{SEL}(1, k) = k \tag{14}$$

$$i = 1 \qquad \text{(contador de iteração)} \tag{15}$$

Passo 2: Calcular a matriz $\mathbf{P}^i$ de projeção no subespaço ortogonal a $\mathbf{z}^i$, onde **I** é uma matriz identidade ($N_{cal} \times N_{cal}$).

$$\mathbf{P}^i = \mathbf{I} - \frac{\mathbf{z}^i (\mathbf{z}^i)^T}{(\mathbf{z}^i)^T \mathbf{z}^i} \tag{16}$$

Passo 3: Calcular os vetores projetados $\mathbf{x}_j^{i+1}$ para todo $j = 1, \ldots, K$.

$$\mathbf{x}_j^{i+1} = \mathbf{P}^i \mathbf{x}_j^i \tag{17}$$

Passo 4: Determinar o índice j^* do maior vetor projetado e armazene este índice no elemento $(i + 1, k)$ da matriz **SEL**:

$$j^* = \arg \max_{j=1,\ldots,K} (\mathbf{x}_j^{i+1}) \tag{18}$$

$$SEL(i+1, k) = j^* \tag{19}$$

Passo 5: Seja $\mathbf{z}^{i+1} = \mathbf{x}_{j^*}^{i+1}$ (vetor que define as operações de projeção para a próxima iteração).

Passo 6: Seja $i = i + 1$. Se $i < M$ volte para o Passo 2.

Ao final da Fase I, essas cadeias são armazenadas em uma matriz **SEL** ($M \times K$) tal que $SEL(1, k)$, $SEL(2, k)$, …, $SEL(M, k)$ correspondem aos índices de M variáveis na k-ésima (k^{th}) cadeia. É importante notar que o procedimento de busca utilizado na Fase I não garante que o máximo geral do determinante para um determinado número de variáveis seja alcançado, mas é uma ótima alternativa para substituir a busca exaustiva, que consome muito tempo de processamento [1].

A segunda fase, Fase II, consiste em avaliar os subconjuntos candidatos de m variáveis a partir de x_k extraídas das cadeias armazenadas na matriz **SEL**. Cada cadeia é definida pelo conjunto de índices $\{SEL(1, k), SEL(2, k), \ldots, SEL(m, k)\}$. Como m varia de 1 a M e k varia de 1 a K e, um total de $M \times K$ subconjuntos de variáveis devem ser avaliados. O melhor subconjunto de variáveis é selecionado com base em uma função de custo relacionada à capacidade de predição do modelo MLR. Normalmente, essa função de custo é calculada como o erro quadrático médio obtido usando validação cruzada ou um conjunto de validação separado (**Eq. 20**). Ao final desta fase é definido um índice de relevância para cada variável pertencente ao subconjunto selecionado [1, 46, 59].

$$RMSEV = \sqrt{\frac{1}{N_{val}} \sum_{i=1}^{N_{val}} \left(y_{val,i} - \hat{y}_{val,i}\right)^2} \qquad (20)$$

A terceira fase, Fase III, consiste em um procedimento de eliminação para trás (*backward*) visando melhorar a parcimônia do modelo. Nesta fase as variáveis são ordenadas de acordo com seu índice de relevância e o RMSEV é recalculado incluindo progressivamente as variáveis ordenadas no modelo MLR. Ao final deste processo a solução é obtida utilizando o menor número de variáveis para que RMSEV não seja significativamente maior que o valor mínimo observado, conforme um teste F [1].

3.3.5.1 Aplicações de SPA-MLR:

O método de seleção de variáveis do tipo SPA, conforme mostrado no capítulo 3.3.5, já foi aplicado em diferentes matrizes e técnicas para seleção de variáveis. Alguns

trabalhos foram utilizados em matrizes de biodiesel e/ou diesel [14, 23, 34, 63], já outros aplicados à produtos naturais como milho [34], frutas cítricas [58], folhas de chá [62], óleos vegetais [64] e água do mar [66]. Além disto, pode se observar na **Tabela 1** a seguir, que alguns trabalhos utilizaram NIR [14, 23, 25, 34, 39, 58, 63-65] e outros UV-Vis [12, 33, 36, 66]

Tabela 1. *Aplicações do* SPA *como método de seleção de variáveis*

Artigo	Ano	Técnica	Matriz	Ref.
Determinação do teor de biodiesel em misturas de biodiesel/diesel usando NIR e espectroscopia visível com seleção variável	2011	NIR and UV-Vis	Biodiesel/diesel	[14]
Espectroscopia de infravermelho e calibração multivariada para monitorar parâmetros de estabilidade e qualidade do biodiesel	2010	NIR and MIR	Biodiesel	[23]
Algoritmo de projeções sucessivas combinado com eliminação de variáveis não informativas para seleção de variáveis espectrais	2008	NIR	Tabaco e tabletes farmacêuticos	[25]
Aspectos do algoritmo de projeções sucessivas para seleção de variáveis em calibração multivariada aplicado à espectrometria de emissão de plasma	2001	ICP-OES	Dados simulados e Aço	[32]
O algoritmo de projeções sucessivas para seleção de variáveis em análise multicomponente espectroscópica	2001	UV-Vis	Misturas sintéticas	[33]
Um método de eliminação de variáveis para melhorar a parcimônia de modelos MLR usando o algoritmo de projeções sucessivas	2008	NIR	Diesel e Milho	[34]
Determinação Espectrométrica Simultânea de Cu^{2+}, Mn^{2+} e Zn^{2+} em Medicamentos Polivitamínicos/Poliminerais Utilizando Algoritmos SPA e GA para Seleção de Variáveis	2005	UV-Vis	Medicamentos Polivitamínicos e Poliminerais	[36]
Seleção de variáveis em espectros de infravermelho visível/próximo para calibrações lineares e não lineares: um estudo de caso para determinar o teor de sólidos solúveis da cerveja	2009	Vis-NIR	Cerveja	[39]
Determinação não destrutiva do teor de sólidos solúveis de frutas cítricas usando tecnologia de transmitância de infravermelho próximo combinada com o algoritmo de seleção variável	2020	Vis-NIR	Frutas cítricas	[58]
Imagens hiperespectrais NIR acopladas à quimiometria para avaliação não destrutiva dos conteúdos de fósforo e potássio em folhas de chá	2020	HSI	Folhas de Chá	[62]
Determinação de enxofre total em óleo diesel empregando espectroscopia NIR e calibração multivariada	2003	NIR	Diesel	[63]
Determinação espectrométrica NIR de parâmetros de qualidade em óleos vegetais usando iPLS e seleção de variáveis	2008	NIR	Óleos vegetais	[64]
Determinando a qualidade de óleos isolantes usando espectroscopia de infravermelho próximo e seleção de comprimento de onda	2011	NIR	Óleos isolantes	[65]
Algoritmo de projeções sucessivas melhorando a determinação espectrofotométrica direta simultânea multivariada de cinco compostos fenólicos na água do mar	2007	UV-Vis	Água do mar	[66]
Seleção de recursos e métodos de regressão linear/não linear para a previsão precisa das atividades inibitórias de glicogênio sintase quinase-3β	2009	Molecular descriptors	Série de Inibidores de Glicogênio Sintase Quinase-3β	[67]
Modelagem QSPR dos coeficientes de sorção do solo (K_{OC}) de pesticidas usando SPA-ANN e SPA-MLR	2009	Molecular descriptors	Pesticidas	[68]

4 Experimental

4.1 Modelo *f*SPA-MLR proposto

Considerando um modelo de regressão linear múltipla cujas variáveis são previamente centradas na média como mostrado na **Eq. 1a**:

$$y = b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots + b_m x_m + \varepsilon \qquad (1a)$$

A estimativa dos coeficientes de regressão pode ser obtida por mínimos quadrados ordinários dados pela **Eq. 3**[44-45]:

$$b = (X^T X)^{-1} X^T y \qquad (3)$$

Observe que calcular o inverso da matriz de covariância exigirá vetores coluna em **X** tão linearmente independentes quanto possível. Mas, ao mesmo tempo, esses vetores coluna devem carregar as informações apropriadas para predizer *y*. Na versão de trabalho atual do SPA-MLR conforme descrito por Soares *et al*. [5], a fase 1 das projeções se concentra em minimizar a multicolinearidade para obter matrizes de covariância invertíveis e o número de variáveis incluídas na cadeia de projeção ser menor ou igual ao número de amostras do conjunto de calibração. No entanto, devido à elevada dimensionalidade da matriz e à possibilidade de incluir apenas um número limitado de variáveis nas cadeias APS, muitas variáveis úteis para fins de calibração são frequentemente deixadas de fora da avaliação baseada em critérios de proporção. Quando o SPA-MLR é usado em dados maiores, é comum que precursores importantes não sejam selecionados e, neste caso, para melhores resultados são necessárias etapas manuais prévias. No método proposto, denominado fAPS-MLR, a etapa do filtro ou etapa zero adicionada consiste na exclusão prévia de variáveis não informativas, medidas pelo

coeficiente de correlação entre a j-ésima coluna da matriz X e o vetor **y** (Eq. 21). Além das variáveis altamente redundantes (colineares) indicadas pelo cálculo do VIF, que é o fator de inflação da variância (**Eq. 22**).

$$r(j) = \frac{1}{J-1} \times \sum_{j=1}^{J} \left(\frac{x_{i,j}-\underline{x_j}}{s_j}\right) \times \left(\frac{y_i-\underline{y}}{s_y}\right) \quad (21)$$

$$VIF\ (j) = \frac{1}{(1-p^2(j))} \quad (22)$$

onde $x_{i,j}$ são os elementos da matriz de resposta instrumental **X** (I × J); $\underline{x_j}$ e s_j são a média e o desvio padrão da j-ésima (j^{th}) coluna de **X**, respectivamente; y_i, $\underline{y}$, e s_y são o vetor, a média e o desvio padrão da variável dependente *y*, respectivamente; *p*(*j*) é o coeficiente de correlação entre x_j e $\hat{x}_j$ calculado por regressão linear usando as variáveis *x* restantes.

Uma variável preditora deve ter valores de VIF baixos e valores de correlação y altos, portanto, após a normalização para valor máximo igual a 1; ambos são combinados (**Eq. 23**) para gerar um índice de avaliação (*J*) das variáveis e posterior definição de um limiar (*J-limite*). Somente variáveis com valor de J maior que *J-limite* serão incluídas na etapa de projeção, fase 1 do SPA.

$$J = |r| \times \frac{1}{VIF} \quad (23)$$

J assume valores positivos maiores que zero até 1 e o *J-limite* é definido minimizando o RMSECV ou RMSEV em um loop interno e o PRESS (soma dos quadrados dos erros de predição – critério de Haaland e Thomas) é usado como critério de escolha do valor ótimo. A matriz de calibração $\mathbf{X_{cal}}$ (I × J) é reduzida a $\mathbf{X_{cal}}$ (I × J^T). Essa etapa de triagem iniciada permite que as etapas subsequentes do SPA sejam mais

eficazes, evitando a necessidade de exclusão manual de partes dos dados que são trabalhosas, não otimizadas e baseadas na experiência do analista no uso desse algoritmo. Observa-se ao utilizar este tipo de parâmetro que um valor de J-limit muito pequeno não resultará em uma remoção eficiente de variáveis não informativas. Já um valor J-limite muito alto pode levar à exclusão excessiva de preditores importantes para o modelo. Na configuração padrão J são avaliados valores de 0,0001 a 0,1 com um passo de 0,01.

4.2 Procedimento quimiométrico e software

Para avaliar os efeitos de diferentes tipos de pré-processamento nos dados para posterior utilização do algoritmo de seleção de variáveis proposto, *f*SPA-MLR, os espectros de ambos os conjuntos de dados experimentais foram pré-processados.

Algumas informações irrelevantes, incluindo ruído, linha de base e dispersão de luz causadas pelo instrumento ou ambiente externo, podem afetar a identificação de informações valiosas do espectro nos modelos quimiométricos. Assim, o pré-processamento de espectros pode ser útil para desenvolver modelos confiáveis e robustos [58]. Neste trabalho, os dados espectrais foram pré-processados usando a primeira derivada da função Savitzky-Golay (SG) [5, 14, 59-61] e a função Standard Normal Variate (SNV) [58, 60-62]. A primeira derivada da função SG é eficaz para reduzir os deslocamentos da linha de base e os picos superpostos e a função SNV é usado para eliminar o efeito do tamanho das partículas sólidas e da mudança do caminho óptico nos espectros originais [58]. As condições específicas utilizadas para cada conjunto de dados experimentais serão descritas na **Seção 4.3**. Além disso, algumas métricas foram utilizadas para avaliar a performance dos modelos de regressão como RMSECV (Raiz Quadrada do Erro Quadrático Médio de Validação Cruzada (Root-Mean-Square Error of Cross-Validation)), R2 (Coeficiente de determinação) e RMSEP (Raiz Quadrada do Erro Quadrático Médio de Predição (Root-Mean-Square Error of Prediction)). Os pré-processamentos de Savitzy-Golay e SNV, cálculos de SPA-MLR, *f*SPA-MLR, PLS e métricas de avaliação de performance foram realizados em Matlab® 2010a (Mathworks).

4.3 Estudos de caso

4.3.1 Estudo de caso I

Seiscentas e cinquenta e quatro amostras de comprimidos farmacêuticos foram analisadas a partir de dois espectrômetros diferentes e os dados foram publicados como um conjunto de dados de espectro "*Shootout*" pela International Diffuse Reflectance Conference (IDRC) em 2002 (disponível em https://eigenvector.com/resources /conjuntos de dados/). A região NIR na faixa de 600 a 1898 nm com 2 nm de incremento foi adotada para este estudo, totalizando 650 comprimentos de onda. Para cada comprimido, três variáveis de resposta foram medidas: a quantidade de ingrediente ativo (nominalmente 200 mg/comprimido), o peso e a dureza. Por se tratar de um conjunto de dados de simulação,os conjuntos de calibração (155 amostras), validação (40 amostras) e teste (460 amostras) foram separados previamente, sem possibilidade de ajuste dos conjuntos. Foram realizados dois pré-processamentos diferentes: a) primeira derivada da função Savitzky-Golay com janela de 13 pontos e polinômios de segunda ordem e; b) função SNV. A derivada resultante e os espectros SNV foram utilizados ao longo do trabalho.

4.3.2 Estudo de caso II

Cem amostras de mistura diesel/biodiesel foram analisadas na faixa espectral de 750 a 2500 nm com 1 nm de resolução, totalizando 1751 comprimentos de onda. A porcentagem de biodiesel no diesel (% m/m) ficou entre 5 a 50% e a densidade foi medida para cada amostra. O conjunto de dados é dividido em conjuntos de calibração (40 amostras) e de teste (60 amostras). No conjunto de dados foi realizada a regressão linear utilizando o teor de biodiesel no diesel (% m/m) e não a densidade. Foram realizados dois pré-processamentos diferentes: a) primeira derivada da função Savitzky-Golay com janela de 21 pontos e polinômios de segunda ordem e; b) função SNV. A derivada resultante e os espectros SNV foram utilizados ao longo do trabalho. As medidas foram realizadas utilizando espectrofotômetro Perkin Elmer modelo 750 Lambda, equipado com célula de quartzo com caminho óptico de 1 cm, fonte de tungstênio e tubo fotomultiplicador R928 e sistemas de detecção de PbS resfriados por Peltier.

5 Resultados e Discussão

5.1 Estudo de caso I

A **Figura 9a, b e c** mostra os espectros sem pré-processamento e pré-processados com primeira derivada da função SG e função SNV, respectivamente, das 155 amostras de comprimidos farmacêuticos do conjunto de calibração. Como pode ser visto na **Figura 9a** os espectros sem pré-processamento exibiram características de linha de base sistemáticas e picos superpostos, que foram removidos pelo procedimento de primeira derivada da função SG (**Figura 9b**). Por sua vez, o procedimento SNV mostrado na **Figura 9c** eliminou o efeito da mudança do caminho óptico nos espectros completos, o que gerou maior similaridade entre os espectros das amostras do mesmo conjunto de dados.

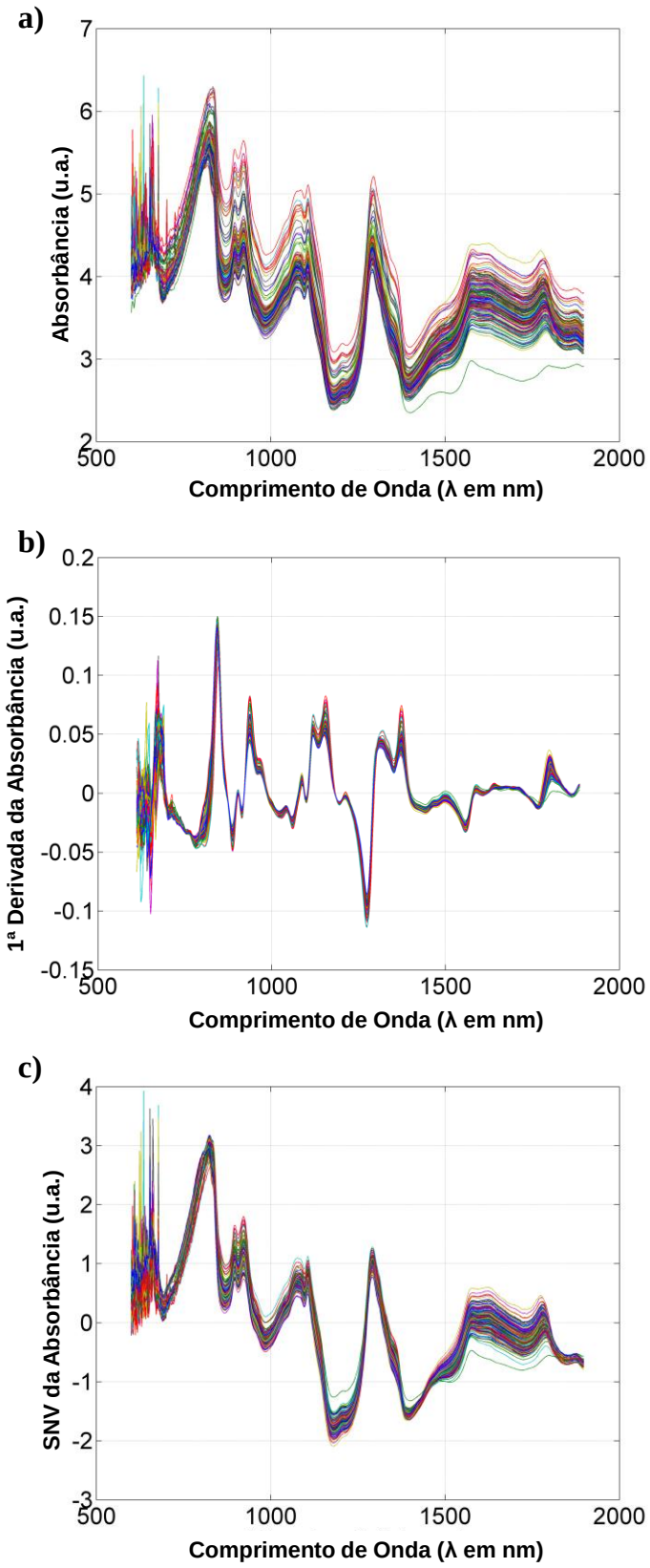

Figura 9. *(a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processamento pela função SNV das 155 amostras de comprimidos farmacêuticos do conjunto de calibração.*

Quando avaliamos as bandas de absorção no infravermelho, podemos observar que as bandas com maior intensidade de absorbância aparecem de forma muito clara nos três espectros da **Figura 9**, independentemente do tipo de pré-processamento realizado. A banda entre 1815 e 1790 nm que está presente nos três espectros é característica de grupos funcionais C=O de cloretos de acila, já as bandas em 1300 e 1100 nm são características de ligações C-O de ésteres insaturados aromáticos. As bandas entre 860 e 800 nm, bem como as bandas entre 900 e 860 nm são características de anéis aromáticos, para 2 H adjacentes e H isolado, respectivamente. Ao longo do trabalho serão mostradas à quais das bandas dos espectros as variáveis selecionadas estão relacionadas.

O conjunto de dados do Estudo de Caso I foi processado de forma a avaliar os modelos de regressão linear múltipla (MLR) para a quantidade de ingrediente ativo (nominalmente 200 mg/comprimido), o peso do comprimido e sua dureza, desta forma, serão apresentados os resultados para cada um dos modelos preditivos separadamente, levando em consideração os diferentes tipos de pré-processamento de espectro em todos os casos.

O valor de J-limite foi definido como o valor de J calculado na etapa de otimização que minimiza RMSECV ou RMSEV em um loop interno e o PRESS. Sendo que na configuração padrão de J são avaliados valores de 0,0001 a 0,1 com um passo de 0,01. A **Figura 10** mostra os valores J calculados para as 155 amostras de comprimidos farmacêuticos, tendo sido escolhidos como J-limite o menor valor de RMSEV, sendo 0,01 para os espectros sem pré-processamento, 0,02 para os espectros pré-processados com primeira derivada da função SG e 0.0051 para espectros pré-processamento pela função SNV.

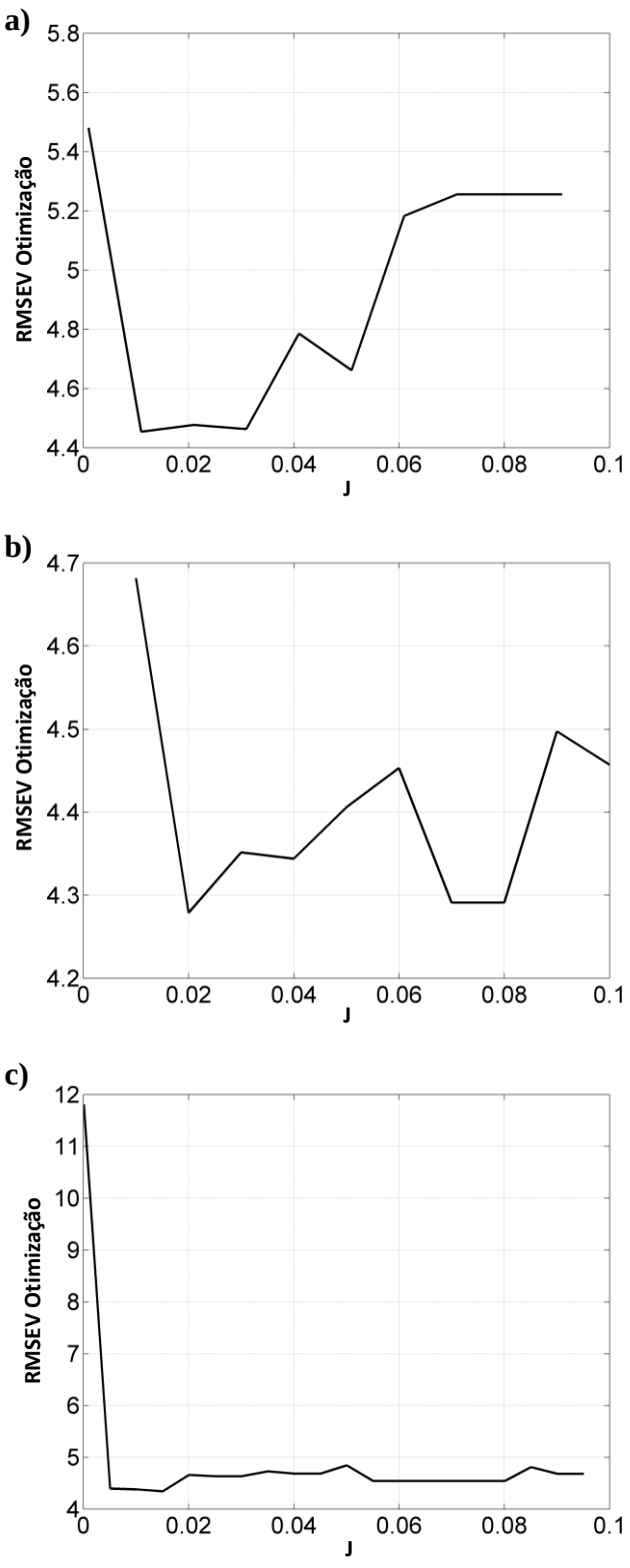

Figura 10. Valores de J calculados para *(a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processamento pela função SNV das 155 amostras de comprimidos farmacêuticos do conjunto de calibração.*

5.1.1 Modelagem para predição de quantidade de ingrediente ativo

A **Figura 11a, b e c**, mostra as variáveis selecionadas para construir o modelo MLR para o parâmetro de ingredientes ativos usando os espectros sem pré-processamento e pré-processados com primeira derivada da função SG e função SNV, respectivamente. Os círculos azuis indicam a variável selecionada do SPA-MLR e os quadrados verdes indicam a variável selecionada do *f*SPA-MLR.

As variáveis selecionadas na **Figura 11a** via *f*SPA-MLR (quadrados verdes) foram 1636, 1326, 1290 e 834 nm, já para o método SPA-MLR (círculos azuis) apenas as variáveis 1282 e 840 nm foram selecionadas. Na **Figura 11b** as variáveis 1372, 1300 e 846 nm foram selecionadas via *f*SPA-MLR e apenas a variável 1296 nm via SPA-MLR. Na **Figura 11c** 1326, 1282 e 836 nm foram as variáveis selecionadas via *f*SPA-MLR e apenas 840 nm via SPA-MLR. De forma geral, observa-se que as bandas de maior importância, independentemente dos métodos propostos, estão relacionadas aos anéis aromáticos com 2 H adjacentes presentes nas amostras (bandas entre 860 e 800) e às ligações C-O de ésteres insaturados aromáticos (bandas em 1300 e 1100 nm), ou seja, estas são as variáveis que melhor caracterizam o parâmetro de princípio ativo das amostras.

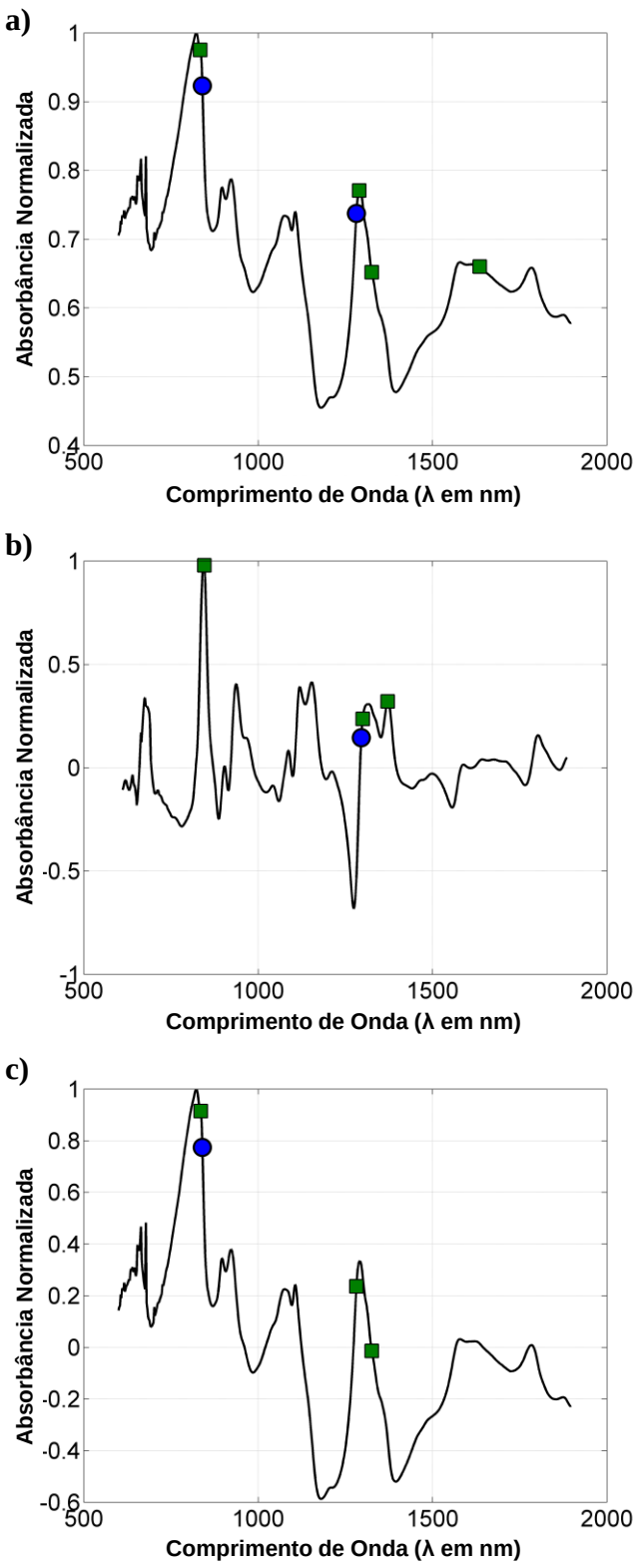

Figura 11. *(a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de ingredientes ativos, do conjunto de calibração (círculos azuis indicam as variáveis selecionadas em* SPA-*MLR e os quadrados verdes indicam as variáveis selecionadas em f*SPA-*MLR).*

Como pode ser visto na **Figura 11**, as variáveis selecionadas no *f*SPA-MLR (quadrados verdes) estão em maior número e mais representativas dos espectros, independentemente do tipo de pré-processamento, este resultado, juntamente com a **Tabela 2,** explica as fragilidades do SPA em sua versão sem a etapa de filtro. A **Tabela 2** resume os resultados de validação cruzada e predição externa obtidos com SPA-MLR, *f*SPA-MLR e PLS com os dados de espectros com e sem pré-processamento. Como pode ser visto, o *f*SPA-MLR supera o SPA-MLR para quaisquer dados de espectros com ou sem pré-processamento, em termos de validação e predição externa. Os parâmetros RMSECV e RMSEP foram menores no modelo *f*SPA-MLR e R^2(cv) e R^2(pred) foram maiores que o SPA-MLR. Por outro lado, o *f*SPA-MLR apresenta desempenho semelhante ao modelo PLS, em termos de validação cruzada e predição externa.

Tabela 2. *Resultados de modelos de predição de quantidade de ingrediente ativo em amostras de comprimidos farmacêuticos.*

Pré-processamento	Modelo	RMSECV	R^2 (cv)	bias (cv)	RMSEP	R^2 (pred)	bias (pred)	Número de variáveis[a]
Sem pré-processamento	**SPA-MLR**	6,477	0,913	0,021	5,520	0,890	-1,401	2
	***f*SPA-MLR**	4,454	0,959	0,001	4,408	0,923	0,471	4
	PLS	5,136	0,953	0,012	4,715	0,912	0,162	4
1ª derivada SG	**SPA-MLR**	5,140	0,945	-0,016	4,740	0,913	0,655	1
	***f*SPA-MLR**	4,279	0,962	0,004	4,132	0,931	0,116	3
	PLS	4,830	0,950	0,054	4,656	0,936	0,080	4
SNV	**SPA-MLR**	11,810	0,709	0,022	11,639	0,584	-3,900	1
	***f*SPA-MLR**	4,394	0,960	0,007	4,255	0,928	0,157	3
	PLS	4,955	0,934	0,060	4,722	0,916	0,262	4

[a] *Variáveis latentes em PLS e variáveis espectrais em SPA-MLR e fSPA-MLR.*

A **Figura 12a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando o espectro sem etapa de pré-processamento. A **Figura 13a, b** e **c**, mostra o gráfico de valores preditos versus valores de referência dos três modelos usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 14a, b** e **c**, mostra o gráfico de valores preditos versus valores de referência dos três modelos usando SNV como pré-processamento do espectro.

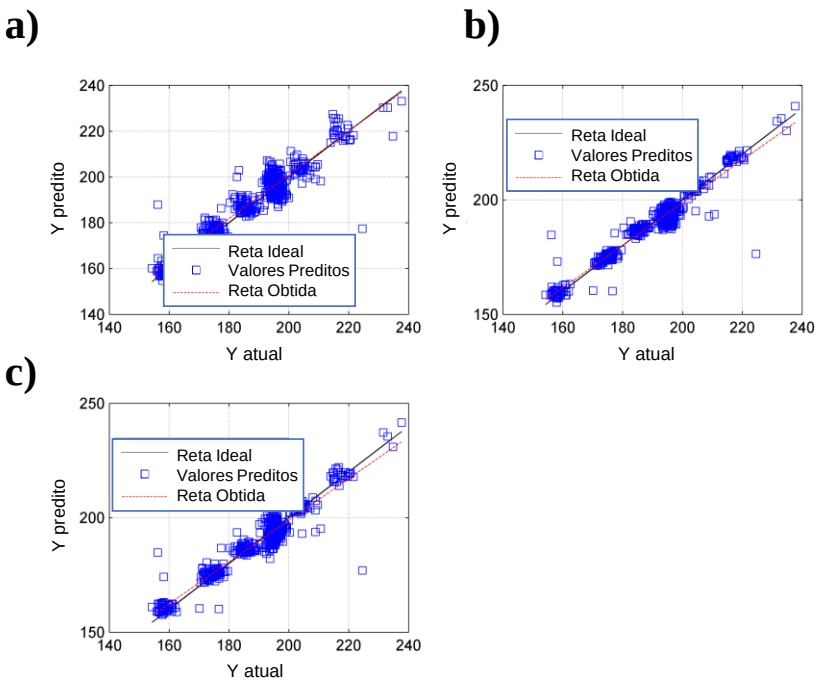

Figura 12. *Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros sem etapa de pré-processamento em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

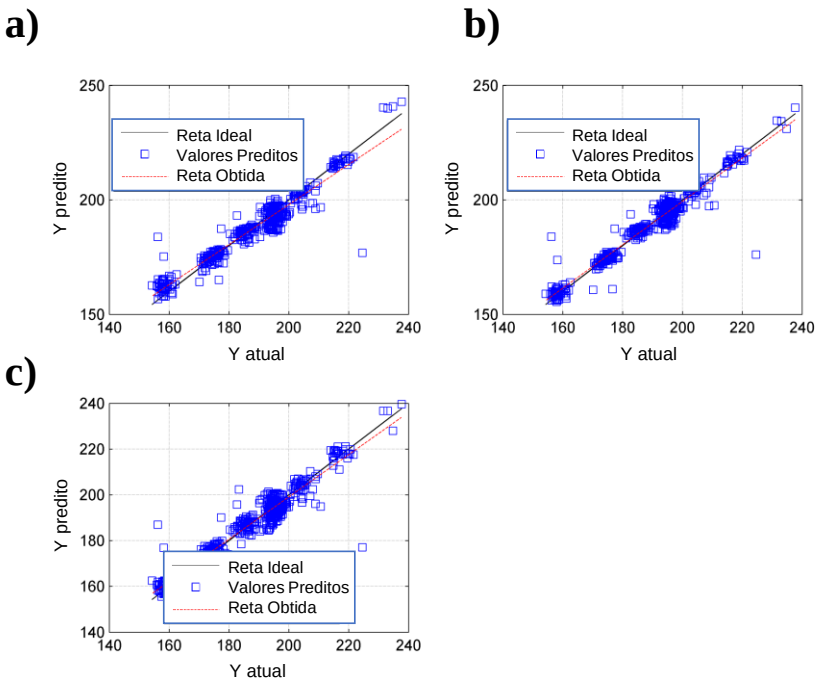

Figura 13. *Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros pré-processados com 1ª derivada SG em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

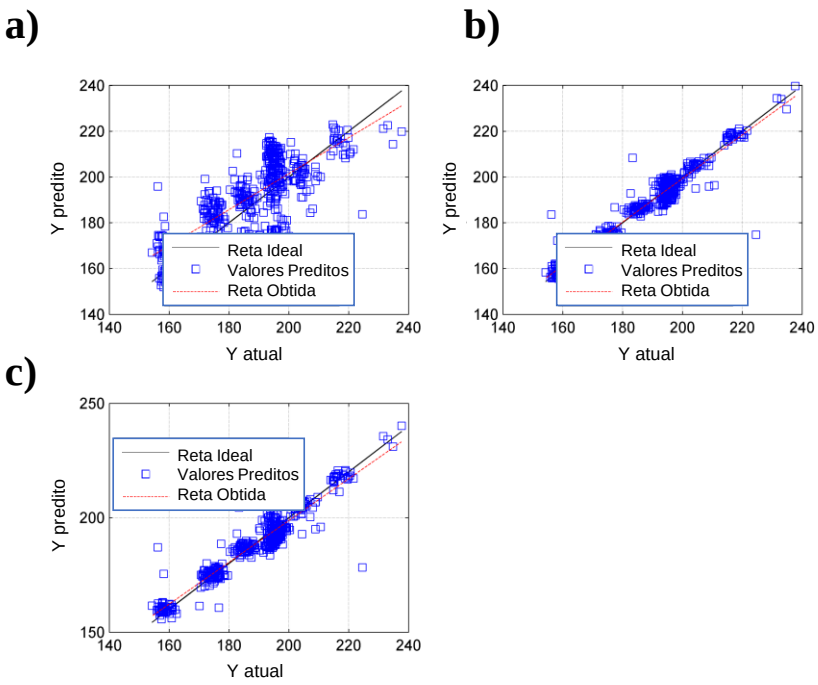

Figura 14. *Gráfico linear dos valores de ingrediente ativo preditos versus reais para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

Como pode ser observado na **Tabela 2**, **Figura 12(a-c), Figura 13(a-c)** e **Figura 14(a-c)**, os melhores resultados foram obtidos a partir dos dados de espectros pré-processados com 1ª derivada SG.

De forma a avaliar a presença ou ausência de erros sistemáticos nos modelos, foram gerados gráficos de erro residual de modelo. A **Figura 15** mostra os erros residuais dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS, usando o espectro sem etapa de pré-processamento. A **Figura 16a, b** e **c** mostra os erros residuais dos três modelos, respectivamente, usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 17a, b** e **c**, mostra os erros residuais usando SNV como pré-processamento do espectro. É importante notar que nas **Figura 15(a-c), Figura 16(a-c)** e **Figura 17(a-c)**, as amostras são distribuídas aleatoriamente, o que indica a ausência de erro sistemático, além disso, o modelo *f*SPA-MLR apresenta erros menores ou semelhantes aos apresentados nos outros dois modelos (SPA-MLR e PLS).

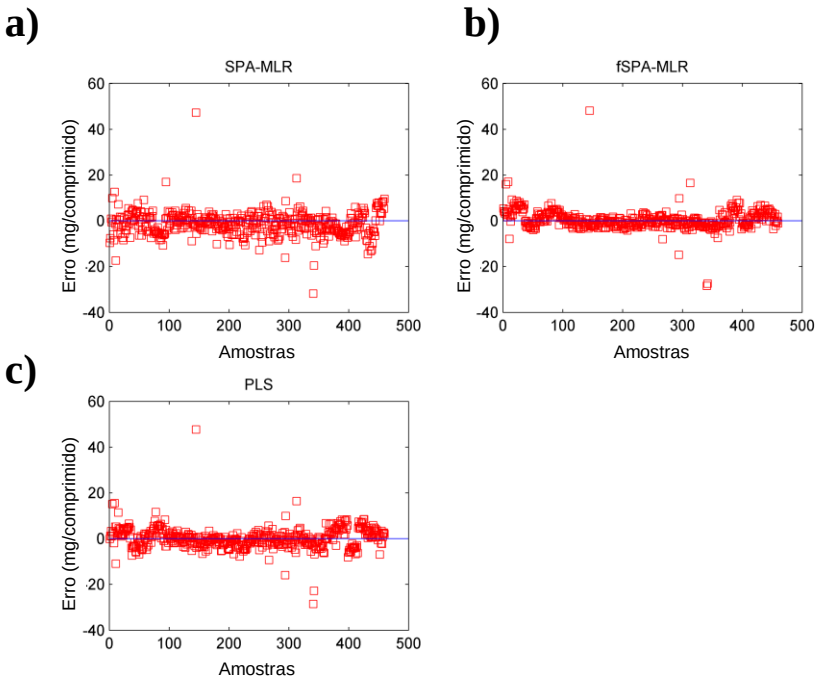

Figura 15. *Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a)* SPA*-MLR, (b) f*SPA*-MLR e (c) PLS, para o parâmetro de quantidade de ingrediente ativo.*

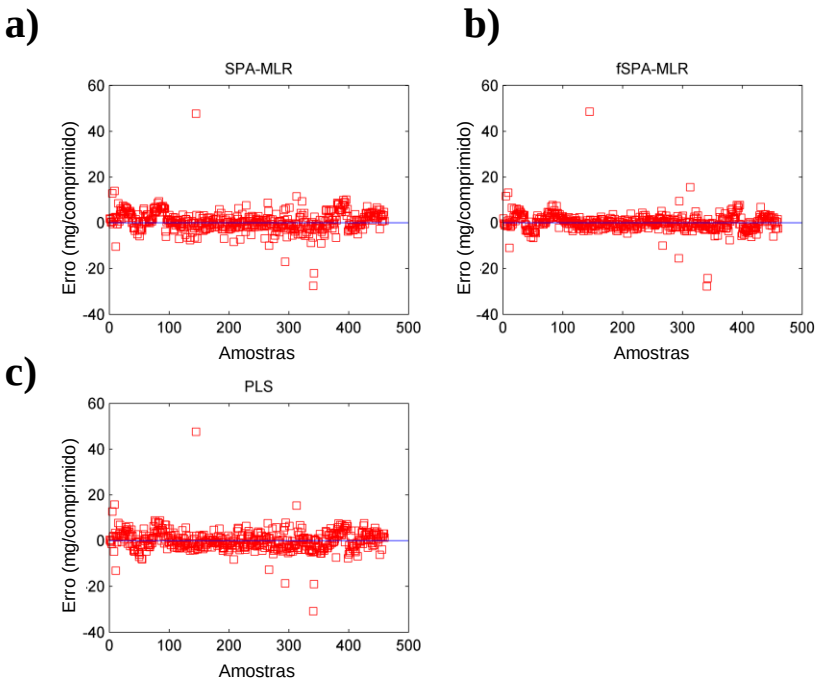

Figura 16. *Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a)* SPA*-MLR, (b) f*SPA*-MLR e (c) PLS para o parâmetro de quantidade de ingrediente ativo.*

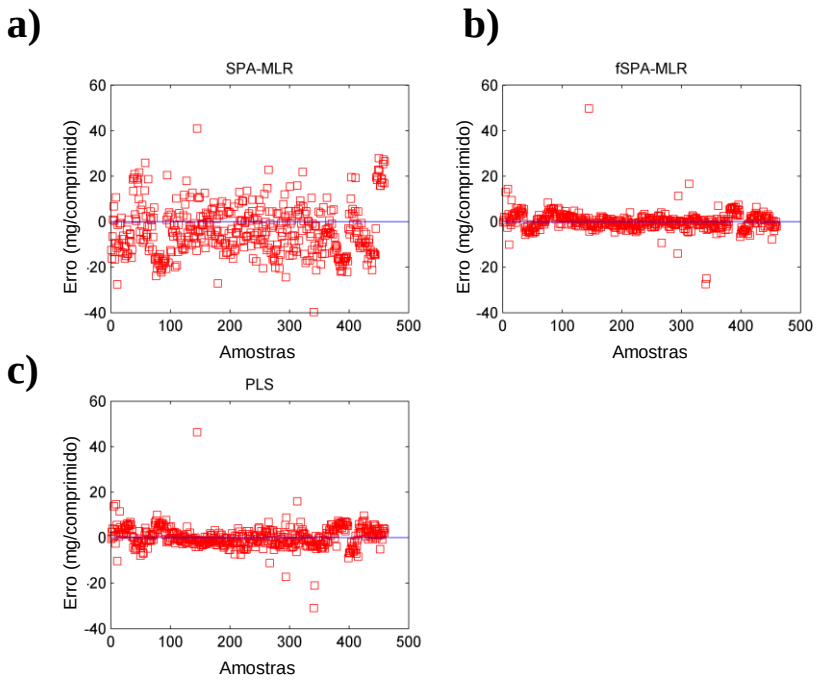

Figura 17. *Erros residuais dos modelos para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS para o parâmetro de quantidade de ingrediente ativo.*

5.1.2 Modelagem para predição do peso do comprimido

A **Figura 18a, b e c**, mostra as variáveis selecionadas para construir o modelo MLR para o parâmetro de peso dos comprimidos usando os espectros sem pré-processamento e pré-processados com primeira derivada da função SG e função SNV, respectivamente. Os círculos azuis indicam a variável selecionada do SPA-MLR e os quadrados verdes indicam a variável selecionada do *f*SPA-MLR.

As variáveis selecionadas na **Figura 18a** via *f*SPA-MLR (quadrados verdes) foram 1766, 1362, 1294, 1098, 886 e 846 nm, já para o método SPA-MLR (círculos azuis) apenas as variáveis 1630 e 1068 nm foram selecionadas. Na **Figura 18b** as variáveis 1808, 1286, 934, 864 e 856 nm foram selecionadas via *f*SPA-MLR e apenas as variáveis 1260 e 824 nm via SPA-MLR. Na **Figura 18c** 1630, 1580, 994, 964 e 866 nm foram as variáveis selecionadas via *f*SPA-MLR e apenas 972 nm via SPA-MLR. Observa-se que, neste caso do parâmetro de peso do comprimido, a alteração dos métodos de pré-processamento gerou diferenças significativas na escolha das variáveis, já que poucas delas aparecem em mais de um dos métodos propostos. Mas é importante salientar que utilizando-se o

método de seleção *f*SPA-MLR observa-se mais variáveis importantes do que utilizando o método via SPA-MLR.

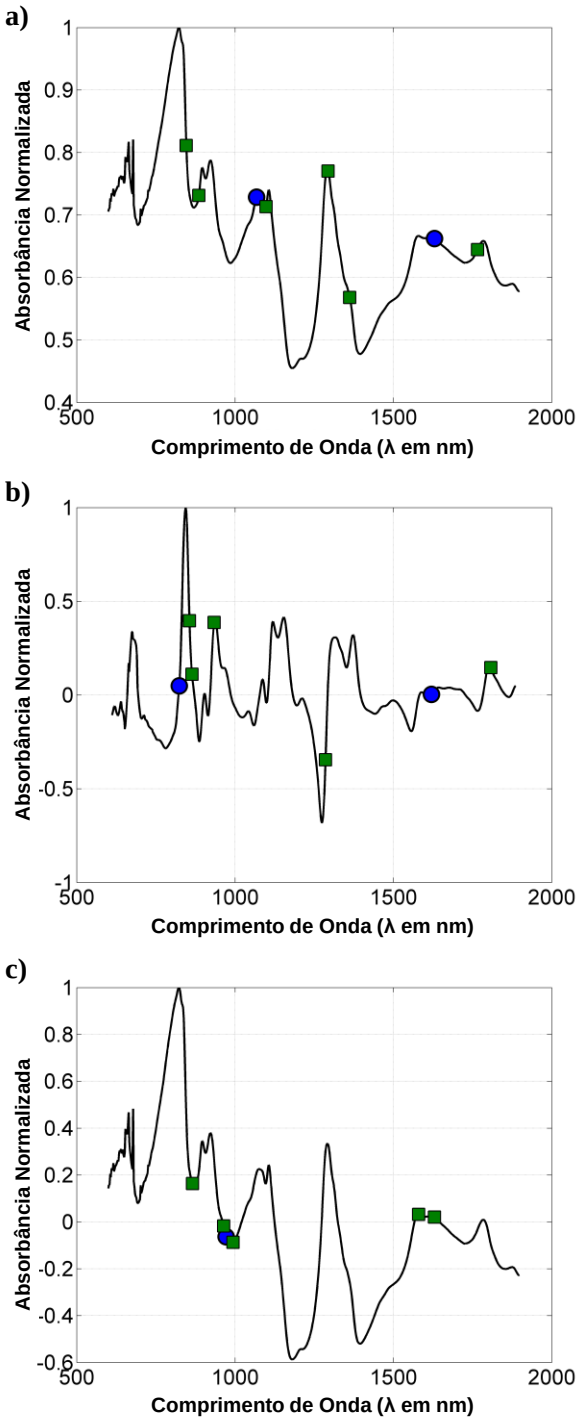

Figura 18. *(a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de peso dos comprimidos, do conjunto de calibração (círculos azuis indicam as variáves selecionadas em* SPA*-MLR e os quadrados verdes indicam as variáveis selecionadas em f*SPA*-MLR).*

Como pode ser visto na **Figura 18**, as variáveis selecionadas no *f*SPA-MLR (quadrados verdes) estão em maior número e mais representativas dos espectros, independentemente do tipo de pré-processamento, este resultado, juntamente com a **Tabela 3**, explica as fragilidades do SPA em sua versão sem a etapa de filtro. A **Tabela 3** resume os resultados de validação cruzada e predição externa obtidos com SPA-MLR, *f*SPA-MLR e PLS com os dados de espectros com e sem pré-processamento. Como pode ser visto, o *f*SPA-MLR supera o SPA-MLR e o PLS para quaisquer dados de espectros com ou sem pré-processamento, em termos de validação cruzada e predição externa. Os parâmetros RMSECV e RMSEP foram menores no modelo *f*SPA-MLR e R^2(cv) e R^2(pred) foram maiores que o SPA-MLR e o PLS.

Tabela 3. *Resultados de modelos de predição do peso de cada comprimido em amostras de comprimidos farmacêuticos.*

Pré-processamento	Modelo	RMSECV	R^2 (cv)	bias (cv)	RMSEP	R^2 (pred)	bias (pred)	Número de variáveis[a]
Sem pré-processamento	**SPA-MLR**	3,974	0,491	-0,001	5,051	0,221	-0,195	2
	***f*SPA-MLR**	2,794	0,748	0,005	3,564	0,485	-0,677	6
	PLS	3,856	0,588	0,007	5,180	0,216	0,016	3
1ª derivada SG	**SPA-MLR**	3,662	0,568	0,008	4,412	0,319	-0,203	2
	***f*SPA-MLR**	2,939	0,721	0,002	3,719	0,433	-0,692	5
	PLS	3,719	0,650	0,008	4,910	0,241	0,086	3
SNV	**SPA-MLR**	3,881	0,514	0,004	4,898	0,239	-0,189	1
	***f*SPA-MLR**	3,362	0,636	0,005	4,571	0,361	-0,093	5
	PLS	3,847	0,609	0,048	4,992	0,244	0,084	2

[a] *Variáveis latentes em PLS e variáveis espectrais em SPA-MLR e fSPA-MLR.*

Para a determinação do parâmetro de peso do comprimido os resultados obtidos via *f*SPA-MLR não foram tão satisfatórios quanto os resultados para o parâmetro de quantidade de ingrediente ativo que tiveram resultados de R^2(cv) e R^2(pred) maiores que 0,923. Esta diferença é explicada pelo princípio da técnica utilizada, pois os espectros NIR trazem informações relacionadas aos compostos presentes na amostra. Os compostos presentes na amostra podem dar uma ideia de peso do comprimido, mas como podemos observar nos resultados obtidos, os modelos possuem eficiência inferior.

A **Figura 19a, b** e **c**, mostra o gráfico de valores preditos versus valores de referência dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS usando o espectro sem etapa de pré-processamento. A **Figura 20a, b** e **c**, mostra o gráfico de valores preditos versus valores de referência dos três modelos usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 21a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando SNV como pré-processamento do espectro.

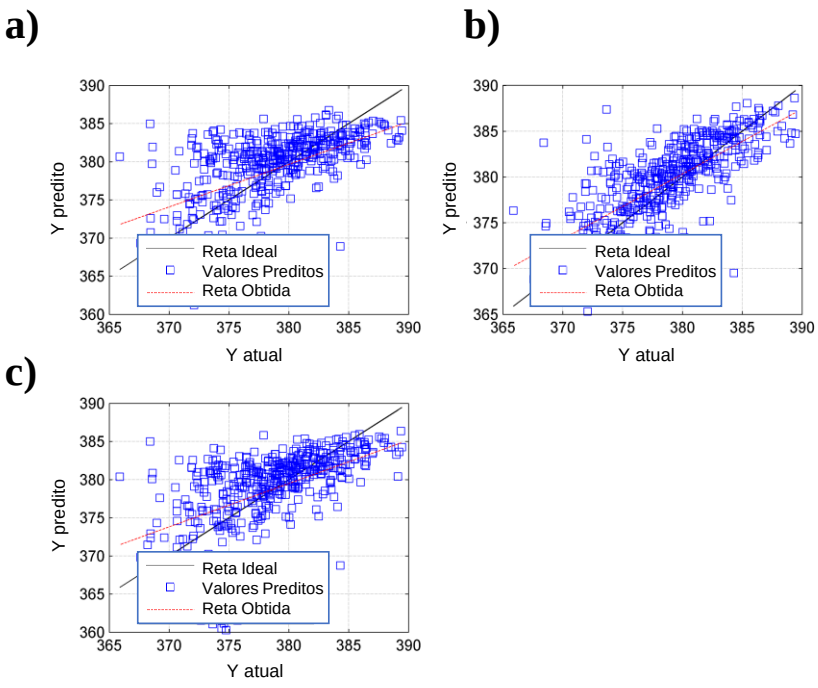

Figura 19. *Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros sem etapa de pré-processamento em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

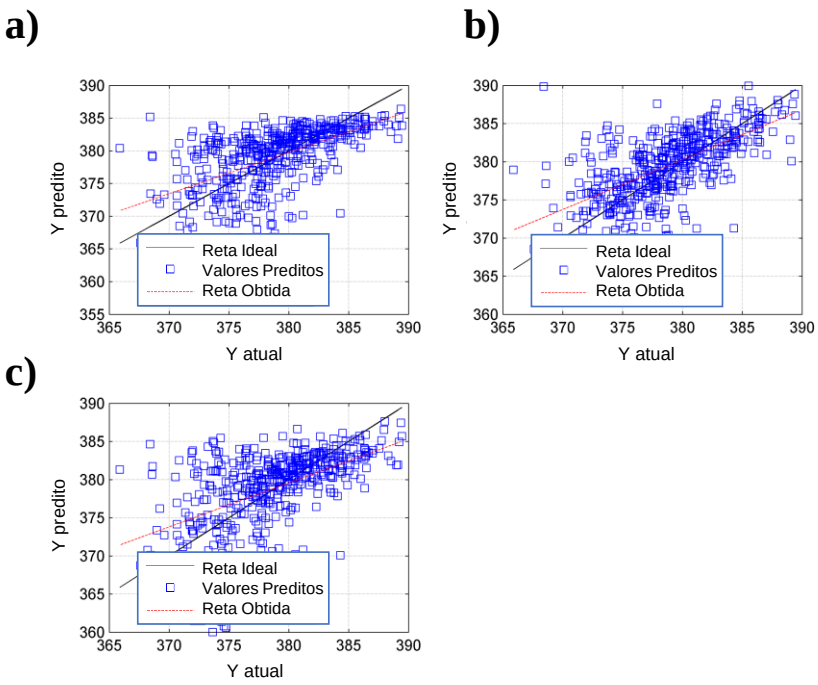

***Figura 20**. Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros pré-processados com 1ª derivada SG em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS.*

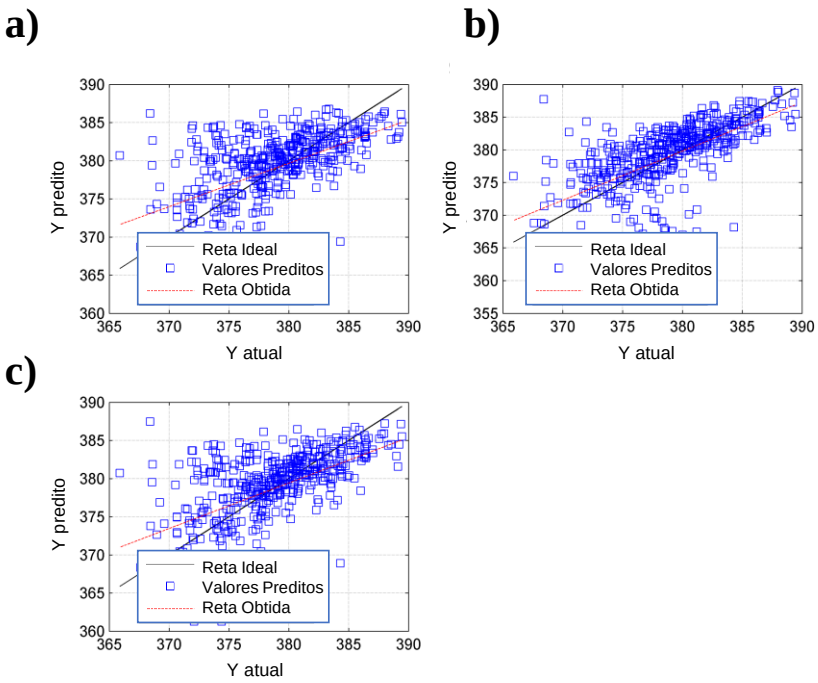

***Figura 21**. Gráfico linear dos valores de peso do comprimido preditos versus reais para espectros pré-processados com SNV em (a) SPA-MLR, (b) fSPA-MLR e (c) PLS.*

Como pode ser observado na **Tabela 3**, **Figura 19(a-c), Figura 20(a-c)** e **Figura 21(a-c)**, os melhores resultados foram obtidos a partir dos dados de espectros pré-processados com 1ª derivada SG ou sem etapa de pré-processamento.

De forma a avaliar a presença ou ausência de erros sistemáticos nos modelos, foram gerados gráficos de erro residual para cada um dos modelos estudados. A **Figura 22** mostra os erros residuais dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS, usando o espectro sem etapa de pré-processamento. A **Figura 23a, b** e **c** mostra os erros residuais dos três modelos, respectivamente, usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 24a, b** e **c**, mostra os erros residuais usando SNV como pré-processamento do espectro. É importante notar que nas **Figura 22(a-c), Figura 23(a-c)** e **Figura 24(a-c)**, as amostras são distribuídas aleatoriamente, o que indica a ausência de erro sistemático, além disso, o modelo *f*SPA-MLR apresenta erros menores ou semelhantes aos apresentados nos outros dois modelos (SPA-MLR e PLS).

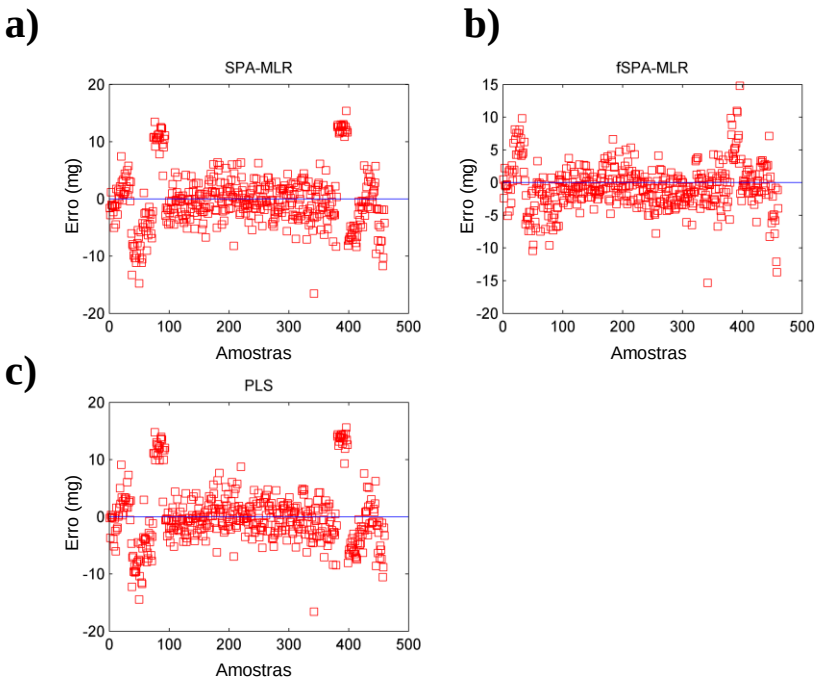

Figura 22. *Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de peso dos comprimidos.*

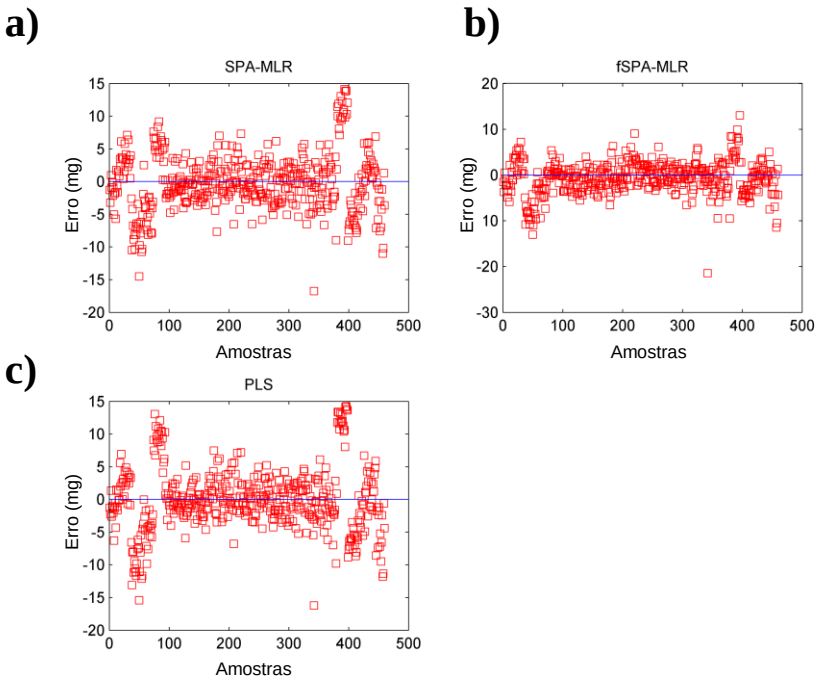

Figura 23. *Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de peso do comprimido.*

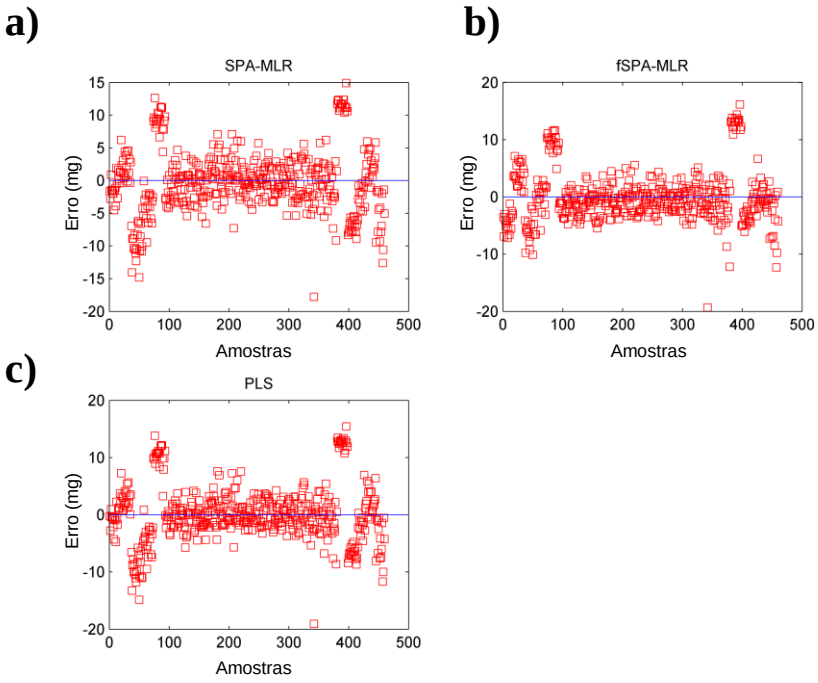

Figura 24. *Erros residuais dos modelos para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de peso dos comprimidos.*

5.1.3 Modelagem para predição da dureza do comprimido

A **Figura 25a, b e c**, mostra as variáveis selecionadas para construir o modelo MLR para o parâmetro de dureza do comprimido usando os espectros sem pré-processamento e pré-processados com primeira derivada da função SG e função SNV, respectivamente. Os círculos azuis indicam a variável selecionada do SPA-MLR e os quadrados verdes indicam a variável selecionada do *f*SPA-MLR.

As variáveis selecionadas na **Figura 25a** via *f*SPA-MLR (quadrados verdes) foram 1798, 1140, 922, 838, 824, 814 e 802 nm, já para o método SPA-MLR (círculos azuis) apenas as variáveis 1808 e 842 nm foram selecionadas. Na **Figura 25b** as variáveis 1856, 1292, 846, 834 e 806 nm foram selecionadas via *f*SPA-MLR e apenas a variável 1836 nm via SPA-MLR. Na **Figura 25c** 1808, 918 e 844 nm foram as variáveis selecionadas via *f*SPA-MLR e apenas 1158 nm via SPA-MLR. Observa-se que, neste caso do parâmetro de dureza do comprimido, a alteração dos métodos de pré-processamento gerou diferenças significativas na escolha das variáveis, já que poucas delas aparecem em mais de um dos métodos propostos.

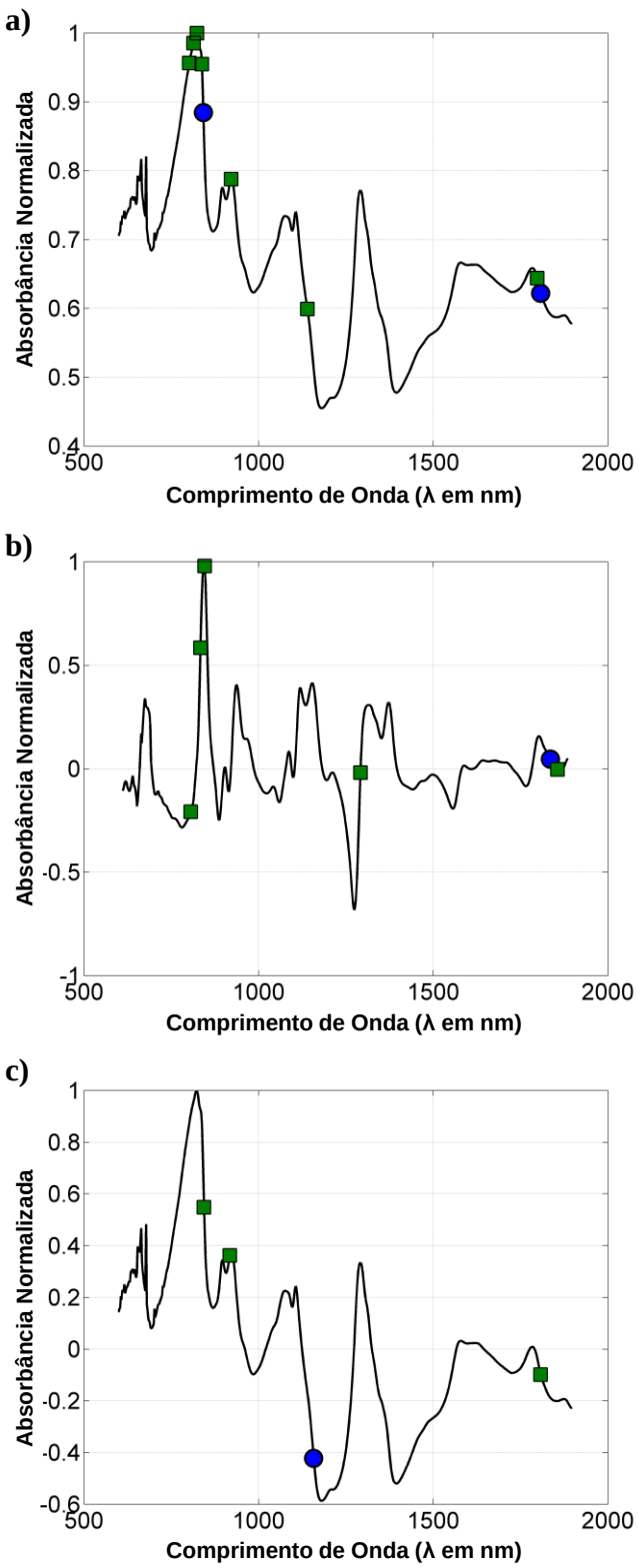

Figura 25. *(a) Espectros sem pré-processamento, (b) espectros pré-processados com primeira derivada da função SG e (c) espectros pré-processados com função SNV das variáveis selecionadas na construção do modelo MLR para o parâmetro de dureza do comprimido, do conjunto de calibração (círculos azuis indicam as variáves selecionadas em* SPA*-MLR e os quadrados verdes indicam as variáveis selecionadas em f*SPA*-MLR).*

Como pode ser visto na **Figura 25**, as variáveis selecionadas no *f*SPA-MLR (quadrados verdes) estão em maior número e mais representativas dos espectros, independentemente do tipo de pré-processamento, este resultado, juntamente com a **Tabela 4**, explica as fragilidades do SPA em sua versão sem a etapa de filtro. A **Tabela 4** resume os resultados de validação cruzada e predição externa obtidos com SPA-MLR, *f*SPA-MLR e PLS com os dados de espectros com e sem pré-processamento. Como pode ser visto, todos os métodos apresentaram desempenho semelhante e não eficiente, em termos de validação cruzada e predição externa. Os parâmetros R^2(cv) e R^2(pred) ficaram extremamente baixos, em torno de 0,300, ou seja, os modelos propostos não são capazes de predizer o parâmetro de dureza dos comprimidos.

Tabela 4*. Resultados de modelos de predição da dureza de cada comprimido em amostras de comprimidos farmacêuticos.*

Pré-processamento	**Modelo**	**RMSECV**	**R^2 (cv)**	**bias (cv)**	**RMSEP**	**R^2 (pred)**	**bias (pred)**	**Número de variáveis[a]**
Sem pré-processamento	**SPA-MLR**	1,224	0,373	-0,001	0,978	0,385	-0,001	2
	***f*SPA-MLR**	1,209	0,392	-0,002	1,047	0,382	0,137	7
	PLS	1,276	0,362	0,001	1,022	0,839	0,020	2
1ª derivada SG	**SPA-MLR**	1,234	0,363	0,002	0,954	0,380	-0,076	1
	***f*SPA-MLR**	1,204	0,395	0,004	1,067	0,326	-0,028	5
	PLS	1,355	0,271	0,002	1,027	0,453	0,103	1
SNV	**SPA-MLR**	1,243	0,354	0,001	0,937	0,392	-0,093	1
	***f*SPA-MLR**	1,207	0,391	0,003	0,994	0,412	0,056	3
	PLS	1,252	0,371	0,001	1,002	0,805	0,002	1

[a] *Variáveis latentes em PLS e variáveis espectrais em SPA-MLR e fSPA-MLR.*

A **Figura 26a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS usando o espectro sem etapa de pré-processamento. A **Figura 27a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 28a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando SNV como pré-processamento do espectro.

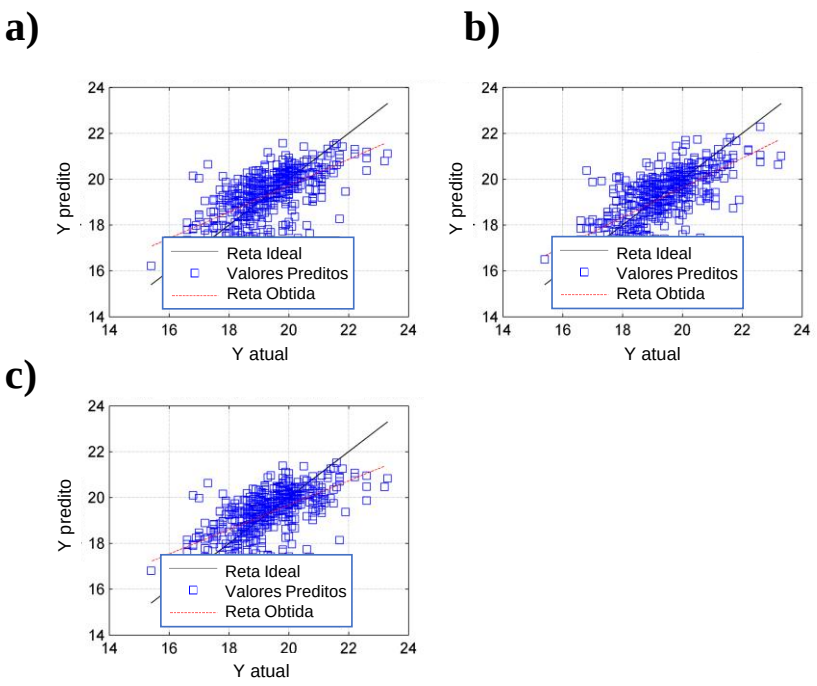

Figura 26. *Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros sem etapa de pré-processamento em (a)* SPA*-MLR, (b) f*SPA*-MLR e (c) PLS.*

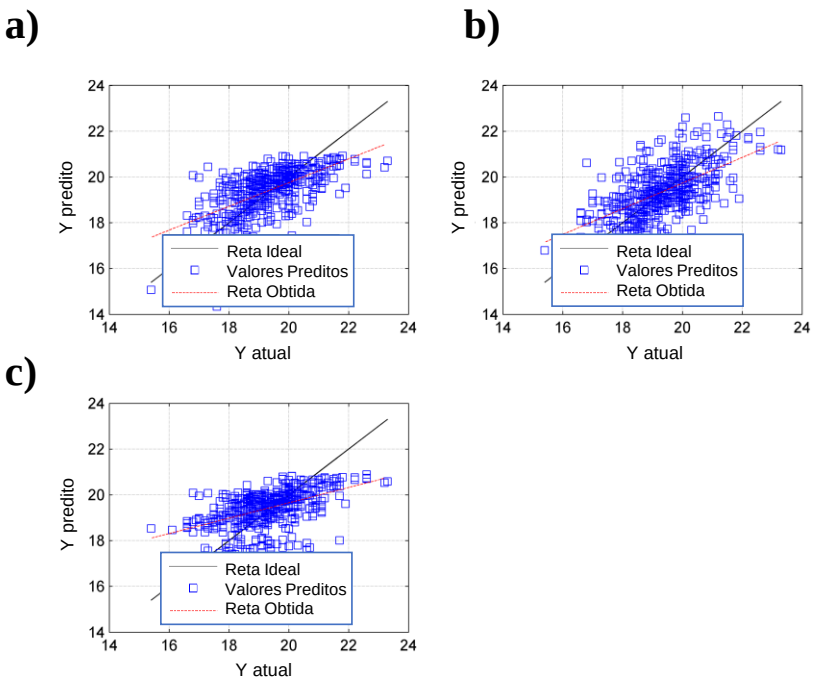

Figura 27. *Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros pré-processados com 1ª derivada SG em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

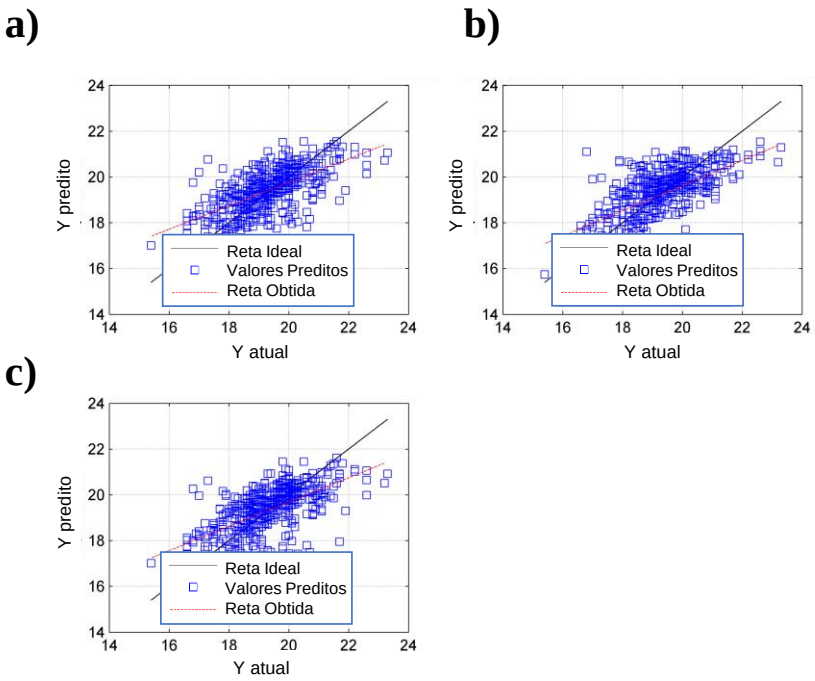

Figura 28. *Gráfico linear dos valores de dureza do comprimido preditos versus reais para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

Como pode ser observado na **Tabela 4, Figura 26(a-c), Figura 27(a-c)** e **Figura 28(a-c)**, os resultados obtidos indicam ineficiência de modelo, acarretando em incapacidade de predição do parâmetro.

De forma a avaliar a presença ou ausência de erros sistemáticos nos modelos, foram gerados gráficos de erro residual para cada um dos modelos estudados. A **Figura 29** mostra os erros residuais dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS, usando o espectro sem etapa de pré-processamento. A **Figura 30a, b** e **c** mostra os erros residuais dos três modelos, respectivamente, usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 31a, b** e **c**, mostra os erros residuais usando SNV como pré-processamento do espectro. É importante notar que nas **Figura 29(a-c), Figura 30(a-c)** e **Figura 31(a-c)**, as amostras são distribuídas aleatoriamente, o que indica a ausência de erro sistemático, além disso, o modelo *f*SPA-MLR apresenta erros menores ou semelhantes aos apresentados nos outros dois modelos (SPA-MLR e PLS).

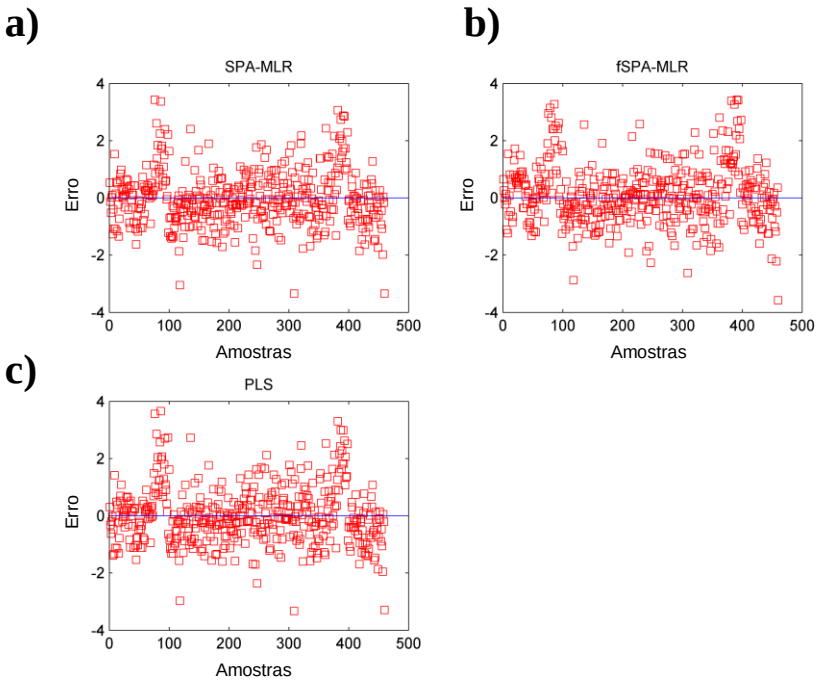

Figura 29. *Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de dureza dos comprimidos.*

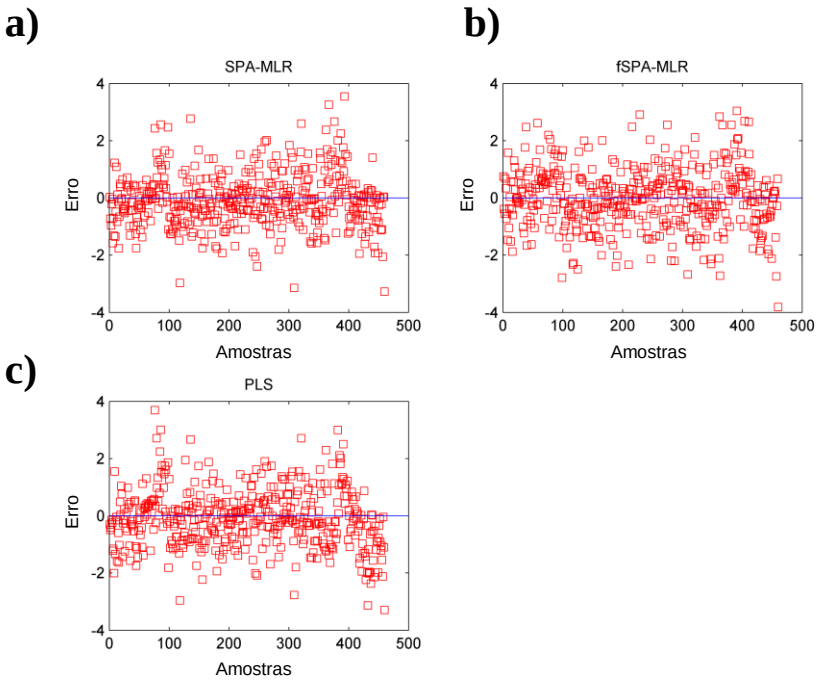

Figura 30. *Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de dureza do comprimido.*

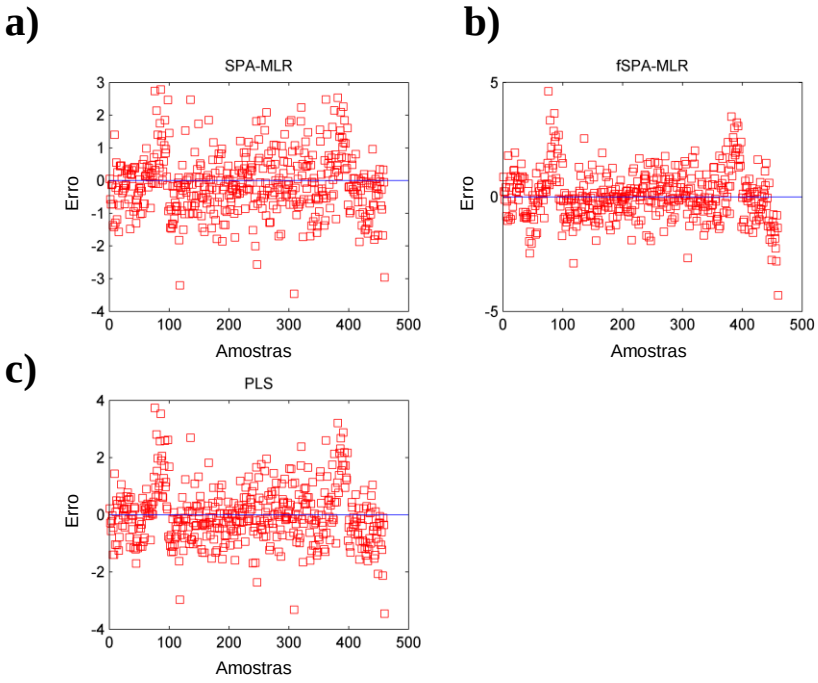

Figura 31. *Erros residuais dos modelos para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS, para o parâmetro de dureza dos comprimidos.*

5.2 Estudo de caso II

As **Figura 32a, b** e **c** mostram os espectros sem pré-processamento e pré-processados com 1ª derivada SG e SNV das 40 amostras de mistura de diesel/biodiesel do conjunto de calibração, respectivamente. Como pode ser visto na **Figura 32a**, os espectros sem pré-processamento exibiram características de linha de base sistemáticas e picos sobrepostos, que foram removidos pelo procedimento de primeira derivada da função SG (**Figura 32b**). Por sua vez, o procedimento SNV mostrado na **Figura 32c** elimina o efeito da mudança do caminho óptico em todo o espectro, o que gerou maior similaridade entre os espectros das amostras do mesmo conjunto de dados.

O valor de J-limite foi definido como o valor de J calculado na etapa de otimização que minimiza RMSECV ou RMSEV em um loop interno e o PRESS. A **Figura 33** mostra os valores J calculados para as amostras de mistura de diesel/biodiesel, tendo sido escolhidos como J-limite o menor valor de RMSEV, sendo 0,07 para os espectros sem pré-processamento, 0,05 para os espectros pré-processados com primeira derivada da função SG e 0,13 para espectros pré-processamento pela função SNV.

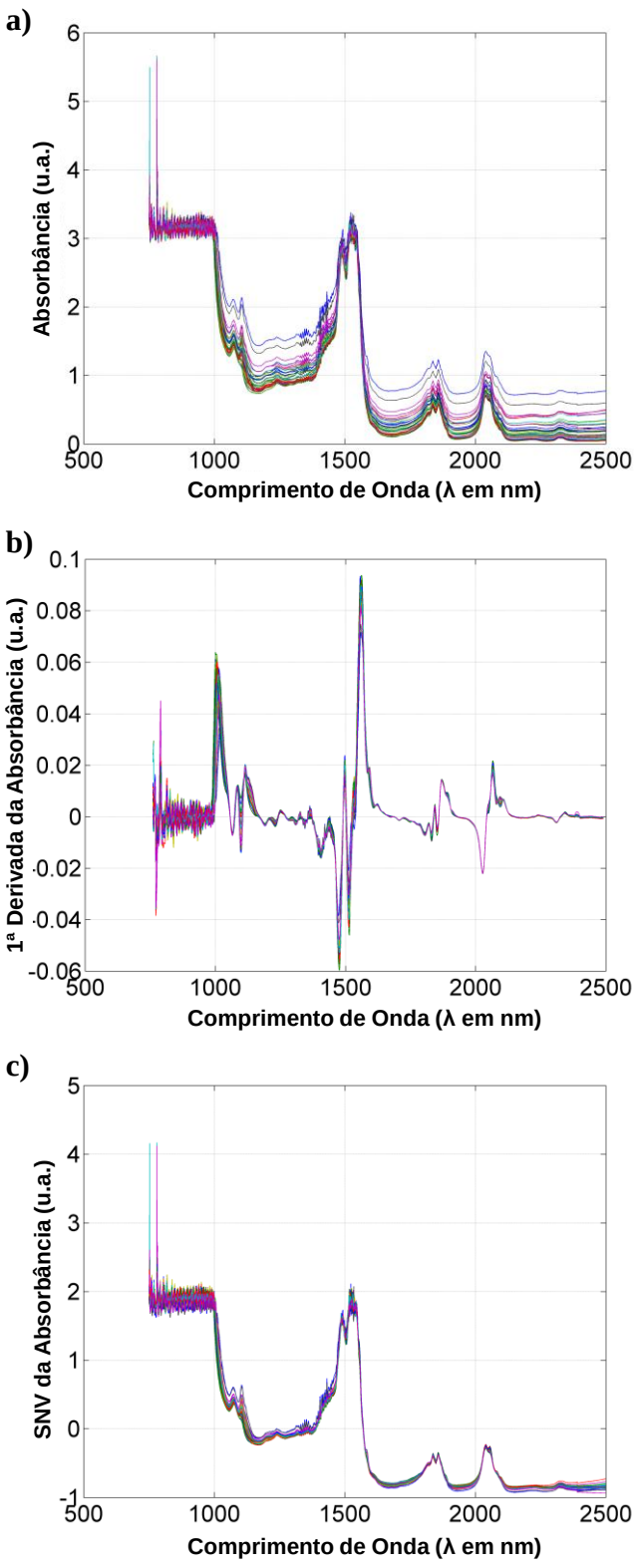

Figura 32. *(a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da função SG e (c) com pré-processamento na função SNV das 40 amostras de mistura diesel/biodiesel do conjunto de calibração.*

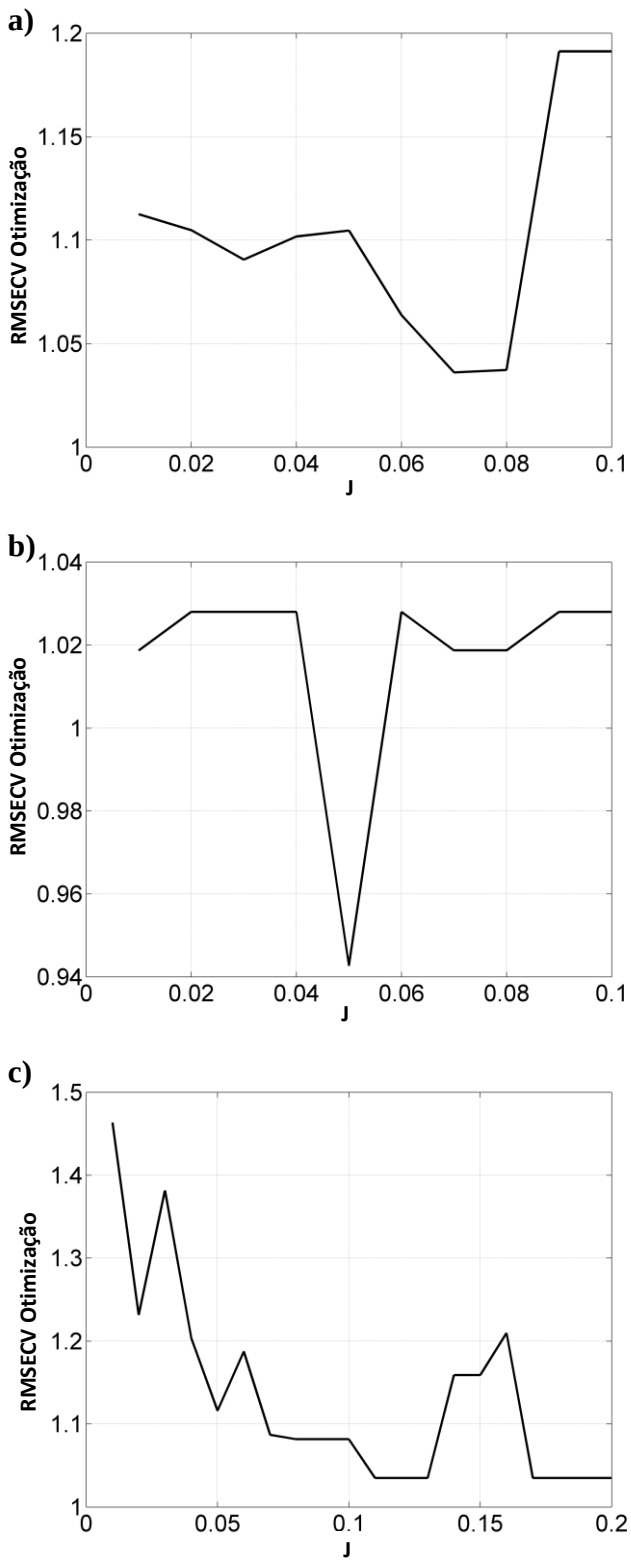

Figura 33. *Valor de J calculado para (a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da função SG e (c) com pré-processamento na função SNV das 40 amostras de mistura diesel/biodiesel do conjunto de calibração.*

Amostras de biodiesel e diesel têm como principal diferença a presença de oxigênio na estrutura do biodiesel e ausência no diesel, conforme pode ser observado na **Figura 34** [70]. Esta diferença estrutural pode causar alteração nas propriedades físicas e químicas das substâncias, bem como no seu espectro de infravermelho. Um espectro de diesel isolado não deveria conter bandas nas regiões características dos estiramentos de ligações C=O (1750 a 1735 nm) ou de ligações C-O (1300 a 1100 nm) [70]. Pode-se observar na **Figura 35** que na região de aproximadamente 1100 nm ocorre um aumento proporcional no tamanho da banda, que está relacionado ao aumento da concentração de biodiesel nas amostras de mistura de diesel/biodiesel. A **Figura 35a, b** e **c**, mostra as variáveis selecionadas para construir o modelo MLR usando espectros sem pré-processamento e pré-processados com 1ª derivada SG e SNV, respectivamente. Os círculos azuis indicam as variáveis selecionadas do SPA-MLR e os quadrados verdes indicam as variáveis selecionadas no *f*SPA-MLR.

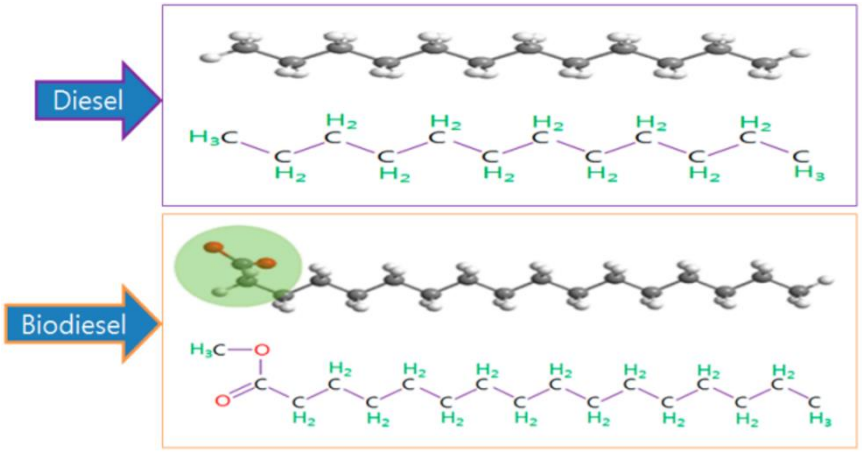

Figura 34. *Estrutura Molecular do Diesel e Biodiesel [70].*

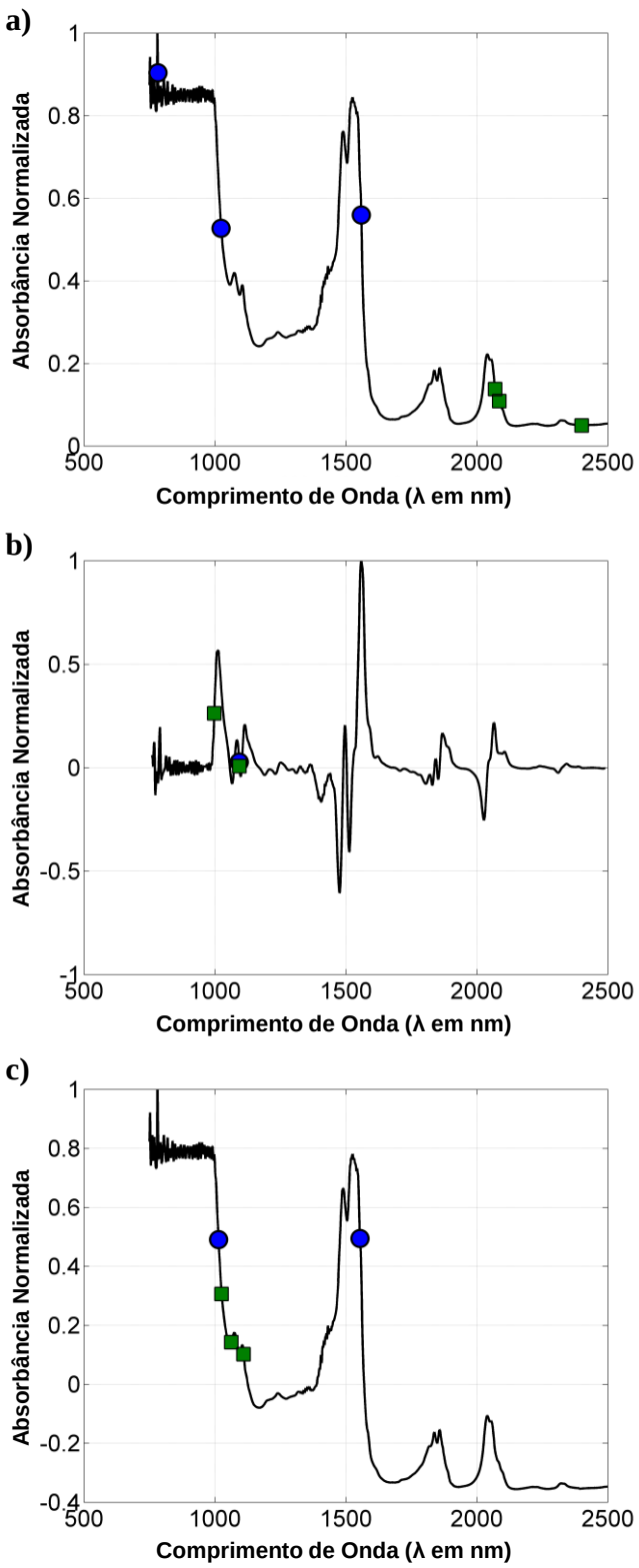

Figura 35. *(a) Espectros sem pré-processamento, (b) com pré-processamento de primeira derivada da funçãoSG e (c) com pré-processamento na função SNV das variáveis selecionadas para construir o modelo MLR a partir do conjunto de calibração (círculos azuis indicam as variáveis selecionadas em* SPA-*MLR; quadrados verdes indicam as variáveis selecionadas em f*SPA-*MLR).*

Como pode ser observado na **Figura 35a,** as variáveis selecionadas via *f*SPA-MLR (quadrados verdes) foram 2401, 2086 e 2070 nm, já para o método SPA-MLR (círculos azuis) as variáveis 1559, 1024 e 783 nm foram selecionadas. Na **Figura 35b** as variáveis1094 e 998 nm foram selecionadas via *f*SPA-MLR e apenas a variável 1093 nm via SPA-MLR. Na **Figura 35c** 1110, 1063 e 1026 nm foram as variáveis selecionadas via *f*SPA-MLR e 1554 e 1014 nm via SPA-MLR. Observa-se que a alteração dos métodos de pré-processamento gerou diferenças significativas na escolha das variáveis, já que poucas delas aparecem em mais de um dos métodos propostos. Além disto, uma quantidade menor de variáveis foi selecionada em ambos os métodos de seleção, diferentemente do Estudo de Caso I, onde um maior número de variáveis foram apresentadas como significativas.

A

Tabela 5 resume os resultados de validação cruzada e predição externa obtidos com SPA-MLR, *f*SPA-MLR e PLS com todos os dados de espectros de pré-processamento. Como pode ser visto, o *f*SPA-MLR supera os modelos SPA-MLR e PLS para quaisquer dados, com ou sem pré-processamento, em termos de validação cruzada e predição externa. Os parâmetros RMSECV e RMSEP foram menores no modelo *f*SPA-MLR e, R^2 (cv) e R^2 (pred) foram maiores que o SPA-MLR.

Tabela 5. *Resultados dos modelos de predição da porcentagem de biodiesel no diesel (% m/m) em amostras de mistura diesel/biodiesel.*

Pré-processamento	Modelo	RMSECV	R^2 (cv)	bias (cv)	RMSEP	R^2 (pred)	bias (pred)	Número de variáveis[a]
Sem pré-processamento	**SPA-MLR**	1,473	0,990	0,025	1,488	0,990	0,444	3
	***f*SPA-MLR**	1,036	0,995	0,019	1,158	0,994	-0,122	3
	PLS	1,972	0,947	0,032	1,553	0,992	0,753	7
1ª derivada SG	**SPA-MLR**	1,019	0,995	0,009	1,032	0,995	0,197	1
	***f*SPA-MLR**	0,943	0,996	-0,001	1,107	0,994	-0,018	2
	PLS	1,307	0,975	0,018	1,135	0,994	0,223	6
SNV	**SPA-MLR**	1,652	0,987	0,018	1,591	0,989	0,539	2
	***f*SPA-MLR**	1,035	0,995	0,021	1,121	0,994	0,221	3
	PLS	1,834	0,945	0,004	1,261	0,994	0,556	6

[a] *Variáveis latentes em PLS e variáveis espectrais em SPA-MLR e fSPA-MLR.*

A **Figura 36a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos modelos (a) SPA-MLR, (b) *f*SPA-MLR e (c) PLS usando o espectro sem etapa de pré-processamento. A **Figura 37a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 38a, b** e **c**, mostra o gráfico de valores preditos *versus* valores de referência dos três modelos usando SNV como pré-processamento do espectro.

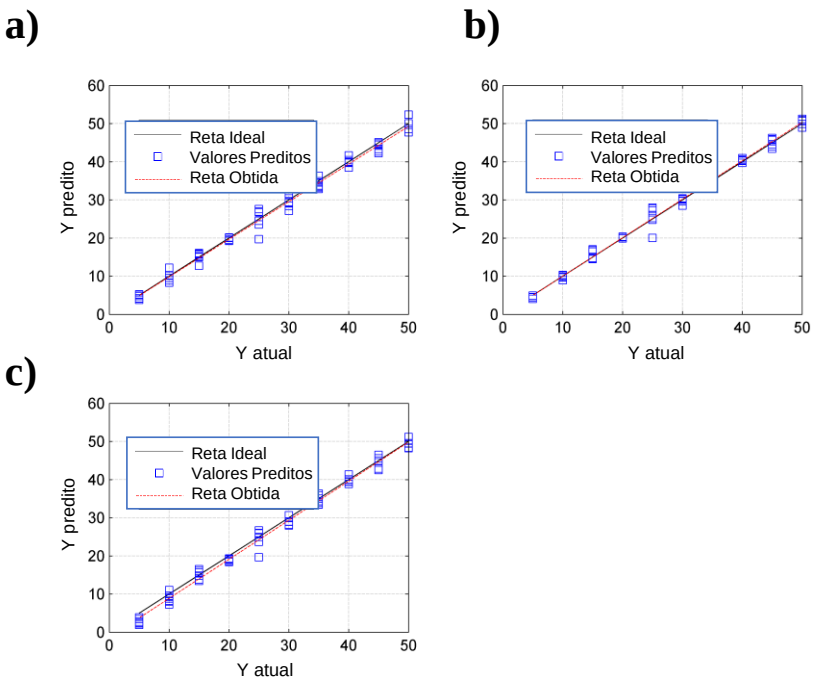

Figura 36. *Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros sem etapa de pré-processamento em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

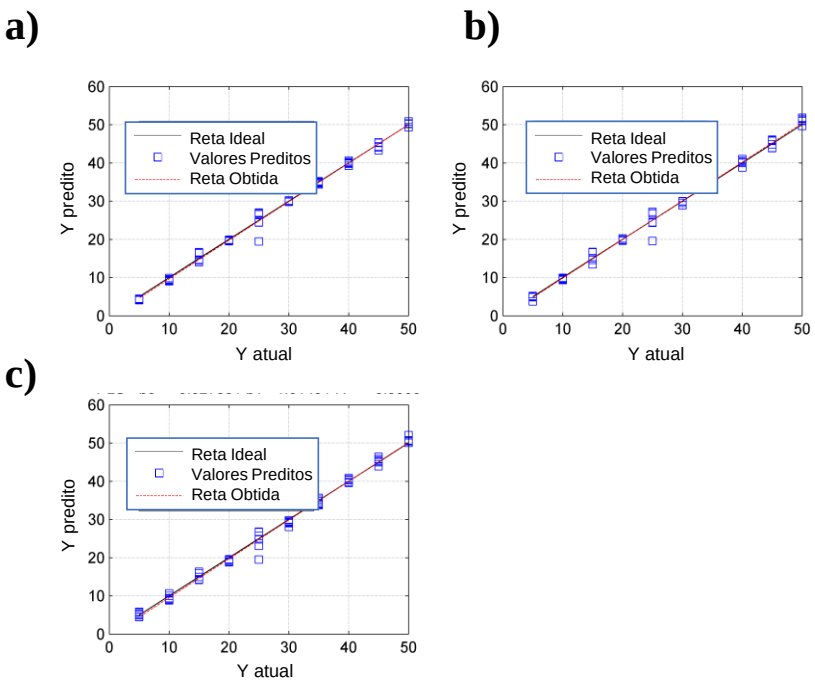

Figura 37. *Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros pré-processados com 1ª derivada SG em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

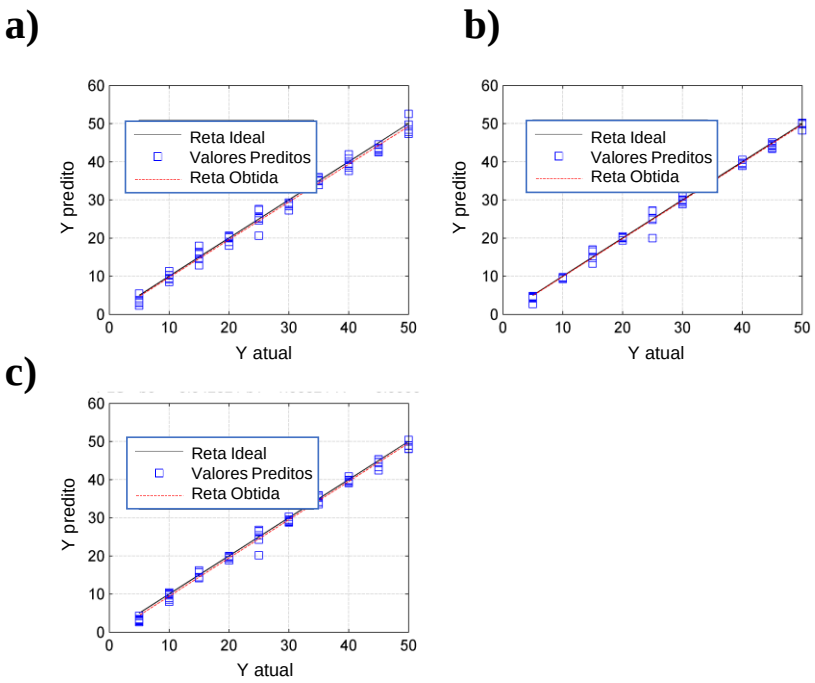

Figura 38. *Gráfico linear da porcentagem predita versus real de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS.*

Como pode ser observado na

Tabela 5, **Figura 36(a-c), Figura 37(a-c)** e **Figura 38(a-c)**, os melhores resultados foram obtidos a partir dos dados de espectros pré-processados com 1ª derivada SG.

De forma a avaliar a presença ou ausência de erros sistemáticos nos modelos, foram gerados gráficos de erro residual para cada um dos modelos estudados. A **Figura 39** mostra os erros residuais dos modelos (a) SPA-MLR, (b) fSPA-MLR e (c) PLS, usando o espectro sem etapa de pré-processamento. A **Figura 40a, b** e **c** mostra os erros residuais dos três modelos, respectivamente, usando a 1ª derivada SG do espectro como pré-processamento. Já a **Figura 41a, b** e **c**, mostra os erros residuais usando SNV como pré-processamento do espectro. É importante notar que nas **Figura 39(a-c), Figura 40(a-c)** e **Figura 41(a-c)**, as amostras são distribuídas aleatoriamente, o que indica a ausência de erro sistemático, além disso, o modelo *f*SPA-MLR apresenta erros menores ou semelhantes aos apresentados nos outros dois modelos (SPA-MLR e PLS). Os melhores resultados também são obtidos a partir dos dados de espectros pré-processados com 1ª derivada SG.

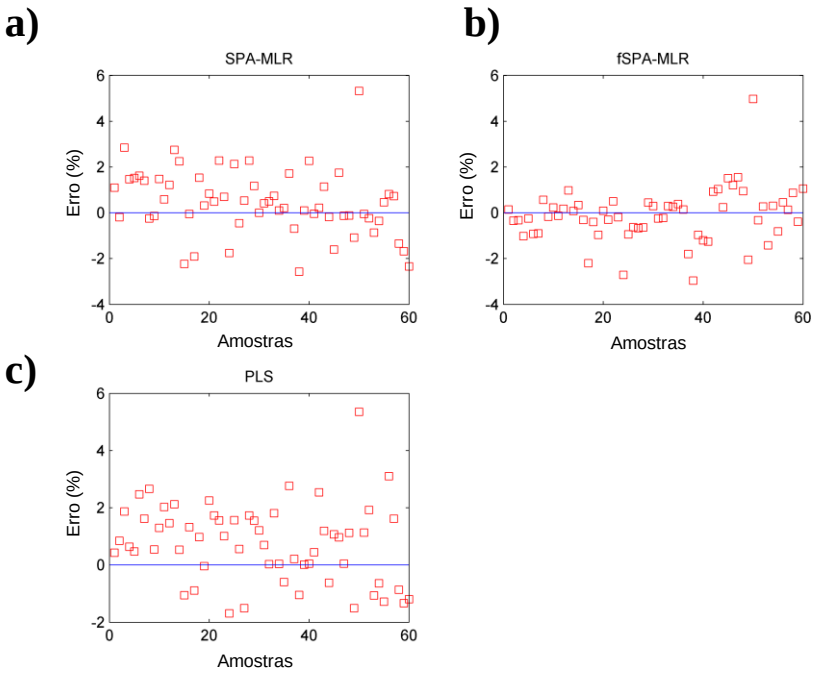

Figura 39. *Erros residuais dos modelos para espectros sem etapa de pré-processamento em (a)* SPA*-MLR, (b) f*SPA*-MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel.*

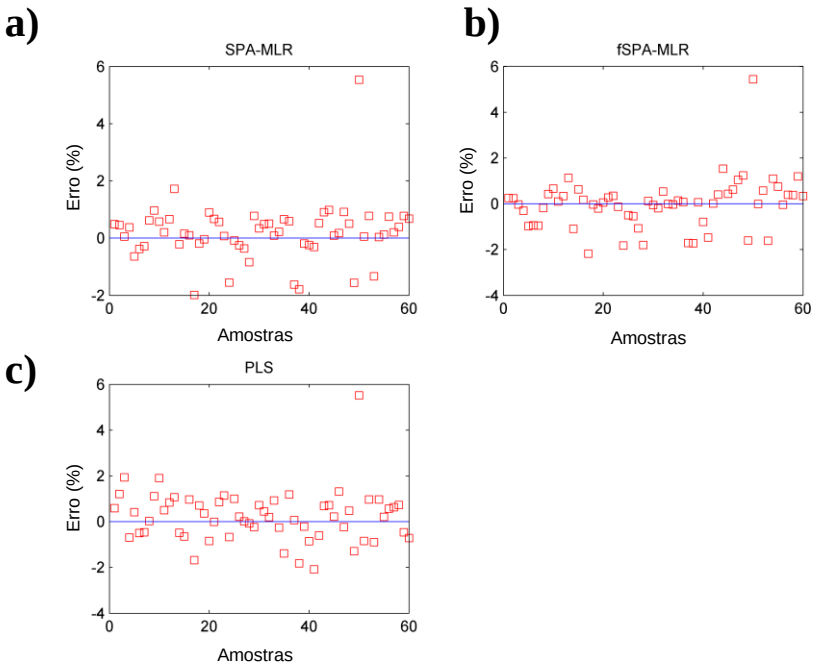

Figura 40. *Erros residuais dos modelos para espectros com pré-processamento de 1ª derivada SG em (a)* SPA*-MLR, (b) f*SPA*-MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel.*

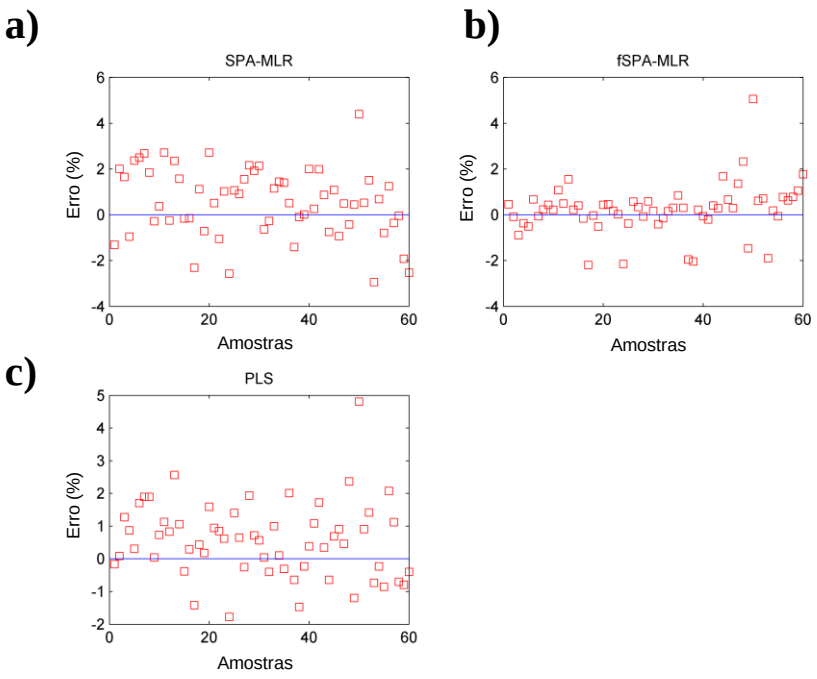

Figura 41. *Erros residuais dos modelos para espectros pré-processados com SNV em (a)* SPA-*MLR, (b) f*SPA-*MLR e (c) PLS para porcentagem de biodiesel em diesel (% m/m) em amostras de mistura de diesel/biodiesel.*

6 Conclusão

A adição da etapa de filtro à versão atual do algoritmo SPA proposta neste trabalho, *f*SPA, visa reduzir o número de variáveis não informativas antes da fase de projeção, bem como adicionar variáveis relevantes. E, desta forma, auxiliar o algoritmo na seleção das melhores variáveis nas etapas subsequentes, obtendo-se cadeias de variáveis tão informativas quanto possível. Os modelos *f*SPA-MLR apresentados para os dois estudos de caso realizados, superam o SPA-MLR original tanto na validação cruzada quanto na predição externa. Em comparação com o PLS, os modelos *f*SPA-MLR demonstram desempenho semelhante ou melhor. Além disso, os modelos *f*SPA-MLR oferecem resultados superiores, independentemente do algoritmo de pré-processamento testado, incluindo primeira derivada da função *Savitzky-Golay* (SG) e função *Standard Normal Variate* (SNV) ou mesmo em dados de espectros brutos.

7 Referências

[1] Y-H. Yun, H-D. Li, B-C. Deng, D-S. Cao, An overview of variable selection methods in multivariate analysis of near-infrared spectra, Trends Analyt Chem. 113 (2019) 102-115. https://doi.org/10.1016/j.trac.2019.01.018

[2] I. Barra, S.M. Haefele, R. Sakrabani, F. Kebede, Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advancess - A review, Trends Analyt Chem. 135 (2021) 116166. https://doi.org/10.1016/j.trac.2020.116166

[3] N. Kumar, A. Bansal, G.S. Sarma, R.K. Rawal, Chemometrics tools used in analytical chemistry: An overview, Talanta. 123 (2014) 186-199. http://dx.doi.org/10.1016/j.talanta.2014.02.003

[4] I. Guyon, A. Elisseeff, An Introduction to Variable and Feature Selection, J Mach Learn Res. 3 (2003) 1157-1182. https://doi.org/10.4028/www.scientific.net/AMR.1044-1045.1258

[5] S.F.C. Soares, A.A. Gomes, A.R. Galvão Filho, M.C.U. Araujo, R.K.H. Galvão, The successive projections algorithm, Trends Analyt Chem. 42 (2013). http://dx.doi.org/10.1016/j.trac.2012.09.006

[6] Z. Xiaobo, Z. Jiewen, M.J.W. Povey, M. Holmes, M. Hanpin. Variables selection methods in near-infrared spectroscopy, Anal Chim Acta. 667 (2010) 14-32. https://doi.org/10.1016/j.aca.2010.03.048

[7] C.M. Andersen, R. Bro, Variable selection in regression - a tutorial, J Chemometr. 24 (2010) 728-737. https://doi.org/10.1002/cem.1360

[8] X. Song, Y. Huang, H. Yan, Y. Xiong, S. Min, A novel algorithm for spectral interval combination optimization, Anal Chim Acta. 948 (2016) 19-29. http://dx.doi.org/10.1016/j.aca.2016.10.041

[9] F.N. Alenezi, T. Mehmood, Majority scoring based PLS filter mixture for variable selection in spectroscopic data, Chemometr Intell Lab. 212 (2021) 104282. https://doi.org/10.1016/j.chemolab.2021.104282

[10] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, NY, 2006.

[11] A.R. Webb, K.D. Copsey, Statistical Pattern Recognition, 3rd ed., Wiley, UK, 2011.

[12] R.M. Balabin, S.V. Smirnov, Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data, Anal Chim Acta. 692 (2011) 63-72. https://doi.org/10.1016/j.aca.2011.03.006

[13] J.A.F. Pierna, O. Abbas, V. Baeten, P. Dardenne, A Backward Variable Selection method for PLS regression (BVSPLS), Anal Chim Acta. 642 (2009) 89-93. https://doi.org/10.1016/j.aca.2008.12.002

[14] D.D.S. Fernandes, A.A. Gomes, G.B. da Costa, G.W.B. da Silva, G. Véras, Determination of biodiesel content in biodiesel/diesel blends using NIR andvisible spectroscopy with variable selection, Talanta. 87 (2011) 30-34. https://doi.org/10.1016/j.talanta.2011.09.025

[15] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics. 23 (2007) 2507. https://doi.org/10.1093/ bioinformatics/btm344

[16] U. Hörchner, J.H. Kalivas, Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection, Anal Chim Acta. 311 (1995) 1-13. https://doi.org/10.1016/0003-2670(95)00163-T

[17] W. Wu, Q. Guo, D.L. Massart, C. Boucon, S. Jong, Structure preserving feature selection in PARAFAC using a genetic algorithm and Procrustes analysis, Chemometr Intell Lab. 65 (2003) 83-95. https://doi.org/10.1016/S0169-7439(02)00105-3

[18] S. Gourvénec, X. Capron, D.L. Massart, Genetic algorithms (GA) applied to the orthogonal projection approach (OPA) for variable selection, Anal Chim Acta. 519 (2004) 11-21. https://doi.org/10.1016/j.aca.2004.05.023

[19] R.L. Carneiro, J.W.B. Braga, C.B.G. Bottoli, R.J. Poppi, Application of genetic algorithm for selection of variables for the BLLS method applied to determination of pesticides and metabolites in wine, Anal Chim Acta. 595 (2007) 51-58. https://doi.org/10.1016/j.aca.2006.12.023

[20] C.B. Lucasius, M.L.M. Beckers, G. Kateman, Genetic algorithms in wavelength selection: a comparative study, Anal Chim Acta. 286 (1994) 135-153. https://doi.org/10.1016/0003-2670(94)80155-X

[21] A. Niazi, R. Leardi, Genetic algorithms in chemometrics, J Chemometr. 26 (2012) 345-351. https://doi.org/10.1002/cem.2426

[22] A. Höskuldsson, Variable and subset selection in PLS regression, Chemometr Intell Lab. 55 (2001) 23-38. https://doi.org/10.1016/S0169-7439(00)00113-1

[23] L.F.B. Lira, M.S.Albuquerque, J.G.A. Pacheco, T.M. Fonseca, E.H.S. Cavalcanti, L. Stragevitch, M.F. Pimentel, Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel, Microchem J. 96 (2010) 126-131. https://doi.org/10.1016/j.microc.2010.02.014

[24] V. Centner, D-L. Massart, Elimination of Uninformative Variables for Multivariate Calibration, Anal Chem. 68 (1996) 3851-3858. https://doi.org/10.1021/ac960321m

[25] S. Ye, D. Wang, S. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemometr Intell Lab. 91 (2008) 194-199. https://doi.org/10.1016/j.chemolab.2007.11.005.

[26] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis, Anal Chim Acta. 699 (2011) 18-25. https://doi.org/10.1016/j.aca.2011.04.061

[27] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimisation: a powerful tool for wavelength selection, J Chemometr. 20 (2006) 146-147. https://doi.org/10.1002/cem.1002

[28] T. Mehmood, K.H. Liland, L. Snipen, S. Sæbø, A review of variable selection methods in Partial Least Squares Regression, Chemometr Intell Lab. 118 (2012) 62–69. http://dx.doi.org/10.1016/j.chemolab.2012.07.010

[29] J.A. Hageman, M. Streppel, R. Wehrens, L.M.C. Buydens, Wavelength selection with Tabu Search, J Chemometr. 17 (2003) 427-437. https://doi.org/10.1002/cem.782

[30] R.K.H. Galvão, M.C.U. Araújo, in: B. Walczak, R. Tauler, S. Brown (Eds), Comprehensive Chemometrics, Elsevier, Oxford, 2009, vol. 3, pp. 233-283.

[31] M. Forina, S. Lanteri, M.C.C. Oliveros, C.P. Millan, Selection of useful predictors in multivariate calibration, Anal Bioanal Chem. 380 (2004) 397-418. http://doi.org/10.1007/s00216-004-2768-x

[32] R.K.H. Galvão, M.F. Pimentel, M.C.U. Araújo, T. Yoneyama, V. Visani, Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry, Anal Chim Acta. 443 (2001) 107-115. https://doi.org/10.1016/S0003-2670(01)01182-5

[33] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemometr Intell Lab. 57 (2001) 65-73. https://doi.org/10.1016/S0169-7439(01)00119-8

[34] R.K.H. Galvão, M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, H.M.Paiva, A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm, Chemometr Intell Lab. 92 (2008) 83-91. https://doi.org/10.1016/j.chemolab.2007.12.004

[35] A.A. Gomes, M.R. Alcaraz, H.C. Goicoecheaa, M.C.U. Araújo, The Successive Projections Algorithm for interval selection in trilinear partial least-squares with residual bilinearization, Anal Chimica Acta. 811 (2014) 13-22. https://doi.org/10.1016/j.aca.2013.12.022

[36] H.A.Dantas Filho, E.S.O.N. de Souza, V. Visani, S.R.R.C. de Barros, T.C.B. Saldanha, M.C.U. Araújo, R.K.H. Galvão, Simultaneous Spectrometric Determination of Cu^{2+}, Mn^{2+} and Zn^{2+} in Polivitaminic/Polimineral Drug Using SPA and GA Algorithms for Variable Selection, J Brazil Chem Soc. 16, 1 (2005) 58-61. https://doi.org/10.1590/S0103-50532005000100009

[37] C. Abrahamsson, J. Johansson, A. Sparén, F. Lindgren. Comparison of different variable selection methods conducted on NIR transmission measurements on intact tablets, Chemometr Intell Lab. 69 (2003) 3-12. https://doi.org/10.1016/S0169-7439(03)00064-9

[38] A. Borin, R.J. Poppi. Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil, Vibrational Spectroscopy. 37 (2005) 27–32. https://doi.org/10.1016/j.vibspec.2004.05.003

[39] F. Liu, Y. Jiang, Y. He. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer, Analytica Chimica Acta. 635 (2009) 45–52. https://doi.org/10.1016/j.aca.2009.01.017

[40] A.C. Rencher, Methods of Multivariate Analysis, Wiley, New York, 2002.

[41] Brereton, R.G. Multivariate classification models. Journal of Chemometrics. 35:e3332, 2021. https://doi.org/10.1002/cem.3332

[42] Miller, N.J.; Miller, J.C. Statistics and Chemometrics for Analytical Chemistry. 6th Edition, Pearson Education Limited, Gosport, 221-224, 2010.

[43] Dardenne, P.; Sinnaeve, G.; Baeten, V. *Multivariate calibration and chemometrics for near infrared spectroscopy: which method? J. Near Infrared Spectrosc.* **8:** (4), 229-237, 2000.

[44] Geladi, P.; Kowalski, B.R. Partial least-squares regression: a tutorial, Anal Chim Acta. 185 (1986) 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

[45] Walmsley, A.D. Improved variable selection procedure for multivariate linear regression, Anal Chim Acta. 354 (1997) 225-232. https://doi.org/10.1016/S0003-2670(97)00450-9

[46] Gomes, A.A. Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS. UFPB/BC, João Pessoa, 2012.

[47] Soares, S.F.C. Um novo critério para seleção de variáveis usando o Algoritmo das Projeções Sucessivas. UFPB/BC, João Pessoa, 2010.

[48] Soares, F. Abordagens de seleção de variáveis para classificação e regressão em química analítica. UFRGS, Porto Alegre, 2017.

[49] Gomes, A.A. Algoritmo das Projeções Sucessivas para Seleção de Variáveis em Calibração de Segunda Ordem. UFPB/BC, João Pessoa, 2015.

[50] Miller, N.J.; Miller, J.C. (2010) Statistics and Chemometrics for Analytical Chemistry. 6th Edition, Pearson Education Limited, Gosport, 216-217.

[51] Youssef, H.; Sait, S.M.; Adiche, H. Evolutionary algorithms, simulated annealing end tabu search: a comparative study. Engineering Aplications of Artificial Inteligence, V. 14, p. 167-181, 2001.

[52] Pereira, G.W. Aplicação da Técnica de Recozimento Simulado em Problemas de Planejamento Florestal Multiobjetivo. UFMG, Belo Horizonte, 2004.

[53] Baskent, E.Z.; Jordan, G.A. Forest landscape management modelling using simulated annealing Forest Ecology and Management, Philadelphia, v. 165, p. 2945, 2002. https://doi.org/10.1016/S0378-1127(01)00654-5

[54] Kirkpatrick , S.; Gelatt, C.D.; Vecchi, M.P. Optimization by Simulated Annealing.

Science, New Series, Vol. 220, No. 4598, p. 671-680, 1983. https://doi.org/10.1016/B978-0-08-051581-6.50059-3

[55] Cerny, V. A thermodynamic approach to the traveling salesman problem: An efficient simulation. J. Optim. Theory Appl, 45, p. 41-51, 1985. https://doi.org/10.1007/BF00940812

[56] Costa Filho, C.A.; Poppi, R.J. Algoritmo Genético em Química. Química Nova 22:405, 1999. https://doi.org/10.1590/S0100-40421999000300019

[57] Neto, W.B. Parâmetros de qualidade de lubrificantes e óleo de oliva através de espectroscopia vibracional, calibração multivariada e seleção de variáveis. Unicamp, Campinas/SP, 2005.

[58] Tian, X.; Li, j.; Yi, S.; Jin, G.; Qiu, X.; Li, Y. Nondestructive determining the soluble solids content of citrus using near infrared transmittance technology combined with the variable selection algorithm, Artificial Intelligence in Agriculture. 4, 48-57, 2020. https://doi.org/10.1016/j.aiia.2020.05.001

[59] Paiva, H.M.; Soares, S.F.C.; Galvão, R.K.H.; Araújo, M.C.U. A graphical user interface for variable selection employing the Successive Projections Algorithm,

Chemometr Intell Lab. 118, 260-266, 2012. https://doi.org/10.1016/j.chemolab.2012.05.014

[60] Tan, W.; Sun, L.; Yang, F.; Che, W.; Ye, D.; Zhang, D.; Zou, B. Study on bruising degree classification of apples using hyperspectral imaging and GS-SVM, Optik. 154, 581-592, 2018. https://doi.org/10.1016/j.ijleo.2017.10.090

[61] Krepper, G.; Romeo, F.; Fernandes, D.D.S.; Diniz, P.H.G.D.; Araújo, M.C.U.; Di Nezio, M.S.; Pistonesi, M.F.; Centurión, M.E. Determination of fat content in chicken hamburgers using NIR spectroscopy and the Successive Projections Algorithm for interval selection in PLS regression (iSPA-PLS), Spectrochim Acta A. 189, 300-306, 2018. https://doi.org/10.1016/j.saa.2017.08.046.

[62] Wang, Y-J.; Jin, G.; Li, L-Q.; Liu, Y.; Kalkhajeh, Y.K.; Ning, J-M.; Zhang, Z-Z. NIR hyperspectral imaging coupled with chemometrics for nondestructive assessment of phosphorus and potassium contents in tea leaves, Infrared Physics & Technology. 108, 103365, 2020. https://doi.org/10.1016/j.infrared.2020.103365

[63] Breitkreitz, M.C.; Raimundo, I.M.; Rohwedder, J.J.; Pasquini, C.; Dantas Filho, H.A.; José, G.E.; Araújo, M.C. Determination of total sulfur in diesel fuel employing NIR spectroscopy and multivariate calibration. The Analyst, 128 9, 1204-7, 2003. https://doi.org/10.1039/B305265F

[64] Pereira, A.F.; Pontes, M.J.; Neto, F.F.; Santos, S.R.; Galvão, R.K.; Araujo, M.C. NIR spectrometric determination of quality parameters in vegetable oils using iPLS and variable selection. Food Research International, 41, 341-348, 2008. https://doi.org/10.1016/j.foodres.2007.12.013

[65] Pontes, M.J.; Rocha, A.M.; Pimentel, M.F.; Pereira, C.F. Determining the quality of insulating oils using near infrared spectroscopy and wavelength selection. Microchemical Journal, 98, 254-259, 2011. https://doi.org/10.1016/j.microc.2011.02.010

[66] Di Nezio, M.S.; Pistonesi, M.F.; Fragoso, W.D.; Pontes, M.J.C.; Goicoechea, H.C.; Araujo, M.C.U.; Band, B.S.F. Successive projections algorithm improving the multivariate simultaneous direct spectrophotometric determination of five phenolic compounds in sea water, Microchemical Journal, Volume 85, Issue 2, 194-200, ISSN 0026-265X, 2007. https://doi.org/10.1016/j.microc.2006.04.021.

[67] Goodarzi, M.T.; Freitas, M.P.; Jensen, R. Feature Selection and Linear/Nonlinear Regression Methods for the Accurate Prediction of Glycogen Synthase Kinase-3β Inhibitory Activities. Journal of chemical information and modeling, 49, 824-832, 2009. https://pubs.acs.org/doi/10.1021/ci9000103#

[68] Goudarzi, N.; Goodarzi, M.; Araujo, M.C.; Galvão, R.K. QSPR modeling of soil sorption coefficients (K(OC)) of pesticides using SPA-ANN and SPA-MLR. J Agric Food Chem. 57, 7153-7158, 2009. https://pubs.acs.org/doi/10.1021/jf9008839

[69] Silverstein, R. M.; Webster, F. X.; Identificação Espectrométrica de Compostos Orgânico, 6a ed, LTC: Rio de Janeiro, 1998.

[70] Ge, J. C.; Yoon, S; Choi, N. Using Canola Oil Biodiesel as an Alternative Fuel in Diesel Engines: A Review. Applied Sciences, 7, 881, 2017. 10.3390/app7090881.