



**UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA**

TESE DE DOUTORADO

**Seleção de variáveis robustas para transferência de
modelos de classificação empregando o Algoritmo das
Projeções Sucessivas**

Karla Danielle Tavares de Melo Milanez

João Pessoa – PB – Brasil

Agosto/2017



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

**Seleção de variáveis robustas para transferência de
modelos de classificação empregando o Algoritmo das
Projeções Sucessivas**

Karla Danielle Tavares de Melo Milanez*

Tese de Doutorado submetida ao Programa de Pós-Graduação em Química da Universidade Federal da Paraíba, como parte dos requisitos para obtenção do título de Doutor em Química.

Orientador: Prof. Dr. Márcio José Coelho de Pontes

***Bolsista CNPq**

João Pessoa – PB – Brasil

Agosto/2017

Catálogo na Publicação
Seção de Catalogação e Classificação

M637s Milanez, Karla Danielle Tavares de Melo.
Seleção de variáveis robustas para transferência de modelos de classificação empregando o algoritmo das projeções sucessivas / Karla Danielle Tavares de Melo Milanez. - João Pessoa, 2017.
132 f. : il.

Orientador: Dr. Márcio José Coelho de Pontes.
Tese (Doutorado) - UFPB/CCEN/PPGQ

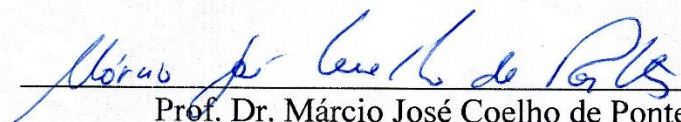
1. Química. 2. Algoritmo das Projeções Sucessivas (SPA). 3. Análise Discriminante Linear (LDA). 4. Etanol hidratado.
I. Título.

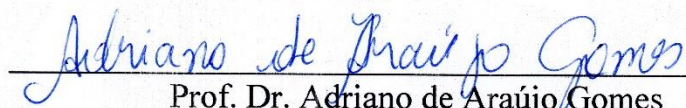
UFPB/BC

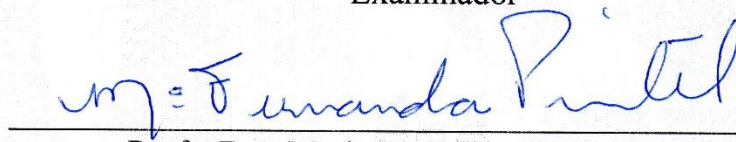
CDU - 54(043)

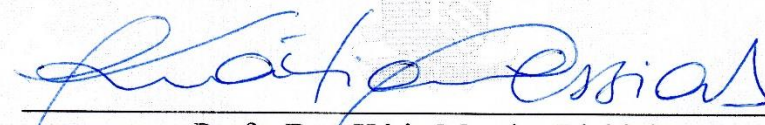
Seleção de variáveis robustas para transferência de modelos de classificação empregando o algoritmo das projeções sucessivas.

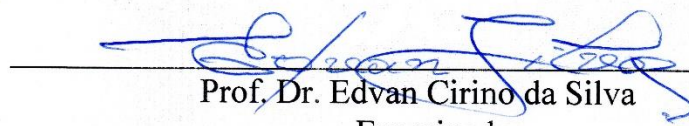
Tese de Doutorado apresentada pela aluna Karla Danielle Tavares de Melo Milanez e aprovada pela banca examinadora em 25 de agosto de 2017.


Prof. Dr. Márcio José Coelho de Pontes
Orientador/Presidente


Prof. Dr. Adriano de Araújo Gomes
Examinador


Profa. Dra. Maria Fernanda Pimentel Avelar
Examinadora


Profa. Dra. Kátia Messias Bichinho
Examinadora


Prof. Dr. Edvan Cirino da Silva
Examinador

*A todos que participaram direta ou
indiretamente deste trabalho e àqueles
que buscam, através da ciência, tornar
o mundo um pouco melhor.*

Com carinho, dedico.

AGRADECIMENTOS

- ❖ Agradeço a Deus, por me manter firme e consciente ao longo de toda jornada acadêmica;
- ❖ Aos meus pais, Luis Carlos e Iricelia, por todo amor e dedicação que me guiaram até aqui;
- ❖ À minha irmã Laura, pelo amor e compreensão;
- ❖ Ao meu esposo Lucas Milanez, por todo amor, apoio, sugestões e paciência;
- ❖ Ao Professor Márcio José Coelho Pontes, pela orientação, dedicação e confiança;
- ❖ Ao Professor Roberto Kawakami Harrop Galvão, pela preciosa ajuda nos momentos decisivos deste trabalho;
- ❖ Ao Professor Mário César Ugulino Araújo, pela acolhida acadêmica;
- ❖ Ao Laboratorio FIA na Universidad Nacional del Sur da Argentina, pelo registro dos espectros UV-Vis e de fluorescência molecular. Em especial à Professora Beatriz Susana Fernandez Band e aos queridos amigos Danielle Silva Nascimento e Matias Insausti, pela parceria e por toda ajuda na realização da parte experimental;
- ❖ Ao Laboratório de Combustíveis (LAC) da Universidade Federal de Pernambuco (UFPE), pelo fornecimento das amostras de etanol hidratado combustível;
- ❖ À EMBRAPA, Centro Nacional de Pesquisa de Algodão, pelo registro dos espectros NIR. Em especial Dr. Everaldo Paulo de Medeiros;
- ❖ A Adenilton Camilo da Silva e José Eduardo Matos Paz, pelo trabalho realizado previamente com as amostras de etanol hidratado combustível e Thiago César Araújo Nóbrega, pela ajuda com as medidas no UV-Vis;
- ❖ Às amigas Marília Gabriela, Clarissa Maia e Danielle Nascimento e ao meu amigo Felipe, pela amizade, carinho, paciência e tantos momentos compartilhados;
- ❖ Às amigas Anna Luiza e Welma e ao amigo Sófacles pelo carinho, compreensão, presteza e por compartilhar conhecimentos e anseios de um aluno de doutorado;
- ❖ A todos os demais membros do LAQA, pelo apoio, amizade e troca de experiências;

- ❖ Aos Professores Edvan Cirino da Silva e Kátia Messias Bichinho, pelas excelentes sugestões na banca de qualificação;
- ❖ À Universidade Federal da Paraíba;
- ❖ Ao CNPq, pela concessão da bolsa;
- ❖ A todos que contribuíram indiretamente para a realização deste trabalho.

RESUMO

Este trabalho propõe dois novos critérios para a seleção de variáveis robustas para transferência de classificação empregando o algoritmo das projeções sucessivas (SPA). Estas variáveis são utilizadas para construir modelos baseados na análise discriminante linear (LDA) robustos às diferenças entre as respostas dos instrumentos envolvidos ou às condições experimentais. Para isso, amostras de transferência são incluídas no cálculo do custo para cada subconjunto de variáveis consideradas. Os métodos propostos são avaliados para quatro conjuntos de dados que envolvem identificação de adulteração de etanol hidratado combustível (EHC) e azeite de oliva extra virgem (AOEV). Para investigar EHC, foi utilizada espectroscopia na região do infravermelho próximo (NIR). No estudo do AOEV, foram utilizadas espectrometria UV-Vis, espectrometria de fluorescência molecular e imagens digitais. Em todos os casos, os melhores resultados de transferência de classificação empregando os dois critérios, obtidos para um conjunto de teste medido no instrumento secundário, foram comparados com a padronização direta (DS) e a padronização direta em etapas (PDS). Quando algum dos critérios foi aplicado ao conjunto de teste medido no instrumento secundário, a acurácia do modelo aumentou cerca de 50%, 33%, 3% e 12% para os dados NIR, UV-Vis, de emissão de fluorescência e de imagens digitais, respectivamente. Esses resultados são compatíveis, em algumas vezes superiores, àqueles obtidos pelos métodos de padronização, demonstrando que, quando as diferenças entre as respostas instrumentais não apresentavam magnitude drasticamente elevada (NIR e UV-Vis), um dos critérios propostos pode ser utilizado para construir modelos robustos como alternativa à padronização de respostas instrumentais para transferência de classificação. Para os dados com grande diferença entre as respostas de cada instrumento (fluorescência e imagens digitais), os espectros precisaram ser corrigidos por meio de padronização DS e PDS para realizar a transferência de classificação efetivamente. Os resultados sugerem que a abordagem proposta é uma alternativa promissora à recalibração completa do modelo ou métodos de padronização, especialmente se os instrumentos primário e secundário não estiverem localizados no mesmo laboratório, se as amostras estiverem deterioradas, quando se tem dificuldade no transporte das amostras ou se o instrumento principal não estiver mais disponível.

Palavras-chave: Transferência de classificação multivariada; modelagem robusta; algoritmo das projeções sucessivas; métodos de padronização; etanol hidratado combustível; azeite de oliva extra virgem.

ABSTRACT

This work proposes two new criteria for selection of robust variables for classification transfer employing the successive projections algorithm (SPA). These variables are used to construct models based on linear discriminant analysis (LDA) that are robust to the differences between the responses of the instruments involved or to the experimental conditions. For this purpose, transfer samples are included in the calculation of the cost for each subset of variables under consideration. The proposed methods are evaluated for four datasets involving identification of adulteration of hydrated ethanol fuel (HEF) and extra virgin olive oil (EVOO). To investigate HEF, near infrared (NIR) spectroscopy was used. In the EVOO study, were used UV-Vis spectrometry, molecular fluorescence spectrometry and digital images. In all cases, better classification transfer results using the two criteria, obtained for a test set measured in the secondary instrument, were compared with direct standardization (DS) and piecewise direct standardization (PDS). When one of the criteria was applied to the test set measured in the secondary instrument, the accuracy of the model increased by about 50%, 33%, 3% and 12% for NIR, UV-Vis, fluorescence emission and digital imaging data, respectively. These results are compatible, sometimes superior to those obtained by the standardization methods, demonstrating that, when the differences between the instrumental responses did not present a drastically high magnitude (NIR and UV-Vis), either of the criteria proposed can be used for building robust models as an alternative to the standardization of spectral responses for transfer of classification. For the data with a large difference between the responses of each instrument (fluorescence and digital images), the spectra needed to be corrected with DS and PDS standardization to perform classification transfer effectively. The results suggest that the proposed approach is a promising alternative to full recalibration of the model or standardization methods, especially if the primary and secondary instruments are not located in the same laboratory, if the samples are deteriorated, when it is difficult to transport the samples or if the primary instrument is no longer available.

Keywords: Multivariate classification transfer; robust modeling; successive projections algorithm; standardization methods; hydrated ethyl fuel; extra virgin olive oil.

LISTA DE FIGURAS

Figura 2.1 - Primeira e segunda PCs obtidas a partir da combinação linear das variáveis originais. (a) As linhas pontilhadas na cor verde correspondem às projeções de uma amostra nas PCs (escores), enquanto as linhas tracejadas na cor amarela representam as projeções nas variáveis originais. (b) O cosseno dos ângulos θ_1 e θ_2 representa os <i>loadings</i> . -----	30
Figura 2.2 - Representação do modelo PLS-DA utilizando o algoritmo PLS 1 para duas classes de objetos. -----	31
Figura 2.3 - Análise discriminante linear para duas categorias de objetos definidas com classe A (▲) e classe B (●). A linha na cor azul representa a função discriminante linear obtida na representação. -----	32
Figura 2.4 - Sequência de projeções realizadas pelo SPA com $Kc = 3$ e $j = 5$. Resultado da primeira (a) e segunda (b) iteração. Neste exemplo, o subconjunto de variáveis selecionadas inicia com x_1 e deverá ser $\{x_1, x_2, x_5\}$. -----	34
Figura 2.5 - Níveis de energia eletrônica molecular e as transições envolvidas. -----	49
Figura 2.6 - Diagrama de Jablonski. -----	51
Figura 2.7 - Imagem analógica colorida e sua representação em imagem digital. -----	53
Figura 4.1 - Sistema usado na aquisição das imagens de AOEV. -----	63
Figura 4.2 - Procedimento de seleção de amostras de transferência a partir do conjunto de treinamento medido no instrumento primário. -----	65
Figura 5.1 - Espectros NIR das 102 amostras de EHC registrados no instrumento primário. Amostras não-adulteradas (—) e amostras adulteradas (—). -----	69
Figura 5.2 - Espectros NIR pré-processados das amostras de EHC medidas nos instrumentos primário (—) e secundário (—). -----	70
Figura 5.3 - Gráfico dos escores em PC 1 $\times$ PC 2 para as 80 amostras de EHC do conjunto de teste. Amostras não-adulteradas (●) e adulteradas (■) no instrumento primário, amostras não-adulteradas (●) e adulteradas (■) no instrumento secundário. -----	70

Figura 5.4 - Gráfico dos *loadings* para as duas primeiras PCs.----- 71

Figura 5.5 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de EHC. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.----- 73

Figura 5.6 - Espectros NIR médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios *C1*: ▲, *C2*: ● e SPA na forma usual: ■.----- 75

Figura 5.7 - Valores de *threshold* obtidos pelos modelos SPA-LDA usual (a), critério *C1* (b) e critério *C2* (c) na classificação das amostras de EHC medidas no instrumento secundário. Amostras não-adulteradas (●) e adulteradas (■).----- 76

Figura 5.8 - Número de erros obtidos pelo SPA-LDA com os critérios *C1* e *C2* e SPA-LDA na forma usual após procedimentos de padronização, aplicados às amostras de teste (S) de EHC. A largura da janela empregada na padronização PDS é indicada entre parênteses. ----- 77

Figura 5.9 - Espectros de absorção molecular no UV-Vis das amostras de AOEV antes (a) e após pré-processamento (b). Espectros adquiridos nos instrumentos A (—) e B (—). ----- 80

Figura 5.10 - Espectros UV-Vis pré-processados registrados nos instrumentos A (a) e B (b) com indicação das duas classes de AOEV. Amostras não-adulteradas (—) e adulteradas (—) no instrumento A; amostras não-adulteradas (—) e adulteradas (—) no instrumento B. ----- 81

Figura 5.11 - Gráficos dos escores nas quatro primeiras PCs para as amostras de AOEV do conjunto de teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento A; Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento B. ----- 81

Figura 5.12 - Gráfico dos *loadings* para as quatro primeiras PCs.----- 82

Figura 5.13 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de AOEV analisadas por espectrometria UV-Vis. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.----- 85

Figura 5.14 - Espectros UV-Vis médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios *C1*: ●, *C2*: ✱ e SPA na forma usual: ■.-----87

Figura 5.15 - Gráficos dos escores da FD 1 obtidos para os modelos SPA-LDA usual (a), critério *C1* (b) e critério *C2* (c). Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento secundário.-----87

Figura 5.16 - Número de erros obtidos pelo SPA-LDA com os critérios *C1* e *C2* e SPA-LDA na forma usual após procedimentos de padronização, aplicados às amostras de teste (S). A largura da janela empregada na padronização PDS é indicada entre parênteses. -----88

Figura 5.17 - Espectros de emissão de fluorescência molecular das amostras de AOEV medidas nos instrumentos A (a) e B (b). Amostras não-adulteradas (—) e adulteradas (—). -----91

Figura 5.18 - Gráficos dos escores em PC 1 × PC 2 para as amostras de AOEV do conjunto teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (■) medidas no instrumento A; Amostras não-adulteradas (✕) e adulteradas (▲) medidas no instrumento B. -----92

Figura 5.19 - Gráfico dos *loadings* para as duas primeiras PCs. -----92

Figura 5.20 - Espectros de emissão de fluorescência molecular das amostras de AOEV do conjunto de teste medido no instrumento secundário antes (a) e após (b) suavização. -----95

Figura 5.21 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de AOEV analisadas por espectrometria de fluorescência molecular. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros. -----96

Figura 5.22 - Espectros médios de fluorescência molecular do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios *C1*: ●, *C2*: ◆ e SPA na forma usual: ■.-----97

Figura 5.23 - Espectros de emissão de fluorescência molecular das amostras de teste (S). São apresentados os espectros brutos e pré-processados antes e após procedimentos de padronização usando $N_{trans} = 8$. ----- 100

Figura 5.24 - Gráficos dos escores obtidos pelo modelo SPA-LDA para as amostras de teste (S) antes (a) e depois (b) da padronização DS com $N_{trans} = 4$. Valores estimados pelo modelo PLS-DA para as amostras de teste (S) antes (c) e após (d) a padronização DS dos dados pré-processados usando $N_{trans} = 6$. Amostras não-adulteradas (✕) e adulteradas (Δ). As linhas horizontais são os limites de decisão estabelecidos pelos modelos SPA-LDA e PLS-DA.-- 101

Figura 5.25 - Histogramas RGB das amostras de AOEV adquiridos na webcam (a) e no smartphone (b). As amostras não-adulteradas estão representadas pelas cores mais escuras e as adulteradas pelas cores mais claras. Os valores indicados para as componentes RGB correspondem à faixa de níveis de cores presentes nos histogramas. ----- 103

Figura 5.26 - Gráficos dos escores nas seis primeiras PCs para as amostras de AOEV do conjunto teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (■) medidas na webcam; Amostras não-adulteradas (◐) e adulteradas (◑) medidas no smartphone. ----- 104

Figura 5.27 - Gráfico dos *loadings* para as seis primeiras PCs. ----- 105

Figura 5.28 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios C1 e C2) aplicados às amostras de teste (S) de AOEV medidas por imagens digitais a partir do smartphone. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.----- 108

Figura 5.29 - Histogramas médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA nos modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios C1: ■, C2: ● e SPA na forma usual: ■. ----- 109

Figura 5.30 - Gráficos dos escores obtidos pelo modelo SPA-LDA para as amostras de teste antes (a) e depois (b) da padronização PDS (5) com $N_{trans} = 6$. Valores estimados pelo modelo PLS-DA para as amostras de teste antes (c) e após (d) PDS (7) dos dados usando $N_{trans} = 4$. Amostras não-adulteradas (◐) e adulteradas (◑). As linhas horizontais são os limites de decisão estabelecidos pelos modelos SPA-LDA e PLS-DA. ----- 112

LISTA DE TABELAS

Tabela 2.1 - Tabela de contingência 2×2 . -----	44
Tabela 2.2 - Principais critérios de qualidade para cada tipo de azeite de oliva. -----	47
Tabela 2.3 - Regiões espectrais do infravermelho.-----	48
Tabela 4.1 - Número de amostras de treinamento e teste selecionadas pelo KS para as duas classes de EHC. N representa o número de amostras em cada classe. -----	64
Tabela 4.2 - Número de amostras de treinamento e teste selecionadas pelo KS para as duas classes de AOEV. N representa o número de amostras em cada classe.-----	65
Tabela 5.1 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário. O número entre parênteses apresenta a quantidade de variáveis empregadas no modelo. N indica o número de amostras de treinamento e teste em cada classe de EHC. -----	71
Tabela 5.2 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário aplicado às amostras de teste medidas no instrumento secundário. N indica o número de amostras de teste em cada classe de EHC. -----	72
Tabela 5.3 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios $C1$ e $C2$ empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário. -----	73
Tabela 5.4 - Erros de classificação obtidos pelos critérios $C1$ e $C2$ aplicados às amostras de teste (S) medidas no instrumento secundário. N indica o número de amostras de teste em cada classe de EHC. -----	74
Tabela 5.5 - Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização direta (DS) e padronização direta em etapas (PDS). A largura da janela utilizada no modelo PDS é apresentada entre parênteses. -----	76
Tabela 5.6 - Valores de sensibilidade, especificidade e acurácia obtidos pelos modelos aplicados ao conjunto de teste (S).-----	78

Tabela 5.7 - Valores medidos e de referência para K_{232} , K_{270} e ΔK obtidas para as sete amostras de AOEV não-adulteradas. Os valores são obtidos após primeira (medida em 270 nm) e segunda (medida em 232 nm) diluição com ciclohexano. Cada lote é representado por um número. -- 79

Tabela 5.8 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento A. O número de variáveis empregadas no modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 83

Tabela 5.9 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento B. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 83

Tabela 5.10 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com base no instrumento primário aplicado às amostras de teste medidas no instrumento secundário. N indica o número de amostras de teste em cada classe de AOEV. - 84

Tabela 5.11 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios $C1$ e $C2$ empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário. ----- 84

Tabela 5.12 - Erros de classificação obtidos pelos critérios $C1$ e $C2$ aplicados às amostras de teste (S) medidas no instrumento secundário. N indica o número de amostras de teste em cada classe de AOEV. ----- 86

Tabela 5.13 - Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização DS e PDS. A largura da janela utilizada no modelo PDS é apresentada entre parênteses. ----- 88

Tabela 5.14 - Valores de sensibilidade, especificidade e acurácia obtidos pelos modelos aplicados ao conjunto de teste (S). ----- 89

Tabela 5.15 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento A. O número de variáveis empregadas no modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 93

Tabela 5.16 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento B. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 93

Tabela 5.17 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário (A) aplicado às amostras de teste medida no instrumento secundário. N indica o número de amostras de teste em cada classe de AOEV. ----- 94

Tabela 5.18 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios $C1$ e $C2$ empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário. ----- 95

Tabela 5.19 - Resultados de classificação obtidos com pelo modelo PLS-DA construído para os dados registrados no instrumento primário A. O número de variáveis latentes empregadas no modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 98

Tabela 5.20 - Resultados de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. N indica o número de amostras de teste em cada classe de AOEV. ----- 98

Tabela 5.21 - Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização dos dados sem e com pré-processamento. A largura da janela empregada na padronização PDS é indicada entre parênteses. ----- 99

Tabela 5.22 - Erros de classificação obtidos pelo modelo PLS-DA aplicado às amostras de teste (S) após padronização dos dados sem e com pré-processamento. A largura da janela empregada na padronização PDS é indicada entre parênteses. ----- 99

Tabela 5.23 - Resultados de classificação obtidos com os modelos SPA-LDA na forma usual construídos para as imagens registradas na webcam. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 105

Tabela 5.24 - Resultados de classificação obtidos com os modelos SPA-LDA na forma usual construídos para as imagens registradas no smartphone. O número de variáveis empregadas em

cada modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 106

Tabela 5.25 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com base nas imagens registradas no instrumento primário (webcam) aplicado às amostras de teste medidas no instrumento secundário (smartphone). N indica o número de amostras de teste em cada classe de AOEV. ----- 106

Tabela 5.26 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios $C1$ e $C2$ empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário. ----- 107

Tabela 5.27 - Resultados de classificação obtidos com o modelo PLS-DA construído para as imagens registradas no instrumento primário. O número de variáveis latentes empregadas no modelo é apresentado entre parênteses. N indica o número de amostras de treinamento e teste em cada classe de AOEV. ----- 110

Tabela 5.28 - Resultados de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. N indica o número de amostras de teste em cada classe de AOEV. ----- 110

Tabela 5.29 - Erros de classificação obtidos pelos modelos SPA-LDA e PLS-DA aplicados às amostras de teste (S) analisadas no instrumento secundário após padronização. A largura da janela empregada na padronização PDS é indicada entre parênteses. ----- 111

LISTA DE SIGLAS E ABREVIATURAS

ANN: Artificial Neural Networks

ANP: Agência Nacional de Petróleo, Gás Natural e Biocombustíveis

ANVISA: Agência Nacional de Vigilância Sanitária

AOCS: American Oil Chemist's Society

AOEV: Azeite de Oliva Extra Virgem

CCFO: Codex Committee on Fats and Oils

DS: Direct Standardization

EAC: Etanol Anidro Combustível

EHC: Etanol Hidratado Combustível

EHCP: Etanol Hidratado Combustível Premium

EVOO: Extra Virgin Olive Oil

FIR: Far Infrared

FN: Falso-Negativos

FP: Falso Positivos

GA: Genetic Algorithm

HEA: Hydrated Ethyl Alcohol

IOOC: International Olive Oil Council

IV: Infravermelho

JK: Kack-Knife

KS: Kennard-Stone

LDA: Linear Discriminant Analysis

MAPA: Ministério de Agricultura, Pecuária e Abastecimento

MIR: Medium Infrared

MLR: Multiple Linear Regression

NIR: Near Infrared

OSC: Orthogonal Signal Correction

PC: Principal Component

PCA: Principal Component Analysis

PCR: Principal Component Regression

PDS: Piecewise Direct Standardization

PLS: Partial Least Squares

PLS-DA: Partial Least Squares Discriminant Analysis

RGB: Red-Green-Blue

RMSET: Root Mean Square Error of Prediction in the Transfer Set

RMSEV: Root Mean Square Error of Validation

RP: Reconhecimento de Padrões

SA: Simulated Annealing

SBC: Slope/Bias Correction

SG: Savitzky-Golay

SIMCA: Soft Independent Modeling of Class Analogy

SPA: Sucessive Projections Algorithm

TCC: Taxa de Classificação Correta

TN: Verdadeiro-Negativos

TOP: Transfer by Orthogonal Projection

TP: Verdadeiro-Positivos

TW: Transformada Wavelet

UV-Vis: Radiação Ultravioleta-Visível

SUMÁRIO

1	INTRODUÇÃO	24
1.1	Aspectos gerais	24
1.2	Objetivos	25
1.2.1	Objetivo geral	25
1.2.2	Objetivos específicos	25
2	FUNDAMENTAÇÃO TEÓRICA	28
2.1	Pré-processamento dos dados	28
2.2	Técnicas de reconhecimento de padrões	29
2.2.1	Análise de componentes principais	30
2.2.2	Análise discriminante pelos mínimos quadrados parciais	30
2.2.3	Análise discriminante linear	31
2.3	Seleção de variáveis	33
2.3.1	Algoritmo das projeções sucessivas	34
2.4	Transferência de modelos multivariados	36
2.4.1	Métodos de padronização das respostas instrumentais	37
2.4.1.1	Padronização direta	38
2.4.1.2	Padronização direta em etapas	38
2.4.2	Aumento da robustez do modelo	40
2.4.3	Seleção das amostras de transferência	42
2.4.3.1	Seleção de amostras pelo algoritmo KS	43
2.5	Parâmetros de desempenho	43
2.6	Estudos de caso	45
2.6.1	Etanol Hidratado Combustível (EHC)	45
2.6.2	Azeite de oliva extra virgem (AOEV)	46
2.6.3	Técnicas analíticas	48
2.6.3.1	Espectroscopia no infravermelho	48
2.6.3.2	Espectrometria UV-Vis	49
2.6.3.3	Espectrometria de fluorescência molecular	50

2.6.3.4	Imagens digitais-----	52
---------	-----------------------	----

3 SPA E TRANSFERÊNCIA DE MODELOS ----- 56

3.1	Os critérios propostos -----	56
-----	------------------------------	----

4 METODOLOGIA ----- 60

4.1 Conjuntos de amostras----- 60

4.1.1	<i>Etanol Hidratado Combustível</i> -----	60
-------	---	----

4.1.2	<i>Azeite de oliva extra virgem (Espectrometria UV-Vis e de espectroscopia de fluorescência molecular)</i> -----	60
-------	--	----

4.1.2.1	Determinação da extinção específica pela absorção na região do ultravioleta -----	60
---------	---	----

4.1.3	<i>Azeite de oliva extra virgem - Imagens digitais</i> -----	61
-------	--	----

4.2 Registro do sinal----- 61

4.2.1	<i>Espectros NIR</i> -----	61
-------	----------------------------	----

4.2.2	<i>Espectros UV-Vis</i> -----	62
-------	-------------------------------	----

4.2.3	<i>Espectros de emissão de fluorescência molecular</i> -----	62
-------	--	----

4.2.4	<i>Histogramas de cores das imagens</i> -----	63
-------	---	----

4.3 Seleção de amostras ----- 64

4.4 Procedimento quimiométrico e software ----- 66

4.5 Notação----- 67

5 RESULTADOS E DISCUSSÃO ----- 69

5.1 Espectroscopia NIR ----- 69

5.1.1	<i>Análise espectral</i> -----	69
-------	--------------------------------	----

5.1.2	<i>Análise exploratória</i> -----	70
-------	-----------------------------------	----

5.1.3	<i>Modelos SPA-LDA</i> -----	71
-------	------------------------------	----

5.1.4	<i>Transferência dos modelos SPA-LDA</i> -----	72
-------	--	----

5.1.5	<i>Considerações finais</i> -----	78
-------	-----------------------------------	----

5.2 Determinação da extinção específica pela absorção na região do ultravioleta ----- 79

5.3 Espectrometria UV-Vis ----- 79

5.3.1	<i>Análise espectral</i> -----	79
-------	--------------------------------	----

5.3.2	<i>Análise exploratória</i>	81
5.3.3	<i>Modelos SPA-LDA</i>	83
5.3.4	<i>Transferência dos modelos SPA-LDA</i>	84
5.3.5	<i>Considerações finais</i>	89
5.4	Espectrometria de fluorescência molecular	91
5.4.1	<i>Análise espectral</i>	91
5.4.2	<i>Análise exploratória</i>	92
5.4.3	<i>Modelos SPA-LDA</i>	93
5.4.4	<i>Transferência dos modelos SPA-LDA</i>	94
5.4.5	<i>Transferência de classificação usando métodos de padronização</i>	97
5.4.6	<i>Considerações finais</i>	101
5.5	Imagens digitais	103
5.5.1	<i>Análise dos histogramas</i>	103
5.5.2	<i>Análise exploratória</i>	104
5.5.3	<i>Modelos SPA-LDA</i>	105
5.5.4	<i>Transferência dos modelos SPA-LDA</i>	107
5.5.5	<i>Transferência dos modelos SPA-LDA usando procedimentos de padronização</i>	109
5.5.6	<i>Considerações finais</i>	112
6	CONCLUSÕES	115
	REFERÊNCIAS	118
	APÊNCICE	130
	Produção científica	130

Capítulo I

INTRODUÇÃO

1 INTRODUÇÃO

1.1 Aspectos gerais

Modelos multivariados têm sido amplamente utilizados em problemas analíticos envolvendo análises quantitativas e qualitativas de uma variedade de materiais ^[1-11]. Além dos métodos bem estabelecidos na literatura, novas ferramentas quimiométricas são constantemente propostas a fim de facilitar a obtenção de informação e fornecer modelos multivariados cada vez mais robustos.

A eficiência de modelos multivariados está associada a fatores como quantidade de amostras analisadas, bem como sua composição química e/ou física e representatividade, instrumentação utilizada e condições experimentais ^[12]. Além disso, as etapas de construção e validação do modelo também são fundamentais para alcançar um desempenho adequado. Ainda que esses fatores sejam cuidadosamente controlados, há situações em que um modelo não consegue determinar ou classificar corretamente amostras analisadas sob novas condições experimentais ou em um instrumento não envolvido na etapa de modelagem. Situações como essas podem ocorrer, por exemplo, em empresas que possuam vários equipamentos (possivelmente de modelos e/ou fabricantes diferentes) distribuídos entre suas instalações ou quando se deseja empregar um instrumento de laboratório de alta qualidade para construir um modelo que posteriormente venha a ser empregado com instrumentos de menor custo.

A inaplicabilidade do modelo é resultado de diferenças entre as respostas instrumentais, ocasionadas por alterações na composição química e/ou física das amostras ou mudanças tanto no ambiente como na função de resposta do instrumento ^[13]. Uma das soluções para esse problema é a recalibração completa do modelo a partir de um conjunto de amostras analisado sob as novas condições analíticas, que na prática é um procedimento dispendioso, que exige tempo e rigoroso controle das condições experimentais ^[13-15]. Alternativamente, a transferência de classificação pode ser realizada usando métodos de padronização das respostas instrumentais ou aumento da robustez dos modelos ^[14].

Os métodos de padronização das respostas espectrais baseiam-se na modificação da resposta de um instrumento (instrumento secundário) para fazê-la corresponder à resposta do instrumento para o qual o modelo foi desenvolvido (instrumento primário) ^[13, 14]. Entre os métodos de padronização, destacam-se a padronização direta (DS: *Direct Standardization*) e padronização direta em etapas (PDS: *Piecewise Direct Standardization*) ^[16]. Em ambos os casos, os sinais de um conjunto representativo de amostras, conhecidas como amostras de

transferência, são registrados nos instrumentos envolvidos e utilizados no procedimento de padronização.

A robustez é um dos parâmetros mais utilizados para avaliar o desempenho dos modelos multivariados, sendo relacionada à capacidade de fornecer resultados confiáveis diante de variações não incluídas na modelagem ^[17]. A robustez pode ser obtida pela incorporação de todas as fontes de variação importantes ainda na etapa de modelagem, pré-processamento adequado dos dados e seleção de variáveis menos sensíveis às condições experimentais ou variações instrumentais ^[18]. Esses procedimentos funcionam quando as diferenças entre as funções de resposta são relativamente pequenas, eliminando a necessidade dos métodos de padronização ^[18, 19].

Nesse contexto, este trabalho propõe dois novos critérios para seleção de variáveis robustas para transferência de classificação empregando o algoritmo das projeções sucessivas (SPA). O SPA tem sido aplicado com sucesso, acoplado à análise discriminante linear (LDA), em problemas de classificação envolvendo técnicas instrumentais, sobretudo para minimizar problemas de multicolinearidade em modelos LDA. Nos casos envolvendo transferência de classificação, a seleção de variáveis robustas pode ajudar na construção de modelos LDA com menor sensibilidade em relação às diferenças entre os instrumentos.

1.2 Objetivos

1.2.1 Objetivo geral

Adaptar o algoritmo das projeções sucessivas (SPA) para seleção de variáveis robustas que, acopladas à análise discriminante linear (LDA), possa ser empregado na transferência de modelos de classificação, fornecendo modelos com menor sensibilidade em relação às diferenças instrumentais.

1.2.2 Objetivos específicos

- Desenvolver, a partir de um determinado instrumento, modelos de classificação multivariada baseados no SPA-LDA, para serem aplicados na classificação de amostras analisadas em um instrumento ou condição de análise diferente;
- Adaptar, através de dois diferentes critérios, o SPA-LDA para seleção de variáveis robustas com propósito de transferência de modelos de classificação;
- Avaliar o desempenho dos critérios propostos (C1 e C2) para as seguintes situações:

- a. *Classificação de amostras de EHC adulteradas com metanol e não-adulteradas utilizando a espectroscopia na região do infravermelho próximo (NIR);*
 - b. *Classificação de amostras de AOEV adulteradas com óleo de soja e não-adulteradas empregando espectrometria UV-VIS, espectrometria de fluorescência molecular e imagens digitais;*
- Avaliar, através de diferentes parâmetros de desempenho (sensibilidade, especificidade e acurácia), as estratégias de transferência baseadas nos critérios propostos aplicadas aos estudos de caso, comparando-as com métodos de padronização (DS e PDS).

Capítulo II
FUNDAMENTAÇÃO
TEÓRICA

2 FUNDAMENTAÇÃO TEÓRICA

A Quimiometria pode ser definida como o conjunto de métodos matemáticos, estatísticos e gráficos utilizados para extrair o máximo de informação de dados químicos ^[7], sendo comumente utilizada para exame de dados, correção de problemas espectrais, seleção de variáveis robustas e construção de modelos utilizados em análises químicas.

A seguir serão discutidas algumas das ferramentas quimiométricas mais utilizadas para análise qualitativa de dados analíticos.

2.1 Pré-processamento dos dados

Medidas instrumentais normalmente são afetadas por variações relacionadas com mudanças nas propriedades químicas e/ou físicas das amostras, instabilidade instrumental ou espalhamento da radiação durante a análise ^[20]. Essas variações podem afetar o sinal analítico ocasionando, por exemplo, diferenças entre a magnitude dos sinais, variação sistemática de linha de base e/ou ruído instrumental, comprometendo a eficiência dos modelos multivariados ^[21].

Essas variações podem ser corrigidas com a utilização de técnicas de pré-processamentos dos dados, que podem ser aplicadas às variáveis ou às amostras previamente à etapa de modelagem.

O pré-processamento das variáveis pode ser realizado pela centralização dos dados na média, escalonamento e auto-escalonamento ^[21]. A centralização dos dados na média consiste em subtrair os elementos de cada linha pela média da sua coluna (variável). No escalonamento, os elementos do vetor coluna são divididos pelo desvio-padrão da respectiva coluna. O auto-escalonamento, que consiste em centrar na média e escalonar os dados, visa atribuir os mesmos pesos às variáveis em consideração ^[22].

As amostras podem ser pré-processadas utilizando-se, entre outras técnicas, normalização, ponderação, correção de linha de base e suavização ^[21, 22]. As técnicas de normalização e ponderação removem variações sistemáticas relativas às amostras através da divisão ou multiplicação, respectivamente, de cada variável do vetor amostra por uma constante ^[23, 21].

A correção de linha de base normalmente é realizada por meio da derivada ^[21]. Enquanto a primeira derivada remove deslocamentos constantes da linha de base e permite uma melhor visualização dos picos existentes no sinal original ^[22, 23], a segunda derivada elimina uma variação linear da linha de base, associada a efeitos de espalhamento ^[23]. Uma vez que a

aplicação da derivada a dados com baixa relação sinal/ruído tende a aumentar os efeitos do ruído, é necessário que os dados sejam submetidos a uma suavização antes da derivada ^[22].

Uma das técnicas mais utilizadas para suavização de dados é a Savitzky-Golay (SG) ^[24]. Esse procedimento visa aumentar a relação sinal/ruído e reduzir ruídos aleatórios de alta frequência. Na suavização SG, um polinômio de baixa ordem é ajustado aos pontos de uma janela pelos mínimos quadrados. A escolha do número de pontos representa um aspecto importante: um número elevado de pontos pode remover informação e um número reduzido pode fazer com que o ruído permaneça ^[22].

A escolha do pré-processamento adequado é realizada com base na natureza dos dados espectrométricos. É importante salientar que, se aplicado incorretamente, o pré-processamento pode remover informações essenciais ou maximizar a presença de informações irrelevantes ^[20].

2.2 Técnicas de reconhecimento de padrões

As técnicas de reconhecimento de padrões (RP) têm sido amplamente utilizadas em análises qualitativas, envolvendo avaliação de conformidade e identificação de práticas fraudulentas ^[6, 7, 25]. Essas técnicas utilizam a informação química obtida para discriminar as amostras em categorias, considerando suas diferenças e similaridades ^[21]. As técnicas de RP são divididas em duas abordagens: não-supervisionadas e supervisionadas.

As técnicas de RP não-supervisionadas avaliam a existência de agrupamentos naturais utilizando apenas informações provenientes de alguma técnica analítica, sem considerar informação sobre os membros das classes ^[21]. Entre as técnicas de RP não-supervisionadas mais utilizadas no contexto da Química Analítica, destaca-se a análise de componentes principais (PCA: *Principal Component Analysis*) ^[21, 26-28]. Basicamente, a PCA tende a facilitar a representação gráfica do conjunto de dados e, conseqüentemente, facilitar a interpretação dos resultados ^[26].

Os métodos de RP supervisionados, ou apenas métodos de classificação, baseiam-se na modelagem de um conjunto de amostras de categoria conhecida, denominado conjunto de treinamento, usando informação adicional sobre os membros das classes ^[21, 22]. O modelo desenvolvido é posteriormente empregado na identificação de amostras desconhecidas. Diversos métodos de classificação são encontrados na literatura, entre os quais, a análise discriminante por mínimos quadrados parciais (PLS-DA: *Partial Least Squares Discriminant Analysis*) e análise discriminante linear (LDA: *Linear Discriminant Analysis*) vêm se destacando com um elevado número de aplicações ^[29-34].

2.2.1 Análise de componentes principais

A PCA é um dos métodos de RP não-supervisionados mais utilizados para compressão de dados, avaliação de agrupamentos naturais e identificação de amostras anômalas [28]. Neste método, toda variação do conjunto de dados é descrita em um espaço multidimensional reduzido, preservando a maior quantidade de informação possível [26]. O novo espaço é resultado da decomposição da matriz de dados em um sistema formado por eixos ortogonais entre si, denominados fatores ou componentes principais (PCs), obtidos por combinação linear das variáveis originais em ordem decrescente de variância [21, 26, 27]. À medida que a variância decresce, diminui o percentual de informação contido nas PCs e, consequentemente, menos PCs serão necessárias para representar o conjunto de dados, obtendo-se a redução da dimensionalidade [21].

A **Figura 2.1a** apresenta duas PCs obtidas a partir de duas variáveis originais. As novas coordenadas das amostras no novo espaço multidimensional são denominadas escores [27].

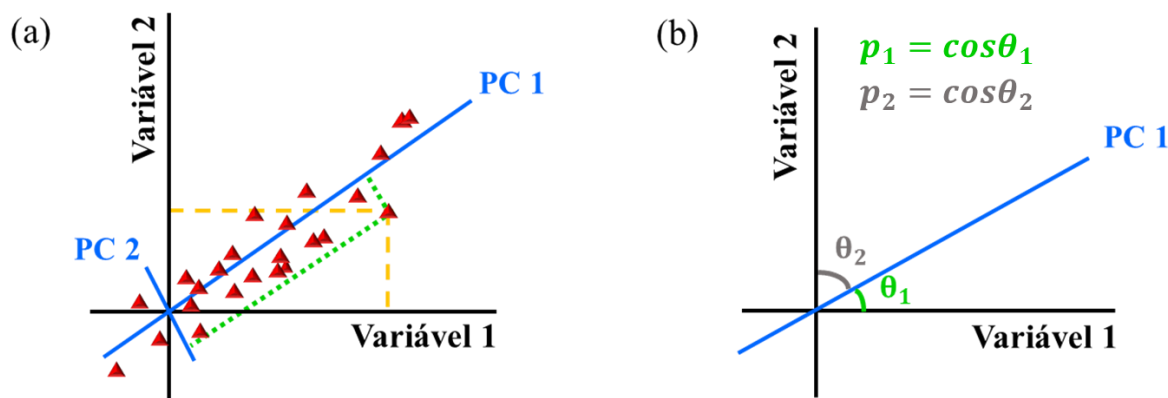


Figura 2.1 - Primeira e segunda PCs obtidas a partir da combinação linear das variáveis originais. (a) As linhas pontilhadas na cor verde correspondem às projeções de uma amostra nas PCs (escores), enquanto as linhas tracejadas na cor amarela representam as projeções nas variáveis originais. (b) O cosseno dos ângulos θ_1 e θ_2 representa os *loadings*.

Fonte: adaptada de Beebe, Pell e Seasholtz (1998).

As PCs são obtidas pela combinação linear das variáveis originais e os coeficientes dessa combinação linear representam uma medida da contribuição de cada variável original para a construção das PCs [27]. Esses valores, conhecidos como *loadings*, estão representado na **Figura 2.1b** e correspondem ao cosseno dos ângulos formados entre a PC e cada variável.

2.2.2 Análise discriminante pelos mínimos quadrados parciais

O modelo PLS-DA baseia-se no método de regressão PLS para construção de modelos de classificação que buscam relacionar as respostas instrumentais com as propriedades de

interesse, utilizando os rótulos das classes como um vetor dependente $\mathbf{y}$ ^[35-37]. Assim como no método PLS, as matrizes $\mathbf{X}$ (matriz de resposta instrumental, $n \times j$) e $\mathbf{Y}$ (matriz das propriedades de interesse, $n \times m$) são decompostas em fatores, ou variáveis latentes, que na quantidade adequada ajudam a minimizar problemas de sobreajuste ^[35].

Problemas de classificação que envolvem apenas duas classes de objetos são solucionados pelo PLS-DA utilizando o algoritmo PLS 1, a partir da codificação do vetor $\mathbf{y}$ com valores de 0 ou 1 para cada classe, como apresentado na **Figura 2.2**. A classificação prossegue obtendo-se valores entre 0 e 1 para cada objeto a ser discriminado, sendo necessário estabelecer um valor que defina a fronteira entre as classes. A fronteira, denominada *threshold*, pode ser calculada pelo teorema de *Bayesian* ou através dos limites de confiança para cada objeto, sendo comumente adotado o valor 0.5 ^[36]. Quando um valor próximo a zero ou abaixo da fronteira é obtido para um objeto, o mesmo é classificado como membro da classe representada pelo valor 0, que neste exemplo corresponde a classe A, caso contrário, pertencerá a classe B.

		Vetor $\mathbf{y}$	Matriz $\mathbf{X}$			
Classe A	▲	0	$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$
	▲	0	$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$
	▲	0	$x_{3,1}$	$x_{3,2}$	...	$x_{3,n}$
Classe B	●	1	$x_{4,1}$	$x_{4,2}$	...	$x_{4,n}$
	●	1	$x_{5,1}$	$x_{5,2}$	...	$x_{5,n}$
	●	1	$x_{6,1}$	$x_{6,2}$	...	$x_{6,n}$

Figura 2.2 - Representação do modelo PLS-DA utilizando o algoritmo PLS 1 para duas classes de objetos.

Quando mais de duas classes de objetos estão envolvidas na classificação, utiliza-se o algoritmo PLS 2 com o número de vetores dependentes $\mathbf{y}$ correspondente ao número de classes envolvidas ^[37].

2.2.3 Análise discriminante linear

Proposta originalmente por Fisher ^[38], a LDA baseia-se no cálculo de funções lineares para discriminar categorias de amostras. As chamadas funções discriminantes lineares são

obtidas por combinação linear das variáveis originais e descrevem os dados na direção de máxima variância entre as classes e mínima variância entre os membros das classes ^[22].

A **Figura 2.3** ilustra a obtenção de uma função discriminante linear e o procedimento de classificação pela LDA quando duas categorias de objetos estão envolvidas. Após a determinação dos centros das classes de objetos a serem discriminados, é obtida uma fronteira chamada função discriminante linear, equidistante e perpendicular à linha que une os centros. Essa função discriminante delimita as regiões correspondentes a cada classe, de modo que, se um objeto qualquer estiver presente na região R1, o mesmo é classificado como membro da classe A. Caso contrário, pertencerá à classe B ^[39]. Neste método, as amostras sempre serão classificadas como membros de uma ou outra classe. A função discriminante será um plano quando estiver relacionada a um espaço tridimensional e um hiperplano m-1 para quatro ou mais dimensões, com m correspondente ao número de dimensões (variáveis) ^[39].

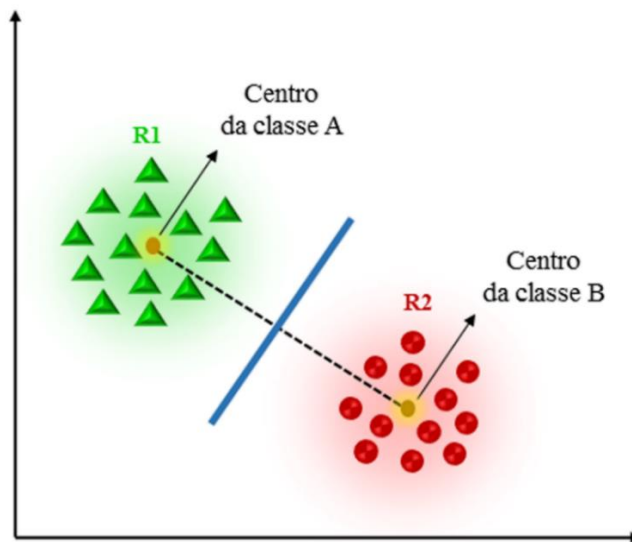


Figura 2.3 - Análise discriminante linear para duas categorias de objetos definidas com classe A (▲) e classe B (●). A linha na cor azul representa a função discriminante linear obtida na representação.
Fonte: adaptada de Bruns e Faigle (1985).

Os métodos de RP supervisionados baseiam-se, essencialmente, no cálculo da distância de um objeto ao centro de uma classe ^[25]. No método LDA, o procedimento de classificação está associado à métrica de Mahalanobis, na qual a correlação entre os conjuntos de dados é considerada ^[40]. Para um estudo envolvendo c classes de objetos, o quadrado da distância de Mahalanobis $r^2(\mathbf{x}, \boldsymbol{\mu}_k)$ entre um objeto $\mathbf{x}$ e o centro da k -ésima classe ($k = 1, 2, \dots, c$) é definido conforme a **Equação 2.1**.

$$r^2(\mathbf{x}, \boldsymbol{\mu}_k) = (\mathbf{x} - \boldsymbol{\mu}_k)^t \cdot \Sigma_k^{-1} \cdot (\mathbf{x} - \boldsymbol{\mu}_k) \quad (2.1)$$

onde $\boldsymbol{\mu}_k (p \times 1)$ e $\boldsymbol{\Sigma}_k (p \times p)$ correspondem ao vetor-média e a matriz de covariância para a classe k , respectivamente. Nos casos em que os valores da média e da covariância são desconhecidos, estimativas desses valores, $\mathbf{m}_k$ e $\mathbf{S}_k$ respectivamente, podem ser obtidas a partir de um conjunto de treinamento com objetos de categoria conhecida. A LDA estima uma única matriz de covariância conjunta $\mathbf{S}$ para todas as classes, o que simplifica o procedimento de classificação e fornece superfícies de decisão lineares no $\Re^p$ [22]. A partir dessas estimativas, o quadrado da distância de Mahalanobis entre o objeto $\mathbf{x}$ e o centro da k -ésima classe é calculado pela **Equação 2.2**.

$$r^2(\mathbf{x}, \mathbf{m}_k) = (\mathbf{x} - \mathbf{m}_k)^t \cdot \mathbf{S}^{-1} \cdot (\mathbf{x} - \mathbf{m}_k) \quad (2.2)$$

Com isso, o objeto $\mathbf{x}$ será considerado membro da classe k mais próxima, para a qual obtiver menor quadrado da distância de Mahalanobis $r^2(\mathbf{x}, \mathbf{m}_k)$. Quando o número de variáveis for maior que o número de classes, $c - 1$ funções discriminantes poderão ser encontradas [41].

É importante destacar que a LDA é bastante afetada por problemas de colinearidade, sendo frequentemente empregada em dados de pequenas dimensões ou que tenham sido submetidos a uma prévia seleção de variáveis [22].

2.3 Seleção de variáveis

Métodos de seleção de variáveis são comumente requisitados por modelos multivariados que não se adequam a matrizes de dados com grandes dimensões ou que apresentam problemas de multicolinearidade. Esses métodos reduzem a dimensionalidade dos dados através da busca por subconjuntos de variáveis capazes de descrever todas as informações relevantes, favorecendo o procedimento de modelagem e o desempenho do modelo [42].

Em problemas de classificação, a seleção de variáveis tenta assegurar que a precisão do modelo não diminua e que a distribuição de classes alcançada com as variáveis selecionadas, seja semelhante a distribuição que seria obtida pelo conjunto completo de variáveis [43].

Diversas estratégias para seleção de variáveis têm sido desenvolvidas, incluindo: Algoritmo Genético (GA: *Genetic Algorithm*) [44], Algoritmo Jack Knife (JK) [45], Stepwise [46], Redes Neurais Artificiais (ANN: *Artificial Neural Networks*) [47] e Algoritmo das Projeções Sucessivas (SPA: *Successive Projections Algorithm*) [48, 49]. A diferença entre essas estratégias está fundamentalmente na forma como os subconjuntos de variáveis são gerados e no critério utilizado para a escolha do melhor subconjunto, que será então acoplado aos métodos multivariados [43].

De modo geral, os métodos de seleção de variáveis seguem uma sequência de passos que consistem, basicamente, na geração de subconjuntos de variáveis, avaliação dos subconjuntos, condição de parada que decide quando finalizar a busca por subconjuntos e validação do subconjunto selecionado ^[43].

2.3.1 Algoritmo das projeções sucessivas

O SPA foi originalmente proposto por Araújo *et al.* (2001) ^[48] no contexto de calibração multivariada, sendo eficientemente utilizado na solução de problemas de multicolinearidade de modelos baseados na regressão linear múltipla (MLR: *Multiple Linear Regression*) ^[21]. Basicamente, o SPA realiza uma sequência de operações de projeções envolvendo as colunas de uma matriz $\mathbf{X}$ ($Kc \times j$) de resposta instrumental, com Kc observações e j variáveis instrumentais, respectivamente, para gerar subconjuntos com número reduzido de variáveis minimamente correlacionadas ^[22]. A **Figura 2.4** ilustra o procedimento de projeções sucessivas das variáveis até a obtenção de uma cadeia com o melhor subconjunto.

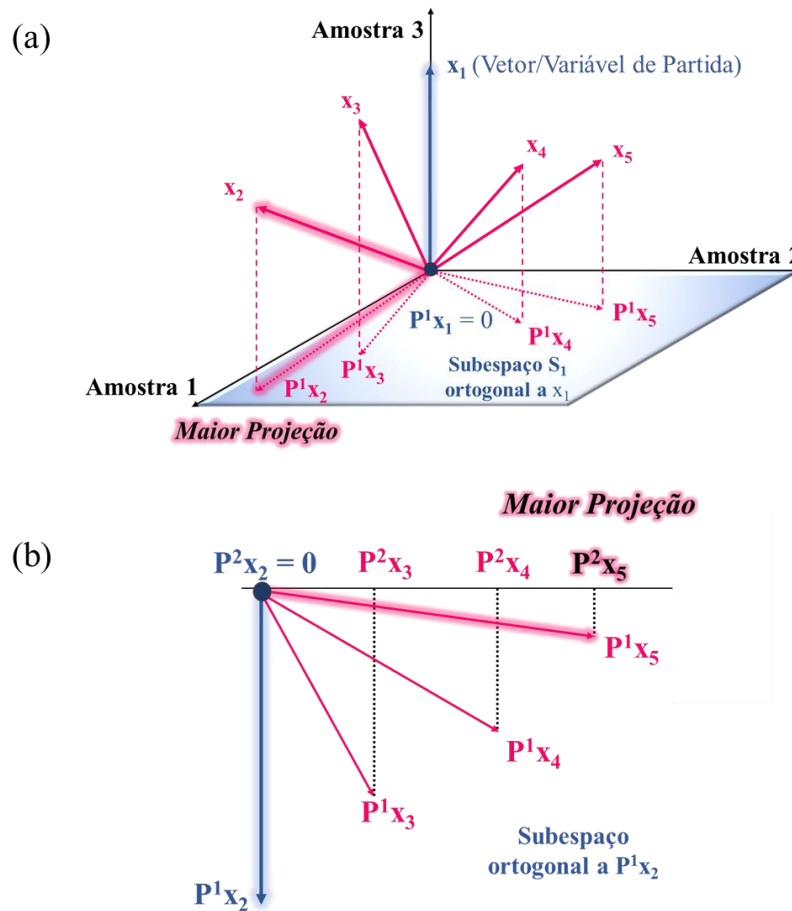


Figura 2.4 - Sequência de projeções realizadas pelo SPA com $Kc = 3$ e $j = 5$. Resultado da primeira (a) e segunda (b) iteração. Neste exemplo, o subconjunto de variáveis selecionadas inicia com x_1 e deverá ser $\{x_1, x_2, x_5\}$.
Fonte: adaptada de Pontes (2009).

A sequência de projeções tem início com um vetor de partida $\mathbf{x}_1$ (vetor que corresponde à variável inicial de seleção) sendo projetado no subespaço S_1 ortogonal a $\mathbf{x}_1$. Os vetores (variáveis) restantes são projetados no subespaço S_1 e o SPA seleciona aquele que apresenta a maior projeção. Este vetor, aqui chamado $\mathbf{x}_2$, é aquele que apresenta menor colinearidade com o vetor de partida $\mathbf{x}_1$, ou seja, é o vetor que apresenta a maior quantidade de informação não contida no vetor de partida. A próxima iteração é restrita ao subespaço ortogonal à projeção do vetor de $\mathbf{x}_2$, aqui denominada $\mathbf{P}^1\mathbf{x}_2$, tendo agora essa projeção como referência. A cada iteração, uma nova variável é selecionada para compor o subconjunto com o número de variáveis previamente estabelecido pelo analista. As projeções seguem sucessivamente com os passos descritos até que uma cadeia de variáveis seja obtida. Além de minimizar a multicolinearidade entre as variáveis, o SPA permite que no máximo K_c variáveis sejam selecionadas, visto que a cada operação de projeção, um grau de liberdade é removido da matriz $\mathbf{X}$ ^[23].

A escolha do melhor subconjunto é feita com base na raiz quadrada do erro médio quadrático para o conjunto de validação (RMSEV: *Root Mean Square Error of Validation*), conforme descrito na **Equação 2.3**.

$$\text{RMSEV} = \sqrt{\frac{1}{K_{\text{val}}} \sum_{k=1}^{K_{\text{val}}} (y_V^k - \hat{y}_V^k)^2} \quad (2.3)$$

onde y_V^k e $\hat{y}_V^k$ correspondem, respectivamente, ao valor de referência e valor previsto para o parâmetro de interesse na k -ésima amostra de validação e K_{val} é o número de amostras de validação ^[50].

O SPA foi adaptado por Pontes *et al.* (2005) ^[49] para seleção de variáveis no contexto de classificação utilizando análise discriminante linear (SPA-LDA). A adaptação consiste na utilização de uma função de custo como critério para seleção do melhor subconjunto de variáveis ^[22]. No SPA-LDA, assume-se que cada objeto pertence a uma classe conhecida com índice Ik ($k = 1, 2, \dots, c$). Dessa forma, a média de cada classe $\mathbf{m}_{Ik}$ e uma matriz de covariância agrupada $\mathbf{S}$ são calculadas para cada subconjunto candidato de variáveis usando o conjunto de amostras de treinamento, como descrito no procedimento de classificação pelo LDA na seção 2.2.3. No SPA-LDA, o critério de avaliação do melhor subconjunto também se baseia no conjunto de validação, agora em termos da função de custo G_{val} definida como

$$G_{\text{val}} = \frac{1}{K_{\text{val}}} \sum_{k=1}^{K_{\text{val}}} g_{\text{val},k} \quad (2.4)$$

Na **Equação 2.4**, $g_{val,k}$, corresponde ao risco de uma classificação incorreta da k -ésima amostra de validação, $\mathbf{x}_{val,k}$, sendo calculado como

$$g_{val,k} = \frac{r^2(\mathbf{x}_{val,k}, \boldsymbol{\mu}_{Ik})}{\min_{Ij \neq Ik} r^2(\mathbf{x}_{val,k}, \boldsymbol{\mu}_{Ij})} \quad (2.5)$$

onde o numerador $r^2(\mathbf{x}_k, \boldsymbol{\mu}_{Ik})$ correspondente ao quadrado da distância de Mahalanobis entre o objeto $\mathbf{x}_k$ (com índice de classe Ik) e a média de sua classe ($\boldsymbol{\mu}_{Ik}$). O denominador da equação é o quadrado da distância de Mahalanobis entre o objeto $\mathbf{x}_k$ e o centro da classe incorreta mais próxima. Com isso, o menor risco será obtido quanto mais distante o objeto estiver do centro da classe incorreta e mais próximo estiver do centro de sua classe verdadeira ^[22].

Uma alternativa foi proposta por Soares *et al.* (2014) ^[51] para problemas de classificação que envolvem um número reduzido de amostras. Nesses casos, o conjunto de validação foi dispensado e o risco de uma classificação incorreta é calculado usando os mesmos objetos de treinamento que foram empregados para calcular as médias de cada classe $\mathbf{m}_{Ik}$ e a matriz de covariância $\mathbf{S}$. A função de custo G_{train} correspondente a essa adaptação é apresentada na **Equação 2.6**.

$$G_{train} = \frac{1}{K_{train} - L - C} \sum_{k=1}^{K_{train}} g_{train,k} \quad (2.6)$$

Na **Equação 2.6**, K_{train} é o número de amostras de treinamento, L é o número de variáveis no subconjunto que está sendo avaliado e C representa o número de classes investigadas. Essa expressão evita possíveis problemas de sobreajuste relacionados ao uso repetido do conjunto de treinamento ^[51].

2.4 Transferência de modelos multivariados

Modelos multivariados desenvolvidos em um determinado instrumento (primário), podem não ser corretamente aplicados a amostras analisadas sob novas condições experimentais ou em um instrumento não envolvido na etapa de modelagem (secundário). A inaplicabilidade do modelo é resultado de diferenças entre as respostas instrumentais, normalmente relacionadas a alterações na composição química e/ou física das amostras, mudanças nas condições de análise, alterações no ambiente do instrumento como variações de temperatura e umidade que podem provocar deslocamento nas bandas de absorção e mudanças

não lineares nas suas intensidades e na sua função de resposta devido à instabilidade do detector, irregularidades de manutenção, reparos ou desgaste natural ^[13, 18, 19].

Uma das soluções para esse problema é a recalibração completa do modelo a partir de um conjunto de amostras analisado sob as novas condições analíticas ^[13, 15]. Na prática, esse procedimento é dispendioso, exige tempo e rigoroso controle das condições experimentais ^[13-15]. Alternativamente, a transferência de classificação pode ser realizada usando métodos de padronização das respostas instrumentais ou aumento da robustez dos modelos.

2.4.1 Métodos de padronização das respostas instrumentais

Nos métodos de padronização, a função de resposta de um instrumento secundário é modificada para corresponder à função de resposta de um instrumento primário, para o qual o modelo foi desenvolvido. Uma série de manipulações matemáticas podem ser realizadas para padronização das respostas instrumentais ^[14].

Em todos os casos, o procedimento de transferência é realizado a partir de um subconjunto de amostras que tenham sido analisadas em todos os instrumentos envolvidos. Essas amostras, conhecidas como amostras de transferência, são selecionadas a partir do subconjunto utilizado para treinar o modelo e devem ser suficientemente representativas para descrever as diferenças instrumentais ^[13]. Considerando-se as matrizes $\mathbf{X}^P$ ($n_{trans} \times j$) e $\mathbf{X}^S$ ($n_{trans} \times j$) contendo os sinais das n_{trans} amostras de transferência, registrados nos equipamentos primário e secundário, respectivamente, a padronização tenta relaciona-las linearmente conforme descrito na **Equação 2.7**:

$$\mathbf{X}^P = \mathbf{X}^S \cdot \mathbf{F} \quad (2.7)$$

onde $\mathbf{F}$ ($j \times j$) é a matriz de transformação, que pode ser estimada com base nas amostras de transferência, n_{trans} . Com a matriz de transformação $\mathbf{F}$, o sinal de uma amostra medida no equipamento secundário pode ser ajustado para $\hat{\mathbf{x}}^a$, de modo a assemelhar-se aos sinais obtidos pelo equipamento primário ^[23].

$$\hat{\mathbf{x}}^a = \mathbf{x}^S \cdot \mathbf{F} \quad (2.8)$$

Dependendo de sua forma, a matriz de transformação $\mathbf{F}$ pode ser usada para corrigir diferenças tanto na magnitude dos sinais como no deslocamento dos comprimentos de onda ^[13]. Os métodos de padronização incluem padronização direta (DS: *Direct Standardization*) e padronização direta em etapas (PDS: *Piecewise Direct Standardization*).

2.4.1.1 Padronização direta

No método proposto por Wang *et al.* (1991) ^[16], as matrizes das respostas instrumentais $\mathbf{X}^P$ ($n_{trans} \times j$) e $\mathbf{X}^S$ ($n_{trans} \times j$) estão relacionadas por meio da matriz de transformação $\mathbf{F}$, conforme descrito na **Equação 2.7**. Neste método, cada coluna da matriz de resposta instrumental obtida no instrumento primário é relacionada, simultaneamente, a todas as colunas da matriz de resposta do instrumento secundário ^[13, 14]. Se os dois equipamentos forem iguais, a matriz de transformação $\mathbf{F}$ é quadrada e suas colunas descrevem, para as respectivas variáveis, as correlações entre os sinais obtidos em cada instrumento ^[23]. Caso $\mathbf{X}^S$ seja invertível, a matriz $\mathbf{F}$ poderá ser calculada por:

$$\mathbf{F} = (\mathbf{X}^S)^{-1} \cdot \mathbf{X}^P \quad (2.9)$$

onde $(\mathbf{X}^S)^{-1}$ é a inversa generalizada de $\mathbf{X}^S$. Uma vez que o número de variáveis é normalmente maior do que o número de amostras de transferência e $\mathbf{X}^S$ é uma matriz retangular, $\mathbf{F}$ pode ser calculado por métodos multivariados como regressão por componentes principais (PCR: *Principal Component Regression*) ^[21] ou PLS. Nestes métodos, as matrizes $\mathbf{X}^P$ e $\mathbf{X}^S$ são projetadas no espaço de coluna de $\mathbf{X}^P$, calculado a partir da decomposição em valores singulares. Os escores obtidos nessas projeções são utilizados nos procedimentos de padronização e calibração ^[13, 14]. Como apresentado na seção anterior, a padronização do sinal da nova amostra medida no instrumento secundário é obtida a partir da matriz $\mathbf{F}$ da seguinte forma:

$$\mathbf{x}^a = \mathbf{x}^S \mathbf{F} \quad (2.10)$$

onde o vetor resposta de uma amostra desconhecida, medida no equipamento secundário, $\mathbf{x}^S$, é padronizado para o vetor resposta esperado no equipamento primário, $\mathbf{x}^a$.

Para garantir que $\mathbf{F}$ represente unicamente variações instrumentais e não alterações na composição das amostras, este método requer que as amostras sejam analisadas nos dois equipamentos exatamente nas mesmas condições. Além disso, o método DS é comumente afetado por sobreajuste na estimativa de $\mathbf{F}$, uma vez que o número de amostras de transferência é muito menor que o número de constantes utilizadas no ajuste do espectro inteiro ^[13].

2.4.1.2 Padronização direta em etapas

Também proposta por Wang *et al.* (1991) ^[16], a padronização PDS é um dos métodos de transferência mais utilizados. Este método considera que as correlações entre os sinais

obtidos nos instrumentos primário e secundário estão limitadas a pequenas regiões, de modo que, utilizando-se uma janela móvel, cada variável j da matriz $\mathbf{X}^P$ ($n_{trans} \times j$) é relacionada a uma pequena região em torno de j na matriz $\mathbf{X}^S$ ($n_{trans} \times j$) [13, 14, 16]. A utilização de uma janela móvel reduz o risco de sobreajuste, uma vez que o número de constantes a serem determinados para estimar $\mathbf{F}$ é menor [23].

No método PDS, as matrizes das respostas instrumentais $\mathbf{X}^P$ e $\mathbf{X}^S$ também se relacionam como na **Equação 2.7**, mas agora, para cada janela é calculado um vetor de regressão $\mathbf{f}$, utilizando PCR e PLS. Os vetores de regressão $\mathbf{f}$ são dispostos em uma matriz diagonal $\mathbf{F}$ de acordo com

$$\mathbf{F} = \text{diag}(\mathbf{f}_1^T, \mathbf{f}_2^T, \dots, \mathbf{f}_j^T, \dots, \mathbf{f}_k^T) \quad (2.11)$$

onde k é o número de variáveis. Com isso, novas amostras podem ser padronizadas pela **Equação 2.10**. O uso de janelas móveis pode acarretar efeitos de borda onde não há dados suficientes para formar uma janela completa, o que pode ser corrigido estimando-se o final do espectro por extrapolação ou, preferencialmente, o descartando [13, 14].

A janela pode ser simétrica ou assimétrica em torno de cada variável. Em problemas envolvendo deslocamentos significativos nos comprimentos de onda, ou desvio espectral, a janela assimétrica melhora o procedimento de transferência, enquanto a janela simétrica é eficiente nos casos em que os deslocamentos de comprimento de onda são desprezíveis [13, 14].

A PDS é o método de transferência normalmente empregado como referência para avaliação de novas técnicas, sobretudo pela sua capacidade de corrigir simultaneamente diferenças de intensidade, deslocamentos de comprimentos de onda e alteração na largura das bandas [13, 14]. Desde a criação do algoritmo original, algumas modificações já foram propostas com o intuito de corrigir diferenças aditivas entre os espectros [52] e variações contínuas de temperatura [53].

Um dos principais problemas apresentados pelo método PDS é a estimativa dos postos locais dos modelos PCR e PLS [13, 18, 54], uma vez que no PDS, o número ideal de componentes principais é estimado com base em janelas com quantidade reduzida de variáveis. O problema na estimativa dos postos locais é atribuído a ruídos de alta frequência nos dados, que podem ser amenizados com um pré-processamento adequado antes da transferência [13, 14]. Um procedimento para detectar e reduzir problemas desse tipo foi proposto por Bouveresse e Massart (1996) [54], empregando-se uma amostra independente do conjunto de dados para exame do espectro antes e depois da transferência [13]. Nesse procedimento, a presença de

problemas nos postos locais é indicada por altos valores residuais entre o espectro original e o transferido [13, 55].

Vários estudos têm sido relatados na literatura usando a transferência de modelos de calibração via padronização das respostas instrumentais [56-62]. Contudo, as técnicas de transferência de modelos multivariados têm sido pouco aplicadas para fins de classificação [63-65].

A PDS foi utilizada por Di Anibal *et al.* (2012) [63] para discriminar diferentes especiarias culinárias adulteradas com corante Sudão dos tipos I, II, III e IV. As amostras foram caracterizadas por espectroscopia UV-Vis e modelos baseados na análise discriminante por mínimos quadrados parciais (PLS-DA: *Partial Least Squares Discriminant Analysis*) [66] foram utilizados para discriminá-las em amostras não-adulteradas e adulteradas com algum dos quatro corantes. O método PDS foi testado usando diferentes janelas e quantidades de amostras de transferência e, após a padronização, resultados positivos foram encontrados pelos modelos PLS-DA na previsão de amostras medidas sob diferentes condições experimentais.

Bergo [64] realizou a discriminação de diferentes espécies de madeira utilizando espectroscopia NIR. Os espectros das amostras foram registrados em dois espectrômetros diferentes (um de bancada e um portátil) e modelos PLS-DA foram desenvolvidos com 100% de classificação correta para cada equipamento. Em seguida, foram testados métodos reversos de padronização que forneceram taxas de classificação correta de 93% a 100% para discriminação de espécies de madeira.

Silva *et al.* (2015) [65] utilizaram os métodos de modelagem independente e flexível por analogia de classe (SIMCA: *Soft Independent Modeling of Class Analogy*) [67] e PLS-DA para a classificação de amostras de gasolina de diferentes origens que são contrabandeadas nas fronteiras brasileiras. Espectros MIR registrados em três diferentes instrumentos foram transferidos empregando-se métodos de padronização direta. O procedimento de transferência minimizou a dispersão entre as amostras de gasolina de origem brasileira e forneceu melhor discriminação entre as gasolinas de diferentes origens.

2.4.2 Aumento da robustez do modelo

A robustez é um dos parâmetros mais utilizadas para avaliar o desempenho de modelos multivariados, sendo comumente associada à capacidade de um modelo de fornecer resultados confiáveis diante de variações não incluídas na modelagem [17]. Além de ser fortemente influenciada pela quantidade de amostras analisadas, instrumentação utilizada e condições

experimentais, a robustez também pode ser alcançada com a incorporação de todas as fontes de variação importantes ainda na etapa de modelagem (modelos globais), pré-processamento adequado dos dados e estratégias de aumento de robustez [12, 18, 14, 68, 69]. Esses procedimentos podem aumentar a robustez de um modelo quando as diferenças entre as funções de resposta são relativamente pequenas, favorecendo a capacidade preditiva e eliminando a necessidade dos procedimentos de padronização [18, 19].

Nos modelos globais, todas as fontes de variação esperadas são intencionalmente calibradas considerando-se uma larga faixa de condições experimentais, incluindo preparo de amostras, variações instrumentais e fatores ambientais, a fim de incluir variabilidade ao modelo [14, 15]. Para que um modelo global seja obtido, é necessário que o analista conheça profundamente o sistema e todas as possíveis fontes de variação. Na prática, isso se torna bastante complicado e tende a fornecer modelos menos precisos, com grandes dimensões e maior complexidade. Caso o analista consiga identificar possíveis fontes de variações futuras, os modelos globais produzidos apresentarão maior robustez em uma ampla gama de situações [13, 14].

As técnicas de pré-processamento são utilizadas para remover variabilidade indesejável no espectro [14, 68]. Além das técnicas mencionadas na seção 2.1, a correção de sinal ortogonal (OSC: *Orthogonal Signal Correction*) [70], a transformada *Wavelet* (TW) [71] e a transferência por projeções ortogonais (TOP: *Transfer by Orthogonal Projections*) [72] são bastante utilizadas para fins de transferência, conforme descrito por Honorato *et al.* (2007) [14].

A seleção das variáveis ideais para o desenvolvimento de um modelo influencia a sua robustez e, conseqüentemente, sua capacidade preditiva. Em problemas de transferência, é fundamental que as variáveis selecionadas apresentem baixa sensibilidade em relação às mudanças nas condições experimentais ou variações instrumentais [18].

A transferência de modelos multivariados pelo aumento da robustez tem sido pouco abordada na literatura [68, 73, 74].

Swierenga *et al.* (1998) [68] investigaram a transferência do modelo baseado nos mínimos quadrados parciais (PLS: *Partial Least-Squares*) [21] utilizando a padronização DS e diferentes técnicas de pré-processamentos, entre as quais, a seleção de variáveis. O modelo foi utilizado para determinar a porcentagem de álcool em misturas alcoólicas contendo metanol, etanol e 1-propanol, analisadas por espectroscopia NIR a partir de três instrumentos diferentes. Os resultados obtidos utilizando os pré-processamentos foram comparáveis aos obtidos pela padronização DS.

Myles *et al.* (2006) ^[73] investigaram a transferência de modelos de classificação multivariados desenvolvidos para discriminação de grãos de café dos tipos Arabica e Robusta analisados por espectroscopia NIR. Para isso, foram avaliadas estratégias de atualização de modelos ^[13], correção de sinal ortogonal (OSC: *Orthogonal Signal Correction*) ^[70] treinada com um vetor discreto de índice de classe e uma versão modificada da correção de slope/bias (SBC: *slope/bias correction*) ^[75]. Com isso, modelos mais robustos foram alcançados com as estratégias de transferência investigadas, sendo obtidos erros de classificação na faixa de 5 – 10%.

Swierenga *et al.* (1998) ^[74] avaliaram a transferência do modelo PLS a partir da seleção de variáveis por recozimento simulado (SA: *Simulated Annealing*) ^[76] para determinar o teor de água em comprimidos. As amostras foram analisadas em dois diferentes instrumentos por espectroscopia NIR e, em comparação a padronização DS, o modelo PLS com seleção de variáveis pelo SA apresentou melhores resultados.

2.4.3 Seleção das amostras de transferência

O procedimento de transferência de um modelo é eficiente quando as amostras de transferência são devidamente selecionadas. Para que as diferenças entre as respostas instrumentais sejam adequadamente contempladas, as amostras de transferência devem ser física e quimicamente estáveis e suficientemente representativas ^[19], caso contrário, os parâmetros de transferência calculados poderão resultar em padronizações não confiáveis ^[23].

O algoritmo originalmente utilizado para seleção das amostras de transferência baseia-se na busca pelas amostras com maior *leverage* ^[13]. Partindo-se da amostra com maior *leverage*, a mesma é removida do conjunto de dados e ortogonalizada contra as demais. Esse procedimento é repetido até que o número de amostras estabelecido pelo analista seja atingido. Como desvantagem, esse método não cobre todo espaço experimental, o que reduz sua representatividade e prejudica a transferência ^[13].

Como alternativa, o algoritmo clássico Kennard-Stone ^[77] tem sido utilizado, visto que permite selecionar amostras de todas as partes do espaço de dados para distribuições não homogêneas ^[13, 14].

Dantas Filho *et al.* (2004) ^[78] propuseram uma adaptação do algoritmo SPA, originalmente utilizado para seleção de variáveis, para ser aplicado à seleção de amostras de calibração, também podendo ser empregado para selecionar amostras de transferência ^[12]. No SPA para seleção de amostras, a matriz da resposta instrumental é transposta (as amostras de

calibração tornam-se colunas) e o SPA seleciona um subconjunto de amostras representativas que exibe mínima multicolinearidade, preservando a informação relevante para a modelagem.

O número ideal de amostras de transferência também representa um aspecto importante. Geralmente, esse número é determinado após a avaliação dos resultados de validação de modo que a quantidade ideal de amostras é aquela que minimiza os erros de previsão ^[13]. Dois aspectos principais devem ser considerados na escolha do número ideal de amostras: a magnitude das diferenças instrumentais e o tipo de padronização utilizado ^[19].

2.4.3.1 Seleção de amostras pelo algoritmo KS

Em quimiometria, a forma com que as amostras são divididas em subconjuntos de treinamento, validação e teste pode comprometer o desempenho dos modelos multivariados, sendo comumente utilizado para este fim, o algoritmo clássico Kennard-Stone (KS) ^[77], sobretudo quando métodos de classificação são empregados.

O algoritmo KS emprega distâncias Euclidianas para selecionar um subconjunto representativo de N amostras a partir de uma dada população. A fim de garantir uma distribuição uniforme para cada subconjunto ao longo do espaço $\mathbf{x}$ de resposta instrumental, o algoritmo KS calcula as distâncias Euclidianas $d_{\mathbf{x}}(p, q)$ entre os vetores $\mathbf{x}$ de cada par de amostras (p, q) , como mostra a **Equação 2.12** ^[79].

$$d_{\mathbf{x}}(p, q) = \sqrt{\sum_{j=1}^J [\mathbf{x}_p(j) - \mathbf{x}_q(j)]^2} ; \quad p, q \in [1, N] \quad (2.12)$$

Para dados de origem espectral, $\mathbf{x}_p(j)$ e $\mathbf{x}_q(j)$ correspondem as respostas instrumentais do j -ésimo comprimento de onda para as amostras p e q , respectivamente, e J representa o número de comprimentos de onda no espectro.

A seleção de amostras pelo KS tem início com o par de amostras que apresenta a maior distância $d_{\mathbf{x}}(p_1, p_2)$ entre si. A cada nova iteração, o algoritmo seleciona amostras que apresentam a maior distância em relação a outra já selecionadas. O procedimento é repetido até que o número de amostras estabelecido pelo analista seja alcançado.

2.5 Parâmetros de desempenho

A validação é uma etapa muito importante em qualquer procedimento de modelagem. Nas análises qualitativas, a avaliação de desempenho dos modelos visa assegurar que as triagens

realizadas no decorrer do processo produtivo ou fora dele, sejam conduzidas de forma rápida, simples, objetiva, com baixo custo e menor risco de erros ^[17]. Alguns parâmetros são bastante utilizados para validação de métodos qualitativos, entre os quais, destacam-se sensibilidade, especificidade e acurácia ^[17, 80], normalmente calculados a partir de uma tabela de contingência 2×2 , construída como apresentado na **Tabela 2.1**.

Tabela 2.1 - Tabela de contingência 2×2 .

Resultados	Adulterante ausente	Adulterante presente	Total
Negativo (amostra não - adulterada)	TN (NA como NA)	FP (NA como A)	TN + FP
Positivo (amostra adulterada)	FN (A como NA)	TP (A como A)	FN + TP
Total	TN + FN	FP + TP	N

FP = quantidade de resultados falso-positivos; FN = quantidade de resultados falso-negativos; TP = quantidade de resultados positivos corretos; TN = quantidade de resultados negativos corretos; N = TP+FP+FN+TN; A = amostras adulteradas; NA = amostras não-adulteradas.

A **Tabela 2.1** exemplifica um problema de classificação que envolve adulteração. Neste caso, tem-se duas categorias de amostras (adulterada e não-adulterada) e duas possíveis situações (adulterante presente e adulterante ausente), que poderiam resultar em quatro tipos de classificação de uma amostra: ser considerada adulterada quando de fato é adulterada (TP), ser considerada adulterada quando o adulterante não está presente (FP), ser considerada não-adulterada sendo não-adulterada (TN) e ser considerada não-adulterada quando houver adulteração (FN).

Enquanto a sensibilidade (**Equação 2.13**) é utilizada para medir a habilidade de um modelo em detectar corretamente as amostras pertence à condição positiva, a especificidade (**Equação 2.14**) avalia sua habilidade em detectar corretamente as amostras pertence à condição negativa ^[81].

$$\text{Sensibilidade} = \frac{TP}{TP+FN} \quad (2.13)$$

$$\text{Especificidade} = \frac{TN}{TN+FP} \quad (2.14)$$

A acurácia, também conhecida como taxa de classificação correta (TCC) é determinada pela **Equação 2.15**.

$$\text{Acurácia (\%)} = \frac{TP+TN}{N} \times 100 \quad (2.15)$$

2.6 Estudos de caso

2.6.1 Etanol Hidratado Combustível (EHC)

O etanol hidratado combustível (EHC) é utilizado em veículos automotores como alternativa aos combustíveis derivados do petróleo, sobretudo pela considerável redução dos impactos ambientais decorrentes da sua utilização. Embora seu poder de combustão seja cerca de 68% menor que o dos combustíveis fósseis, o etanol é um combustível limpo e, consequentemente, emite menor quantidade de substâncias tóxicas ^[82].

O etanol combustível pode ser obtido com diferentes matérias-primas a partir da fermentação dos açúcares encontrados em alguns vegetais (ex.: cana-de-açúcar e beterraba) ou polissacarídeos como amido e celulose, que podem ser convertidos em açúcares ^[82, 83]. Alternativamente, o álcool pode ser obtido a partir da biomassa lignocelulósica proveniente de resíduos de produtos naturais como o sabugo e a palha do milho, o bagaço e a palha da cana-de-açúcar ^[83]. Nestes casos, o produto é chamado de etanol de segunda geração, 2G, sendo comumente associado à sustentabilidade.

No Brasil, a cana-de-açúcar é a principal matéria-prima usada na fabricação de etanol. Sua elevada utilização é atribuída aos baixos custos no processo produtivo, alto rendimento em termos de volume percentual de etanol obtido por área plantada e às vantagens em termos de balanço energético de produção quando comparado à produção de álcool combustível envolvendo outras matérias-primas ^[84].

A fiscalização e regulamentação da qualidade dos combustíveis comercializados no território brasileiro é realizada pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). De acordo com a resolução ANP nº 19/2015 ^[85], como combustível, o etanol pode ser utilizado na forma anidra (etanol anidro combustível – EAC), misturado à gasolina em quantidade regulamentada ou na forma hidratada (EHC), para a qual, o teor alcoólico deve estar na faixa entre 92,5% a 94,6% (m/m) e 95,5% a 96,5% (m/m) para o produto *premium* (EHCP) ^[85].

O EHC tem sido alvo frequente de adulteração envolvendo adição de água e metanol acima dos valores permitidos, substituição do etanol pela mistura de etanol anidro com água, ou em casos mais graves, substituição do etanol pelo metanol ^[84, 86]. Além dos prejuízos econômicos, a adulteração com metanol pode provocar sérios problemas de saúde decorrentes da sua alta toxicidade ^[87].

De modo geral, os métodos recomendados pela ANP para determinação dos teores de água ^[88-90] e metanol ^[91, 92] são laboriosos, demorados e envolvem reagentes e instrumentos

caros. Muitos esforços têm sido voltados ao desenvolvimento de novas metodologias analíticas capazes de verificar de forma rápida e segura a autenticidade de amostras de EHC. Nesse sentido a espectroscopia na região do infravermelho próximo (NIR), associada aos métodos de análise multivariada, tem se mostrado uma alternativa rápida, segura, não destrutiva e que não necessita de pré-tratamento das amostras ^[93-95].

Um dos grandes obstáculos à utilização da espectroscopia NIR associadas aos métodos multivariados de análise é a inaplicabilidade de um modelo às amostras analisadas sob novas condições experimentais ou em um instrumento não envolvido na etapa de modelagem. Este problema torna-se ainda mais grave quando envolve controle de qualidade de combustíveis no Brasil, que em virtude de sua grande extensão territorial, tem a fiscalização dificultada. Nesse contexto, é de fundamental importância assegurar que os modelos desenvolvidos em diferentes condições possam ser transferidos e aplicados à novas amostras.

2.6.2 *Azeite de oliva extra virgem (AOEV)*

O azeite de oliva do tipo extra virgem (AOEV) é um dos principais óleos extraídos do fruto da oliveira e, devido às suas estimadas propriedades nutricionais e características organolépticas, seu consumo tem sido amplamente difundido ao longo dos séculos ^[96].

Suas propriedades nutricionais são atribuídas principalmente aos altos teores de alguns ácidos graxos insaturados, que atuam positivamente sobre os níveis de colesterol e auxiliam na proteção contra doenças coronárias e inflamatórias, compostos fenólicos, conhecidos pela ação antioxidante, e tocoferóis, que promovem sua estabilidade oxidativa ^[96, 97]. Além disso, suas características organolépticas também são muito apreciadas, com destaque para o típico odor proporcionado pela presença de compostos voláteis ^[98].

A **Tabela 2.2** apresenta os valores estabelecidos para alguns critérios de qualidade dos AOEVs em comparação com os demais azeites. Esses valores são estabelecidos pelo Conselho Oleícola Internacional (IOOC: *International Olive Council*) ^[99], organização intergovernamental que regulamenta o comércio internacional desses produtos, e pelo Comitê Codex de Óleos e Gorduras (CCFO: *Codex Committee on Fats and Oils*) ^[100], que estabelece normas internacionais de qualidade para as diferentes categorias de azeites de oliva. No Brasil, a ANVISA ^[101] e o MAPA ^[102] são responsáveis pela regulamentação e fiscalização dos azeites de oliva comercializados.

Tabela 2.2 - Principais critérios de qualidade para cada tipo de azeite de oliva.

Critério de qualidade	Azeite de oliva extra virgem	Azeite de oliva virgem	Azeite de oliva	Azeite de oliva refinado
<u>Acidez livre</u>	≤ 0.8	≤ 2.0	≤ 1.0	≤ 0.3
<u>Valor de peróxido</u>	≤ 20	≤ 20	≤ 15	≤ 5
<u>Absorbância no ultravioleta ($K_{1\text{cm}}^{1\%}$)</u>				
270 nm*	≤ 0.22	≤ 0.25	≤ 0.90	≤ 1.10
ΔK	≤ 0.01	≤ 0.01	≤ 0.15	≤ 0.16
232 nm	≤ 2.50	≤ 2.60		

*A absorbância em 270 nm é verificada quando o solvente utilizado é o ciclohexano. Quando o isoctano é utilizado, verifica-se a absorbância em 268 nm. ΔK representa a variação das absorbâncias no ultravioleta calculada para cada comprimento de onda avaliado.

Fonte: International Olive Council (2015).

Alguns aspectos naturais exercem influência direta na composição química e qualidade dos AOEVs, como a variedade da azeitona, o estágio de maturação e fatores climáticos e geográficos^[103]. Além disso, os métodos empregados em sua produção, desde o cultivo do fruto até a obtenção do óleo, também são fundamentais.

Os AOEVs são obtidos a partir do fruto da árvore oliveira (*Olea europaea* L.) apenas por meios físicos e processos mecânicos (em condições térmicas adequadas), dispensando a utilização de solventes^[99]. A produção envolve complexas e duradouras etapas que resultam em um baixo rendimento do produto final, que, somado aos custos de produção e elevado valor nutricional, fazem com que o produto apresente alto valor de mercado^[104], tornando sua adulteração uma prática bastante atrativa. As adulterações relatadas até agora envolvem azeites e óleos vegetais de qualidade e custo inferiores, como óleo de soja, milho, amendoim, girassol, açafrão, canola e palma^[105].

Diversos trabalhos têm sido reportados na literatura com a finalidade de desenvolver novas metodologias analíticas que, em alternativa aos métodos clássicos cromatográficos de análise^[106, 107], possam caracterizar e/ou autenticar AOEVs de forma eficiente e segura, incluindo: voltametria^[108], espectroscopia de fluorescência^[109, 110], espectroscopia na região do infravermelho^[96, 111, 112], calorimetria de varrimento diferencial^[113], ressonância magnética nuclear^[114], entre outros^[115, 116].

Diante do exposto, é de suma importância assegurar que os modelos de classificação desenvolvidos para identificar adulteração em AOEV possam ser transferidos.

2.6.3 Técnicas analíticas

2.6.3.1 Espectroscopia no infravermelho

A região espectral no infravermelho (IV) compreende a faixa de radiação com números de onda no intervalo de aproximadamente 12.800 a 10 cm^{-1} [117]. Seu espectro é dividido em infravermelho próximo (NIR), infravermelho médio (MIR) e infravermelho distante (FIR), cujos limites aproximados são mostrados na **Tabela 2.3**.

Tabela 2.3 - Regiões espectrais do infravermelho.

Região	Intervalo de comprimento de onda (nm)	Região do número de onda (cm^{-1})
NIR	780 – 2.500	12.800 – 4.000
MIR	2.500 – 5.000	4.000 – 200
FIR	5.000 – 100.000	200 - 10

Fonte: adaptada de Skoog *et al.* (2006).

A espectroscopia no IV, assim como as demais formas de espectroscopia, baseia-se nos efeitos da interação entre matéria e radiação eletromagnética [118]. Ao absorver energia proveniente da radiação, os átomos e grupos funcionais presentes na matéria podem vibrar com amplitude apreciável ao redor de suas ligações covalentes [119]. Uma vez que a radiação no IV não é suficientemente energética, como acontece com as radiações ultravioleta, visível e raios X, ela não promove transições eletrônicas e sua absorção está restrita a espécies moleculares que têm diferenças de energia pequena entre vários estados vibracionais e rotacionais [117].

Uma molécula só absorve radiação IV se a mesma sofrer uma variação no momento de dipolo, como consequência do movimento vibracional ou rotacional. Nesses casos, o campo elétrico alternado da radiação é capaz de interagir com a molécula e causar variações na amplitude de um de seus movimentos [117].

A radiação na região do infravermelho próximo (NIR – *Near Infrared*) tem sido bastante utilizada como alternativa aos métodos analíticos tradicionais para análises dos padrões de qualidade de uma variedade de materiais, por fornecer uma grande quantidade de informações de forma rápida, simples, com baixo custo e sem requisitar pré-tratamento das amostras [120, 22].

As ocorrências espectrais nesta região decorrem de sobretons de transições vibracionais e combinações harmônicas das vibrações fundamentais de ligações C-H, O-H, N-H e S-H provocados pela grande anarmonicidade dessas vibrações [120, 23]. Acoplamentos ou ressonância entre diferentes vibrações do mesmo grupo funcional também podem ocorrer [23].

Comparado ao espectro na região MIR, o espectro NIR apresenta bandas de absorção com menor intensidade (cerca de 10 a 1000 vezes) e maior sobreposição, o que pode ser resolvido por meio de instrumentação adequada (que forneça o aumento da relação sinal/ruído) e utilização de métodos multivariados de análise, respectivamente [20, 22].

Os espectros podem ser obtidos por medidas de transmitância/absorbância amostras transparentes), transfectância, reflectância difusa (amostras sólidas), interactância e transmitância por dispersão média (amostras sólidas densas), sendo bastante utilizados para determinação de água, proteínas, hidrocarbonetos de baixo peso molecular e gorduras [117, 23].

2.6.3.2 Espectrometria UV-Vis

A espectroscopia de absorção molecular no ultravioleta-visível (UV-Vis) utiliza a radiação delimitada pelos comprimentos de onda de 190 a 1100 nm, sendo amplamente utilizada nas mais variadas aplicações analíticas.

A absorção de radiação UV-Vis pela molécula do composto, produz uma modificação na estrutura eletrônica da molécula como resultado de transições eletrônicas que geralmente envolvem elétrons π , σ e n [117]. A **Figura 2.5** apresenta os níveis de energia eletrônica molecular e as transições envolvidas.

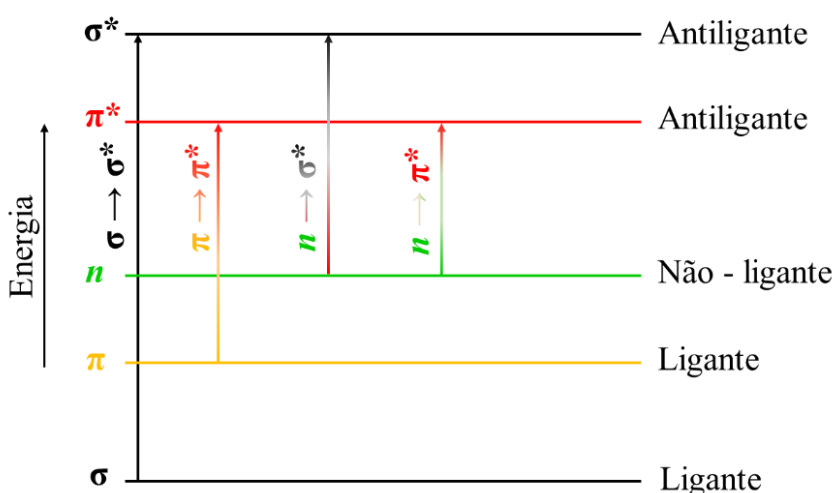


Figura 2.5 - Níveis de energia eletrônica molecular e as transições envolvidas.
Fonte: adaptada de Skoog, Holler e Neiman (2006).

Essa absorção está restrita a moléculas que apresentam grupos funcionais, chamados *cromóforos*, que contém elétrons de valência com energias de excitação relativamente baixas, que podem ser promovidos a níveis de energia maiores [117, 121].

As transições eletrônicas que acontecem nos compostos orgânicos normalmente são dos tipos $n \rightarrow \pi^*$ e $\pi \rightarrow \pi^*$, em razão das energias necessárias para esses processos. Em ambos os casos, as transições requerem a presença de pelo menos um grupo funcional insaturado (cromóforo) para fornecer os orbitais π [117].

As bandas de absorção caracterizam-se pela sua posição e intensidade. Enquanto a posição corresponde ao comprimento de onda (λ) da radiação eletromagnética responsável pela transição, a intensidade está associada, principalmente, a energia dos orbitais moleculares e probabilidade de transições [121].

Embora a absorção de radiação UV-Vis dependa da presença do cromóforo e da estrutura molecular, mudanças significativas na posição e intensidade das bandas podem ocorrer em decorrência da conjugação, por exemplo, entre ligações duplas ou entre grupos de cromóforos presentes na estrutura da molécula [121].

Além desses fatores, a polaridade do solvente utilizado pode influenciar na posição e na intensidade das bandas. Picos associados a transições $n \rightarrow \pi^*$ geralmente são deslocados para comprimentos de onda menores ao se aumentar a polaridade do solvente, efeito chamado *deslocamento hipsocrômico*. O comportamento oposto é observado para transições $\pi \rightarrow \pi^*$, sendo chamado *deslocamento batocrômico* [117]. Em geral, os espectros de absorção UV-Vis apresentam bandas largas resultantes da sobreposição dos sinais provenientes de transições vibracionais e rotacionais aos sinais associados às transições eletrônicas [121].

2.6.3.3 Espectrometria de fluorescência molecular

A absorção de um quantum de luz promove um elétron do estado energético vibracional mais baixo S_0 (estado fundamental) a níveis superiores de energia. Ao retornar ao estado fundamental, parte da energia absorvida é dissipada através da emissão de um fóton de luz na região do UV-Vis, processo conhecido como *luminescência*. Quando essa emissão se dá a partir do primeiro estado singleto excitado S_1 , ocorre o fenômeno chamado *fluorescência* [122]. Os processos de excitação e desativação molecular (volta ao estado fundamental com liberação de energia) que ocorrem após a absorção de radiação eletromagnética estão esquematizados no diagrama de Jablonski na **Figura 2.6**.

No diagrama, a linha horizontal grossa na parte inferior corresponde a energia do estado fundamental S_0 da molécula. As linhas horizontais grossas superiores representam os níveis de energia para os estados fundamentais vibracionais de três estados eletrônicos excitados: S_1 (singleto), S_2 (singleto) e T_1 (triplete) [117]. As flechas verticais retas indicam processos

associados a absorção ou emissão de radiação, como os fenômenos de fluorescência e fosforescência. As flechas verticais sinuosas representam os processos de desativação não-radioativos, que incluem relaxação vibracional, conversão interna, conversão externa, cruzamento intersistema e fosforescência [117, 121].

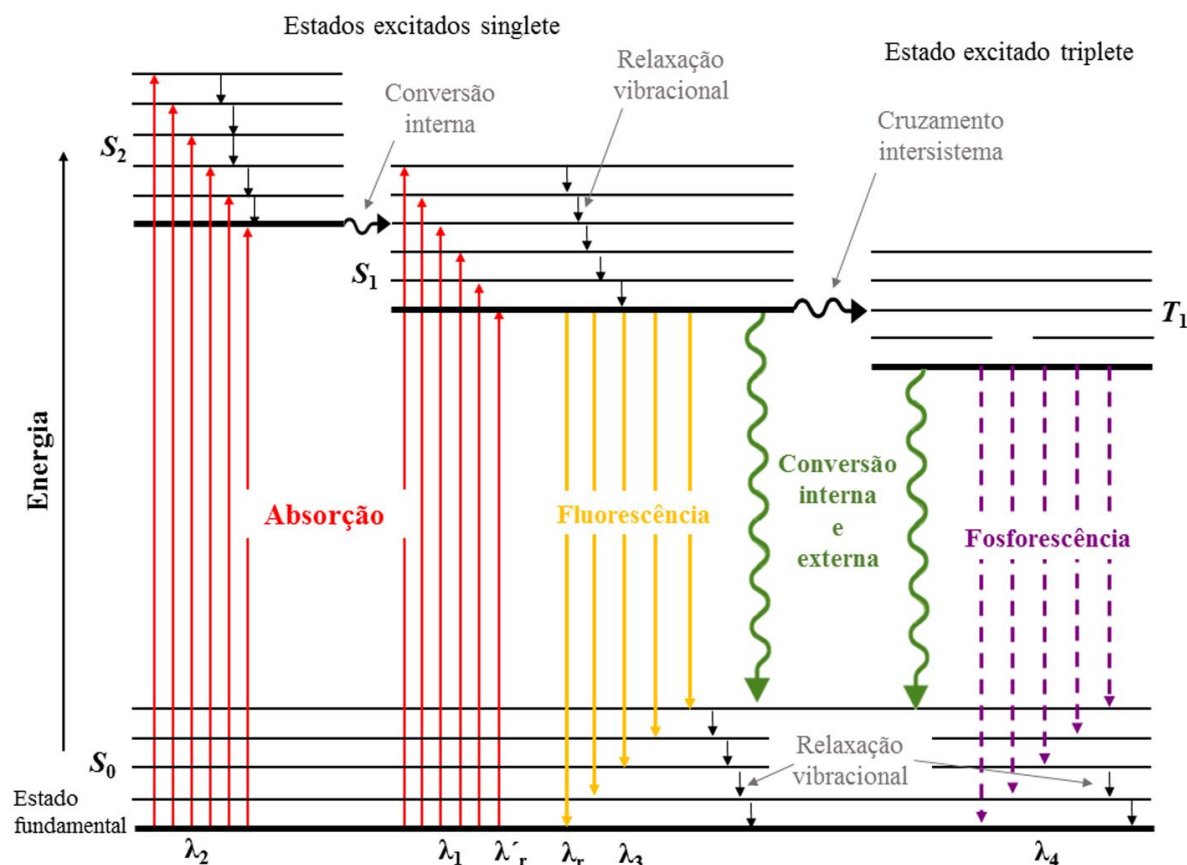


Figura 2.6 - Diagrama de Jablonski.

Fonte: adaptada de Skoog, Holler e Neiman (2006).

O processo de desativação pode se dar pela fluorescência, fosforescência ou ainda por algum dos processos não radioativos. O processo que irá ocorrer deve ser aquele que minimiza o tempo de vida do estado excitado [121].

A fluorescência de um composto é influenciada, sobretudo, pela sua estrutura molecular e está predominantemente associada às transições entre os orbitais $\pi \rightarrow \pi^*$ (com emissão no processo inverso), desde que estas sejam as transições menos energéticas [117, 122]. As moléculas com elevado número de elétrons π são potencialmente fluorescentes. Nesse sentido, os compostos contendo grupos funcionais aromáticos com transições $\pi \rightarrow \pi^*$ de baixa energia apresentam a fluorescência mais intensa. Compostos contendo estruturas alifáticas, alicíclicas

carbonílicas ou estruturas de ligações duplas altamente conjugadas também podem apresentar fluorescência, mas em quantidade inferior quando comparado ao dos sistemas aromáticos ^[117].

Além de sua estrutura molecular, para ocorrer a fluorescência, a molécula precisa estar inserida em um ambiente químico favorável. Entre os fatores que favorecem a emissão, destacam-se a temperatura, o pH, os efeitos dos substituintes e o solvente ^[117].

Três principais tipos de espectros de fluorescência podem ser obtidos: espectro de emissão, espectro de excitação e espectro de excitação/emissão sincronizado ^[122]. O espectro de emissão é obtido através da irradiação da amostra com um comprimento de onda de excitação fixo enquanto a intensidade da luz reemitida numa faixa espectral definida é medida. Para a obtenção de espectros de excitação, mantém-se constante o comprimento de onda de emissão enquanto é realizada uma varredura no comprimento de onda de excitação. Neste caso, o comprimento de emissão é escolhido com base na posição do máximo do espectro de emissão do composto. O registro do espectro sincronizado é usado quando se tem uma mistura de fluoróforos resultando em uma sobreposição espectral. Consiste na varredura simultânea de uma faixa espectral de comprimentos de onda de excitação e de emissão apresentando um intervalo ($\Delta\lambda$) constante entre eles ^[122].

2.6.3.4 Imagens digitais

As técnicas baseadas em imagens digitais têm encontrado grande aplicação na química analítica, sendo predominante utilizadas em problemas que envolvem controle de qualidade e métodos de reconhecimento de padrões. A disponibilidade e o baixo custo dos computadores pessoais, câmeras digitais e aparelhos de digitalização, além da simplicidade e rapidez com que as análises são realizadas, têm contribuído para o aumento do número de ensaios que empregam esses equipamentos ^[123].

Além das vantagens citadas acima, as imagens digitais têm sido muito utilizadas nas análises de rotina dos padrões de qualidade, pois permitem monitorar parâmetros como tamanho, cor e textura, eliminando a subjetividade da análise e a dependência do sistema visual humano, que por sua vez, é profundamente influenciado pelas condições ambientes, podendo levar a uma série de inconsistências ^[124, 125].

Uma imagem digital é o resultado da conversão de uma imagem analógica em uma matriz ($m \times n$) de números digitais que pode ser processada por computadores. Nessa matriz, cada elemento é denominado *pixel* ou *elemento de imagem* ^[123, 126]. A **Figura 2.7** apresenta a representação matricial, ou digital, de uma imagem analógica colorida.

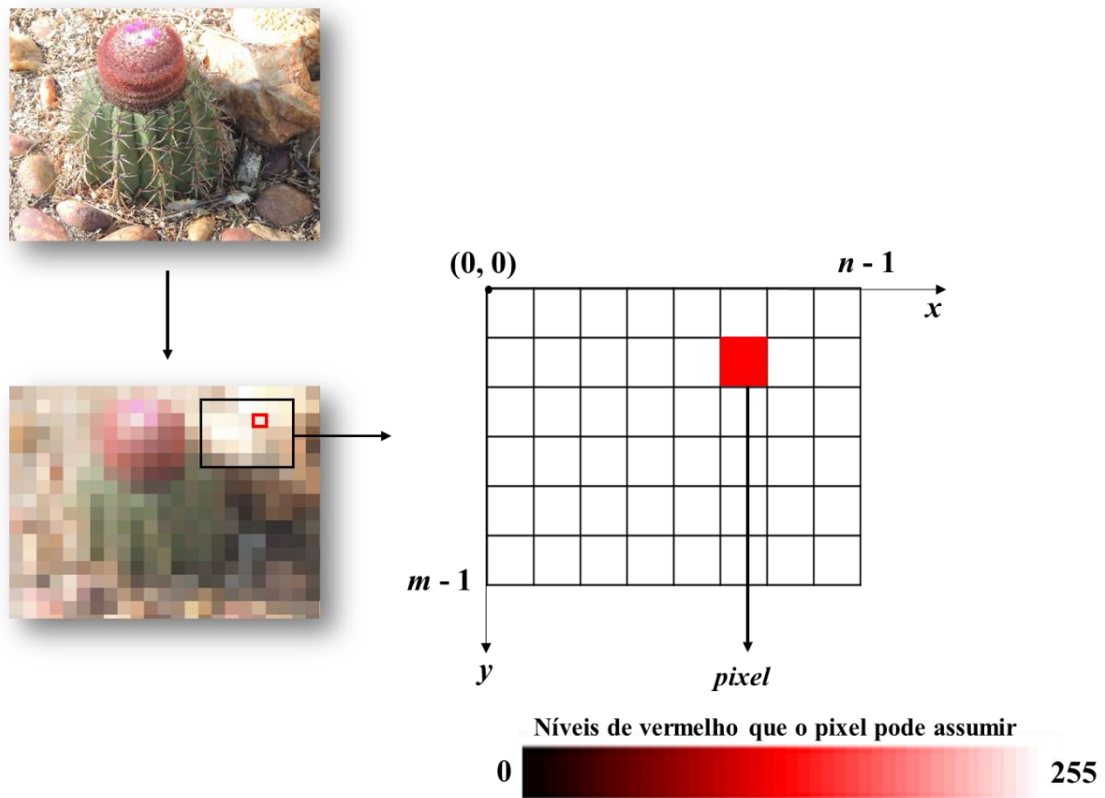


Figura 2.7 - Imagem analógica colorida e sua representação em imagem digital.
Fonte: adaptada de Queiroz e Gomes (2001).

Cada pixel apresenta um único valor numérico correspondente ao nível de sinal naquele ponto ^[127]. A conversão do conjunto numérico formado pelos pixels a uma imagem é realizada por meio de um mapa de cores, onde é designado um nível (ou tom) específico de cor para cada ponto numérico da imagem. Os níveis das cores variam num intervalo de 0 a 255 de acordo com o nível do sinal ^[123, 127].

O conceito de qualidade de uma imagem está relacionado a dois parâmetros principais: resolução espacial e quantidade de bits utilizada para quantificação da informação da imagem ^[123].

A resolução espacial expressa o tamanho da matriz correspondente à imagem. De forma simplificada, quantas linhas e colunas, ou seja, quantos pixels são usados para cobrir o espaço visual captado pela imagem ^[127].

O número de bits define o número possível de valores de cores que um pixel pode assumir ^[127]. Por exemplo, em 1 bit por pixel ($2^1 = 2$), apenas duas cores podem ser assumidas: preto e branco. No caso das imagens no sistema RGB, a utilização de 24 bits permite alocar 8 bits para cada componente da cor, totalizando 256 níveis de cores por componente e 16,7 milhões de cores totais na imagem.

Em aplicações analíticas qualitativas, normalmente são utilizados histogramas de cores obtidos a partir de um dado modelo de cor. Os histogramas representam a distribuição de frequência dos valores que um pixel pode assumir. Trata-se do percentual de pixels naquela imagem que apresenta um dado valor de cor ^[123, 127].

Capítulo III

SPA E TRANSFERÊNCIA DE MODELOS

3 SPA E TRANSFERÊNCIA DE MODELOS

Os modelos SPA-MLR e SPA-LDA foram inicialmente desenvolvidos sem mecanismos explícitos para procedimentos de transferência dos modelos resultantes.

Em 2005, Honorato *et al.* ^[12] propuseram uma modificação no SPA-MLR para que o mesmo pudesse ser utilizado na transferência de modelos de calibração, com seleção de variáveis robustas em relação às diferenças entre os dois instrumentos. A modificação consistiu na inclusão de um conjunto de amostras de transferência K_{trans} na seleção do melhor subconjunto de variáveis. A métrica proposta foi definida como

$$E = \frac{1}{2}(\text{RMSEV} + \text{RMSET}) \quad (3.1)$$

com RMSEV calculado usando objetos de validação medidos no instrumento primário, como na **Equação 2.3**. O RMSET corresponde a raiz quadrada do erro médio quadrático da predição no conjunto de transferência (RMSET: *Root Mean Square Error of prediction in the Transfer set*), definido conforme a **Equação 3.2**.

$$\text{RMSET} = \sqrt{\frac{1}{K_{trans}} \sum_{k=1}^{K_{trans}} (y_{trans,k} - \hat{y}_{trans,k})^2} \quad (3.2)$$

onde $y_{trans,k}$ corresponde ao valor de referência do parâmetro em consideração na k -ésima amostra de transferência e $\hat{y}_{trans,k}$ denota o correspondente valor previsto obtido pela aplicação do MLR às medidas realizadas no instrumento secundário.

A adaptação foi avaliada para dois estudos de caso: determinação de umidade em milho e previsão da temperatura de destilação na qual 90% das amostras de gasolina analisadas tenham sido evaporadas. Em ambos casos, espectros NIR das amostras foram coletados em dois instrumentos diferentes. Os resultados obtidos com a adaptação do SPA-MLR foram comparáveis e ligeiramente melhores que os obtidos para o modelo PLS com padronização PDS do conjunto de amostras medidas no instrumento secundário ^[12].

Até o presente momento, nenhuma adaptação tem sido proposta para o SPA com o propósito de transferência de classificação.

3.1 Os critérios propostos

A modificação do SPA-LDA proposta no presente trabalho segue as linhas da formulação do SPA-MLR para transferência de calibração descrita na seção anterior, com

inclusão da função de custo para um conjunto de amostras de transferência K_{trans} na seleção do melhor subconjunto de variáveis. A métrica proposta é definida como

$$G = \frac{1}{2}(G_{\text{train}} + G_{\text{trans}}) \quad (3.3)$$

com G_{train} obtido para os objetos de treinamento medidos no instrumento primário, como na **Equação 2.6**, e G_{trans} calculado para as amostras de transferência registradas no instrumento secundário usando a seguinte equação

$$G_{\text{transf}} = \frac{1}{K_{\text{trans}}} \sum_{k=1}^{K_{\text{trans}}} g_{\text{trans},k} \quad (3.4)$$

onde $g_{\text{trans},k}$ é definido como na **Equação 2.5** mas usando o k -ésimo objeto de transferência $\mathbf{x}_{\text{trans},k}$ registrado no instrumento secundário, em vez do k -ésimo objeto de validação $\mathbf{x}_{\text{val},k}$.

A expressão expandida da **Equação 3.3** é apresentada abaixo.

$$G = \frac{1}{2} \left[\left(\frac{1}{K_{\text{train}} - L - C} \sum_{k=1}^{K_{\text{train}}} g_{\text{train},k} \right) + \left(\frac{1}{K_{\text{trans}}} \sum_{k=1}^{K_{\text{trans}}} g_{\text{trans},k} \right) \right] \quad (3.5)$$

Após calcular os valores de G_{train} e G_{trans} para cada subconjunto candidato de variáveis, dois critérios possíveis serão considerados aqui:

- Critério C1: Seleciona o subconjunto candidato de variáveis correspondente ao menor valor do custo $G = \frac{1}{2}(G_{\text{train}} + G_{\text{trans}})$

- Critério C2: Seleciona o subconjunto candidato de variáveis correspondente ao menor valor do custo G_{trans} .

Assim como na **Equação 3.1**, o critério C1, voltado para problemas de classificação, busca obter um modelo com bom desempenho de classificação tanto no instrumento primário quanto no secundário. A utilização do critério C2 será investigada como uma alternativa que leva em conta a classificação das amostras de transferência, entendendo-se que o modelo resultante será utilizado apenas no instrumento secundário.

Uma das principais vantagens dos critérios propostos é que as amostras de transferência empregadas na avaliação da métrica G_{trans} só precisam ser medidas no instrumento secundário, bem como acontece com a métrica do RMSET para problemas de calibração. Na prática, esta é uma vantagem importante em relação aos métodos de padronização como DS e PDS,

especialmente se os instrumentos primário e secundário não estiverem localizados no mesmo laboratório, se as amostras estiverem deterioradas, quando se tem dificuldade no transporte das amostras ou se o instrumento principal não estiver mais disponível.

Capítulo IV

METODOLOGIA

4 METODOLOGIA

4.1 Conjuntos de amostras

4.1.1 Etanol Hidratado Combustível

Um total de 102 amostras de EHC (52 não-adulteradas e 50 adulteradas) foram utilizadas neste estudo de caso. A quantidade de amostras em cada subconjunto é detalhada na **Tabela 4.1**. As amostras não-adulteradas foram fornecidas pelo Laboratório de Combustíveis (LAC) da Universidade Federal de Pernambuco (UFPE) e todas encontravam-se dentro das especificações da ANP.

As adulterações foram preparadas em laboratório pela adição de metanol na faixa de 1,86 a 13,0 % (m/m). Para incorporar variabilidade ao conjunto de dados, as adulterações foram realizadas com diferentes amostras e de forma aleatória. As adulterações e as medidas das amostras de EHC foram previamente realizadas em um trabalho de mestrado realizado por Silva (2013) ^[84].

4.1.2 Azeite de oliva extra virgem (Espectrometria UV-Vis e de fluorescência molecular)

Duas classes de AOEV foram utilizadas neste trabalho: não-adulteradas e adulteradas. As amostras não-adulteradas foram adquiridas nos supermercados da cidade de João Pessoa com diferentes lotes, sendo do mesmo fabricante. A quantidade de amostras analisadas em cada técnica é apresentada na **Tabela 4.2**. O fabricante foi escolhido com base em uma investigação realizada por uma Associação Brasileira de Defesa do Consumidor ^[128], que avalia a qualidade dos produtos comercialmente disponíveis.

As adulterações foram realizadas em laboratório pela adição de óleo vegetal comestível de soja (também de um único fabricante) na proporção de 1,0 a 30% (m/m) com intervalo de 5%. O procedimento foi realizado de forma aleatória a fim de incluir variabilidade ao conjunto de dados.

4.1.2.1 Determinação da extinção específica pela absorção na região do ultravioleta

Com o intuito de avaliar a autenticidade das amostras de AOEV não-adulteradas, foi realizada a determinação da extinção específica por absorção na região ultravioleta (Método Oficial AOCS 5-91), utilizando ciclohexano de grau espectrofotométrico como solvente, conforme descrito em Zenebon, Pascuet e Tiglea (2008) ^[129].

Este método utiliza a região ultravioleta para obter informações sobre a qualidade e estado de conservação. Os sistemas dieno e trieno conjugados, cuja presença em AOEV pode

ser atribuída à oxidação e/ou refino, são avaliados nos comprimentos de onda de 232 e 270 nm^[129]. Os valores da extinção específica (K_{232} e K_{270}) e da variação da extinção específica (ΔK) obtidos nestes comprimentos de onda, são comparados com os limites previamente estabelecidos^[99] e podem ser utilizados para diferenciar os AOEV das outras categorias de óleos. Os valores de K e ΔK para cada comprimento de onda são calculados da seguinte forma:

$$K_{1\text{ cm}}^{1\%} = K_{\lambda} = \frac{A_{\lambda}}{c \times 1} \quad (4.1)$$

$$\Delta K = \lambda - \frac{K_{\lambda-4} + K_{\lambda+4}}{2} \quad (4.2)$$

onde K_{λ} , A_{λ} , c e l correspondem à extinção específica no comprimento de onda λ , absorvância medida no comprimento de onda λ , concentração da solução em g/100 mL e caminho óptico da cubeta em cm, respectivamente.

Os valores medidos para K_{232} , K_{270} e ΔK são apresentados nos resultados. Em decorrência da dificuldade de se encontrar amostras com lotes diferenciados, apenas 7 amostras não-adulteradas foram adquiridas, e o conjunto de dados foi expandido pela mistura dessas amostras em diferentes proporções.

4.1.3 Azeite de oliva extra virgem - Imagens digitais

Neste caso, foi utilizado um fabricante diferente do utilizado na seção 4.1.2, mas que também foi escolhido com base em uma investigação realizada pela PROTESTE^[128]. As adulterações foram realizadas pela adição de óleo vegetal comestível de soja (também de um único fabricante) na proporção de 0,5 a 10% (m/m), conforme realizado em 4.1.2. A quantidade de amostras analisadas é apresentada na **Tabela 4.2**.

Todas as amostras foram armazenadas em frascos de vidro âmbar, protegidas da luz e mantidas a uma temperatura de aproximadamente 23 ± 2 °C até a momento da análise. Não foi realizado nenhum pré-tratamento das amostras.

4.2 Registro do sinal

4.2.1 Espectros NIR

Os espectros NIR foram registrados na faixa de 12.004 a 4.000 cm^{-1} utilizando-se um espectrômetro Spectrum GX FTIR Perkin Elmer (localizado no Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria – LAQA da Universidade Federal da Paraíba) que foi considerado o instrumento primário, e um espectrômetro NIR dispersivo Foss Analytical XDS (localizado na Embrapa, Centro Nacional de Pesquisa de Algodão, Campina

Grande, PB, Brasil), utilizado como instrumento secundário. Ambos os instrumentos foram equipados com uma célula de quartzo com 1,0 mm de percurso óptico. Neste caso, apenas as amostras de transferência e teste foram analisadas no instrumento secundário.

Os espectros foram registrados com uma média de 32 varreduras em ambos os espectrômetros, com resolução espectral de 8 cm^{-1} para o espectrômetro Spectrum GX FTIR Perkin Elmer e $0,5\text{ cm}^{-1}$ para o um espectrômetro NIR dispersivo Foss Analytical XDS. Os espectros do branco foram obtidos usando uma célula vazia limpa. A temperatura foi controlada a $23 \pm 1\text{ }^{\circ}\text{C}$ durante todo o processo de aquisição espectral.

4.2.2 Espectros UV-Vis

Os espectros UV-Vis das amostras de AOEV foram registrados no Brasil (Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria da Universidade Federal da Paraíba - instrumento A) e na Argentina (Laboratório FIA, INQUISUR-CONICET, da Universidad Nacional del Sur - instrumento B) utilizando dois espectrofotômetros de absorção molecular Hewlett-Packard modelo HP 8453 UV-Vis. Em ambos os casos, o espectrofotômetro foi equipado com uma célula de quartzo (caminho óptico de 10 mm) e os espectros foram coletados, em triplicata, na faixa de 190 - 1100 nm, com resolução de 1 nm. O ajuste do sinal de transmitância foi realizado utilizando isoctano como branco.

Neste caso, as diferenças espectrais estão associadas sobretudo às condições ambientais, nas quais as amostras foram analisadas, e às diferenças entre o tempo de vida de cada instrumento.

4.2.3 Espectros de emissão de fluorescência molecular

Os espectros foram adquiridos nas mesmas condições experimentais utilizando um espectrofluorímetro Jasco FP-6500 (instrumento A) e um espectrofluorímetro Aminco Bowman Série 2 (instrumento B), ambos localizados no Laboratório FIA, INQUISUR-CONICET, da Universidad Nacional del Sur na Argentina. Essa situação envolve dois instrumentos com diferença significativa de qualidade do sinal registrado. Para o modelo Aminco Bowman Série 2, a qualidade é insatisfatória devido ao elevado tempo de vida, manutenções periódicas irregulares e qualidade inferior desse instrumento quando comparado ao modelo Jasco FP-6500.

Os espectros foram registrados entre 300 a 750 nm, com resolução de 1 nm, usando comprimento de onda de excitação na faixa de 280 a 480 nm com intervalo de 10 nm. A faixa

de trabalho de 350 a 600 nm de emissão, a 330 nm de excitação, foi empregada no procedimento quimiométrico.

Os equipamentos foram equipados com uma lâmpada de xenônio de 150 W, um fotomultiplicador de 400 W e uma célula de quartzo com $10 \times 10 \times 45$ mm para geometria em ângulo reto. A largura da fenda foi de 3 nm para excitação e 5 nm para emissão. O intervalo de aquisição e o tempo de integração foram mantidos a 1 nm e 0.5 s, respectivamente.

4.2.4 Histogramas de cores das imagens

As imagens digitais foram registradas usando dois dispositivos com resoluções diferentes: uma webcam (Microsoft LifeCam Studio com resolução HD 1280 x 720), considerada o instrumento primário, e um smartphone (Samsung Galaxy S3 mini com resolução 1536 x 2560), utilizada como instrumento secundário. Em ambos os casos, um compartimento foi utilizado para captura das imagens, construído como mostra a **Figura 4.1**. Esse ambiente foi utilizado com o intuito de facilitar o controle da luz incidente sobre as amostras e garantir a qualidade e uniformidade das imagens obtidas.

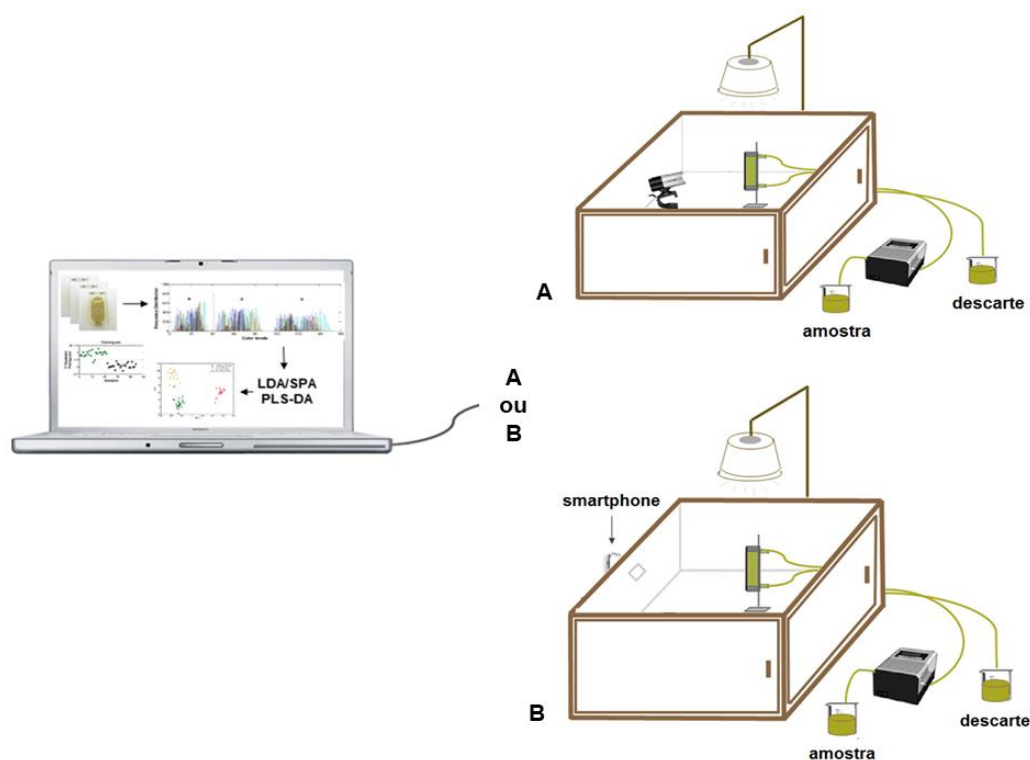


Figura 4.1 - Sistema usado na aquisição das imagens de AOEV.

Basicamente, o sistema empregado é formado por um compartimento, no interior do qual foram colocadas a webcam (quando a configuração A foi utilizada) posicionada

frontalmente a 10 cm da cubeta de fluxo de quartzo de 1 cm de caminho óptico; uma bomba peristáltica (modelo IPC com 8 canais e 12 roletes) usada para propulsão das amostras de forma rápida e sem necessidade de manipulação do sistema; um notebook e uma lâmpada de LED de 5 W na cor branca e temperatura fria, responsável pela iluminação do sistema. A captura das imagens pelo smartphone foi realizada com o mesmo no lado externo do compartimento a uma distância de 20 cm da cubeta (configuração **B**). As imagens foram registradas através de uma abertura com 1 cm × 2 cm, equivalente as dimensões da câmera do smartphone.

Para cada amostra, três imagens sequenciais foram registradas com 24 bits (16,7 milhões de cores) e armazenadas em formato JPEG. Nenhum pré-tratamento foi realizado nas imagens. Após a captura, as imagens foram decompostas em histogramas correspondentes às cores vermelha, verde e azul, que constituem o modelo RGB ^[130]. Os histogramas correspondiam a uma região específica da imagem, com 10 mm de raio, localizada no centro da cubeta. Essa região foi mantida constante ao longo de toda análise.

A média dos três histogramas, referentes às distribuições de frequência (ou taxa de ocorrência) dos índices de cores, foi calculada e utilizada como resposta instrumental referente àquela amostra em todo procedimento quimiométrico. A matriz de dados era então formada pelas amostras, dispostas nas linhas, e pelas variáveis correspondentes aos níveis de cores obtidos para cada componente de cor, dispostas nas colunas.

4.3 Seleção de amostras

As amostras analisadas no instrumento primário foram divididas em conjuntos de treinamento (EHC: 60%; AOEV: 70%) e teste (EHC: 40%; AOEV: 30%) usando o algoritmo clássico Kennard - Stone (KS) ^[77]. As **Tabelas 4.1** e **4.2** apresentam o número de amostras de treinamento e teste em cada classe, para cada conjunto de dados.

Tabela 4.1 - Número de amostras de treinamento e teste selecionadas pelo KS para as duas classes de EHC. *N* representa o número de amostras em cada classe.

Classes	<i>N</i>	Subconjuntos	
		Treinamento	Teste
Não-adulteradas	52	32	20
Adulteradas	50	30	20
Total	102	62	40

Tabela 4.2 - Número de amostras de treinamento e teste selecionadas pelo KS para as duas classes de AOEV. N representa o número de amostras em cada classe.

Classes	UV-Vis			Fluorescência molecular			Imagens digitais		
	Subconjuntos			Subconjuntos			Subconjuntos		
	N	Treinamento	Teste	N	Treinamento	Teste	N	Treinamento	Teste
Não-adulterada	49	34	15	48	34	14	21	14	7
Adulterada	40	28	12	40	28	12	33	23	10
Total	89	62	27	88	62	26	54	37	17

As amostras de treinamento foram empregadas nos procedimentos de modelagem, incluindo a seleção de variáveis SPA para LDA, enquanto as amostras de teste, medidas em ambos os instrumentos, foram utilizadas apenas na avaliação final dos modelos de classificação. Os modelos LDA foram desenvolvidos com base nas variáveis selecionadas pelo algoritmo de projeções sucessivas (SPA) adaptado para validação interna ^[51].

O algoritmo KS também foi utilizado para selecionar subconjuntos de N_{trans} amostras de transferência a partir do conjunto de treinamento. Diferentes N_{trans} (4, 6, 8 e 10) foram investigados, sendo cada subconjunto constituído pelo mesmo número de amostras de cada classe. O procedimento de seleção de amostras de transferência é ilustrado na **Figura 4.2**.

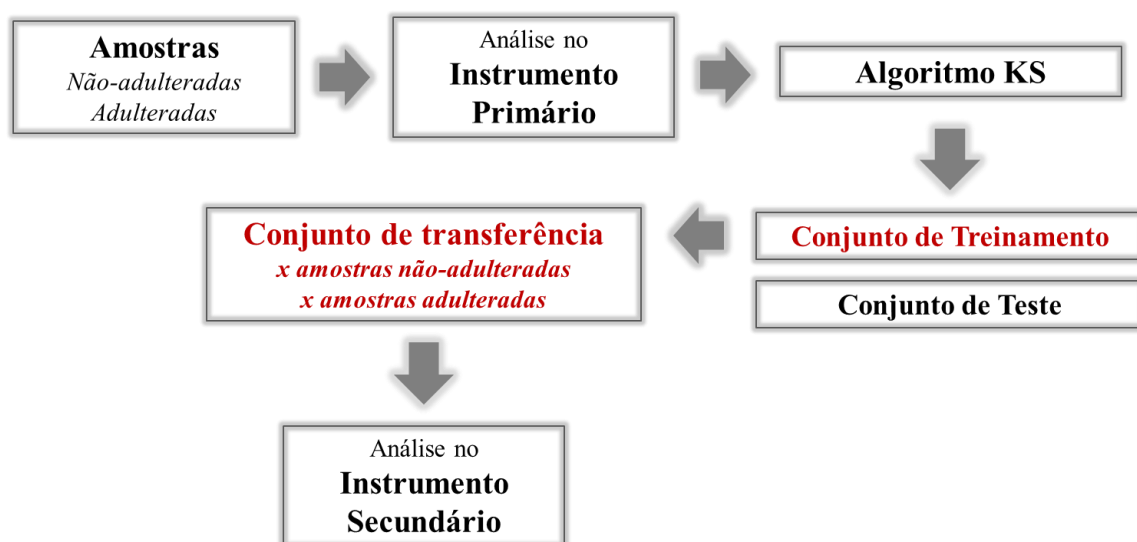


Figura 4.2 - Procedimento de seleção de amostras de transferência a partir do conjunto de treinamento medido no instrumento primário.

4.4 Procedimento quimiométrico e *software*

Estudos preliminares foram realizados com o intuito de avaliar a existência de agrupamentos naturais ou tendências de comportamento. Para isso, a PCA foi aplicada às amostras do conjunto de teste medidas nos dois instrumentos.

Os modelos de classificação foram desenvolvidos a partir do SPA-LDA descrito por Soares *et al.* (2014) ^[51] com a introdução dos dois critérios aqui propostos. Conforme mencionado na seção 3.1, esses critérios envolvem a inclusão das amostras de transferência no cálculo do custo para cada subconjunto de variáveis consideradas. Os resultados obtidos para os três modelos, sem o uso de amostras de transferência e com cada um dos dois critérios para o uso das amostras de transferência, foram comparados em termos de número de erros para os conjuntos de treinamento e teste.

Duas estratégias de padronização foram utilizadas: padronização direta (DS) e padronização direta em etapas (PDS) do conjunto de teste medido no instrumento secundário. O algoritmo PDS foi executado com diferentes tamanhos de janela (3, 5, 7, 9 e 11). Os algoritmos DS e PDS foram realizados usando o PLS Toolbox (versão 3.5) para Matlab.

Ainda a título de comparação, valores de sensibilidade, especificidade e acurácia foram calculados, em termos do conjunto de teste medido no instrumento secundário, para os critérios e para o modelo SPA-LDA na forma usual após os procedimentos de padronização.

No caso dos dados de fluorescência molecular e imagens digitais, os resultados obtidos pelos modelos SPA-LDA na forma usual e PLS-DA após procedimentos de padronização, foram comparados em termos do número de erros para os conjuntos de teste medidos nos instrumentos secundários antes e após a padronização. Além disso, também foram avaliados o número de amostras de transferência empregadas em cada modelo.

Para o modelo PLS-DA, foi adotado um valor limiar de 0,5. Quando um valor acima de 0,5 é previsto, a amostra é considerada como pertencente à classe em estudo, enquanto um valor abaixo de 0,5 indica que a amostra não pertence aquela classe.

As informações analíticas contidas nos histogramas das imagens digitais foram obtidas usando uma interface gráfica codificada em Matlab® 7.1. Os algoritmos KS e SPA-LDA foram codificados em Matlab (Mathworks, EUA). Os modelos PLS-DA e as correções espectrais foram realizadas utilizando o Unscrambler X.1 (CAMO S/A).

4.5 Notação

Nas discussões dos resultados, a notação "erros de teste (P)" será usada para representar os erros de teste obtidos aplicando o modelo de classificação aos dados do instrumento primário, enquanto que os "erros de teste (S)" indicarão os erros de teste obtidos pela aplicação do modelo de classificação aos dados do instrumento secundário. Os critérios propostos serão representados pelo símbolo $C1$ para a **Equação 3.3** e $C2$ para o custo representado na **Equação 3.4**.

Capítulo V

RESULTADOS E DISCUSSÃO

5 RESULTADOS E DISCUSSÃO

5.1 Espectroscopia NIR

5.1.1 Análise espectral

A **Figura 5.1** mostra os espectros NIR, na região de 12.004 a 4.000 cm^{-1} , das 102 amostras de EHC medidas no instrumento primário. Os espectros apresentam bandas atribuídas ao primeiro sobretom (6.000 – 5.000 cm^{-1}) e regiões de combinação (4.600 – 4.000 cm^{-1}) de estiramento C - H ^[131]. Adicionalmente, a 5.128 cm^{-1} ocorrem bandas de combinação de O - H em amostras de álcool ^[132].

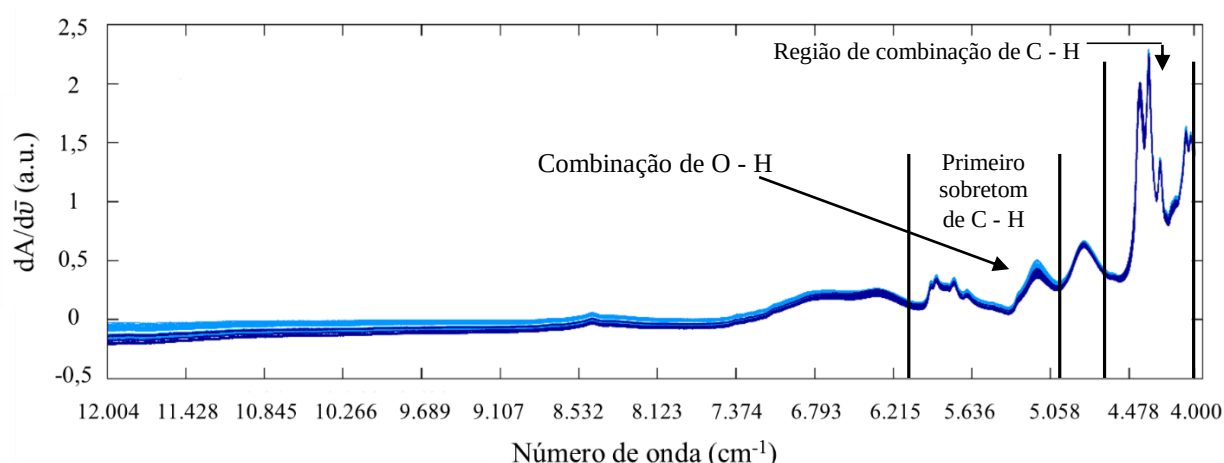


Figura 5.1 - Espectros NIR das 102 amostras de EHC registrados no instrumento primário. Amostras não-adulteradas (—) e amostras adulteradas (—).

Como pode ser observado, os espectros apresentaram variações sistemáticas na linha de base, que foram corrigidas com a primeira derivada usando um filtro Savitzky-Golay com polinômio de segunda ordem e janela de 21 pontos. Os espectros resultantes, que foram empregados em todo procedimento quimiométrico, são apresentados na **Figura 5.2** para as amostras medidas nos dois instrumentos.

É possível observar que a diferença entre os dois espectros é acentuada nas regiões em torno de 5.854 cm^{-1} e abaixo de 4.380 cm^{-1} .

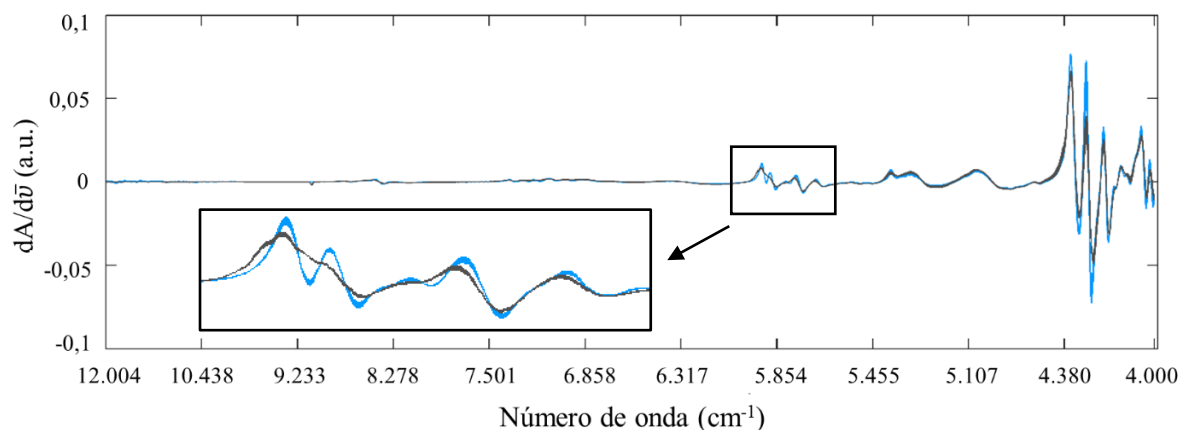


Figura 5.2 - Espectros NIR pré-processados das amostras de EHC medidas nos instrumentos primário (—) e secundário (—).

5.1.2 Análise exploratória

A **Figura 5.3** apresenta o gráfico dos escores obtidos por PC 1 $\times$ PC 2 para as amostras de teste medidas em ambos os instrumentos. O objetivo é avaliar como o sinal registrado para essas amostras é afetado pelas diferenças instrumentais.

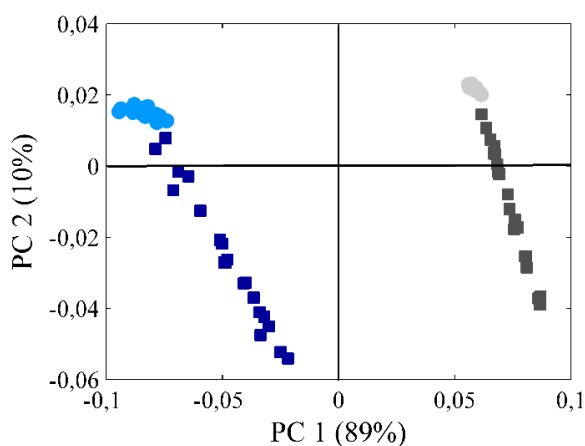


Figura 5.3 - Gráfico dos escores em PC 1 $\times$ PC 2 para as 80 amostras de EHC do conjunto de teste. Amostras não-adulteradas (●) e adulteradas (■) no instrumento primário, amostras não-adulteradas (●) e adulteradas (■) no instrumento secundário.

Ao longo da PC 1, que corresponde a 89% da variância explicada dos dados, ocorre a separação de grupos em função do instrumento no qual foram medidos, com as respostas obtidas no instrumento secundário assumindo valores positivos de escores. Na PC 2, observa-se uma tendência de discriminação das duas classes de EHC, com as amostras não-adulteradas apresentando geralmente valores positivos de escores. Neste caso, verifica-se uma dispersão das amostras adulteradas provavelmente relacionada aos diferentes níveis de adulteração analisados.

A **Figura 5.4** apresenta o gráfico dos *loadings* usados para verificar a contribuição das variáveis para as duas primeiras PCs. É possível observar que para as PCs 1 e 2, as variáveis de maior peso aparecem em torno de 4.380 cm^{-1} , região onde ocorre combinação de estiramento de C - H. Para a PC 1, o peso nessa região é ainda maior, indicando que a diferença entre os instrumentos é superior à diferença entre as classes de amostras.

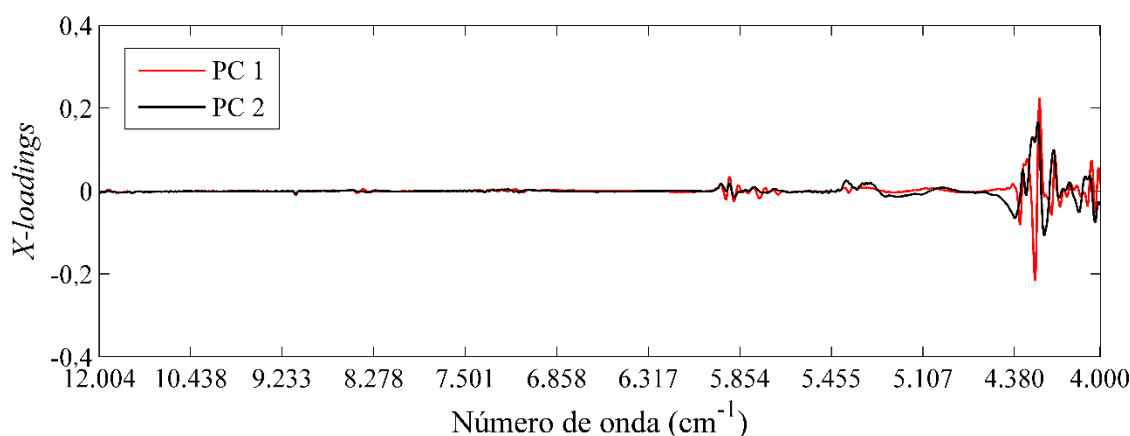


Figura 5.4 - Gráfico dos *loadings* para as duas primeiras PCs.

5.1.3 Modelos SPA-LDA

A **Tabela 5.1** apresenta os resultados de classificação obtidos pelo modelo SPA-LDA na sua forma usual construído para o instrumento primário.

Tabela 5.1 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário. O número entre parênteses apresenta a quantidade de variáveis empregadas no modelo. *N* indica o número de amostras de treinamento e teste em cada classe de EHC.

Conjunto de Treinamento				
Instrumento primário (1)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Classe verdadeira	<i>N</i>			97 ^a
Não-adulterada	32	32	-	100 ^b
Adulterada	30	2	28	93 ^c
Conjunto de Teste				
Instrumento primário (1)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Classe verdadeira	<i>N</i>			100 ^a
Não-adulterada	20	20	-	100 ^b
Adulterada	20	-	20	100 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Os acertos na classificação são obtidos quando os valores encontrados na “classe prevista” são iguais aos valores exibidos na “classe verdadeira”. Em contrapartida, os erros de classificação correspondem aos valores diferentes encontrados nas classes prevista e verdadeira.

Como pode ser observado, o modelo construído com base no instrumento primário apresentou um ótimo desempenho de classificação, com apenas duas amostras de treinamento incorretamente classificadas.

A **Tabela 5.2** apresenta os resultados obtidos com a utilização do modelo primário na classificação das amostras de teste medidas no instrumento secundário.

Tabela 5.2 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário aplicado às amostras de teste medidas no instrumento secundário. *N* indica o número de amostras de teste em cada classe de EHC.

Conjunto de teste secundário				
Classe Verdadeira	<i>N</i>	Classe Prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	20	20	-	100 ^b
Adulterada	20	20	-	0 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Como mostrado na tabela, 50% das amostras foram incorretamente classificadas quando o modelo primário foi aplicado às amostras medidas no instrumento secundário. Mais especificamente, todas as amostras adulteradas foram classificadas como não-adulteradas, indicando a necessidade do procedimento de transferência.

5.1.4 Transferência dos modelos SPA-LDA

Na **Tabela 5.3** são apresentados os erros para os conjuntos de treinamento e teste (P) obtidos pelos modelos SPA-LDA, com base no instrumento primário, na sua forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} .

Para o conjunto de treinamento, o número de amostras incorretamente classificadas pelo critério *C1* usando 6, 8 e 10 amostras de transferência foi maior em apenas uma unidade, quando comparado ao modelo usual, enquanto para *C2*, os erros foram um pouco maiores. No conjunto de teste (P), embora não tenham sido obtidos erros pelo modelo usual, uma quantidade pequena de erros foi encontrada pelos critérios, sobretudo por *C2*.

Tabela 5.3 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário.

Parâmetros	N	Resultados de classificação dos modelos SPA-LDA					
		Usual	Crítérios	$N_{trans} 4$	$N_{trans} 6$	$N_{trans} 8$	$N_{trans} 10$
Erros de treinamento	62	2	<i>C1</i>	8	3	3	3
			<i>C2</i>	6	7	7	7
Erros de teste (P)	40	0	<i>C1</i>	2	1	1	1
			<i>C2</i>	1	3	3	4

Esse aumento do número de erros é atribuído à inclusão das amostras de transferência no cálculo do custo G . O aumento é maior para o critério *C2*, tendo em vista que nesse caso, o custo baseia-se unicamente nas amostras medidas no instrumento secundário, conforme apresentado na seção 3.1.

A **Figura 5.5** apresenta os erros obtidos após aplicação dos modelos ao conjunto de amostras de teste medido no instrumento secundário. Os resultados para *C1* e *C2* são apresentados para diferentes valores de N_{trans} . O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.

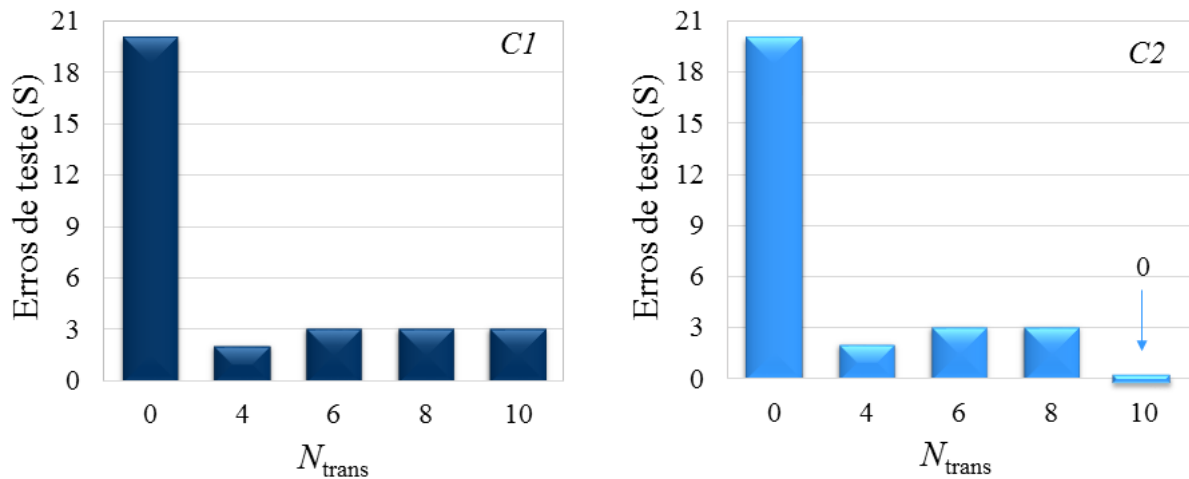


Figura 5.5 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de EHC. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.

Usando o modelo SPA-LDA na forma usual ($N_{trans} = 0$), o número de erros de teste (S) é muito maior do que o número de erros de teste (P) (**Tabela 5.3**), o que indica a necessidade de um procedimento de transferência.

Ainda na **Figura 5.5**, quando os modelos obtidos com os critérios *C1* e *C2* foram aplicados aos dados do instrumento secundário, verificou-se uma diminuição substancial dos erros de teste (S). Em particular, quando o critério *C2* foi aplicado com $N_{trans} = 10$, o número de erros de teste (S) foi reduzido a zero, isto é, todas as amostras de teste foram corretamente classificadas.

Os tipos de erros de classificação encontrados para cada critério são mostrados na **Tabela 5.4**. Para ambos os critérios, os mesmos tipos de erros foram encontrados quando 4, 6 e 8 amostras de transferência foram empregadas. Para o critério *C2*, nenhum erro foi obtido usando $N_{trans} = 10$.

Tabela 5.4 - Erros de classificação obtidos pelos critérios *C1* e *C2* aplicados às amostras de teste (S) medidas no instrumento secundário. *N* indica o número de amostras de teste em cada classe de EHC.

Critério <i>C1</i>									
Classe Verdadeira	<i>N</i>	<i>N_{trans} 4</i>		<i>N_{trans} 6</i>		<i>N_{trans} 8</i>		<i>N_{trans} 10</i>	
		Classe Prevista		Classe Prevista		Classe Prevista		Classe Prevista	
		NA	A	NA	A	NA	A	NA	A
NA	20	18	2	17	3	17	3	17	3
A	20	-	20	-	20	-	20	-	20

Critério <i>C2</i>									
Classe Verdadeira	<i>N</i>	<i>N_{trans} 4</i>		<i>N_{trans} 6</i>		<i>N_{trans} 8</i>		<i>N_{trans} 10</i>	
		Classe Prevista		Classe Prevista		Classe Prevista		Classe Prevista	
		NA	A	NA	A	NA	A	NA	A
NA	20	18	2	17	3	17	3	20	-
A	20	-	20	-	20	-	20	-	20

NA: Amostras não-adulteradas
A: Amostras adulteradas

A **Figura 5.6** mostra que, para todos os modelos, uma única variável foi selecionada. No entanto, o uso de amostras de transferência, quer com o critério *C1* ou critério *C2*, afetou a posição da variável selecionada.

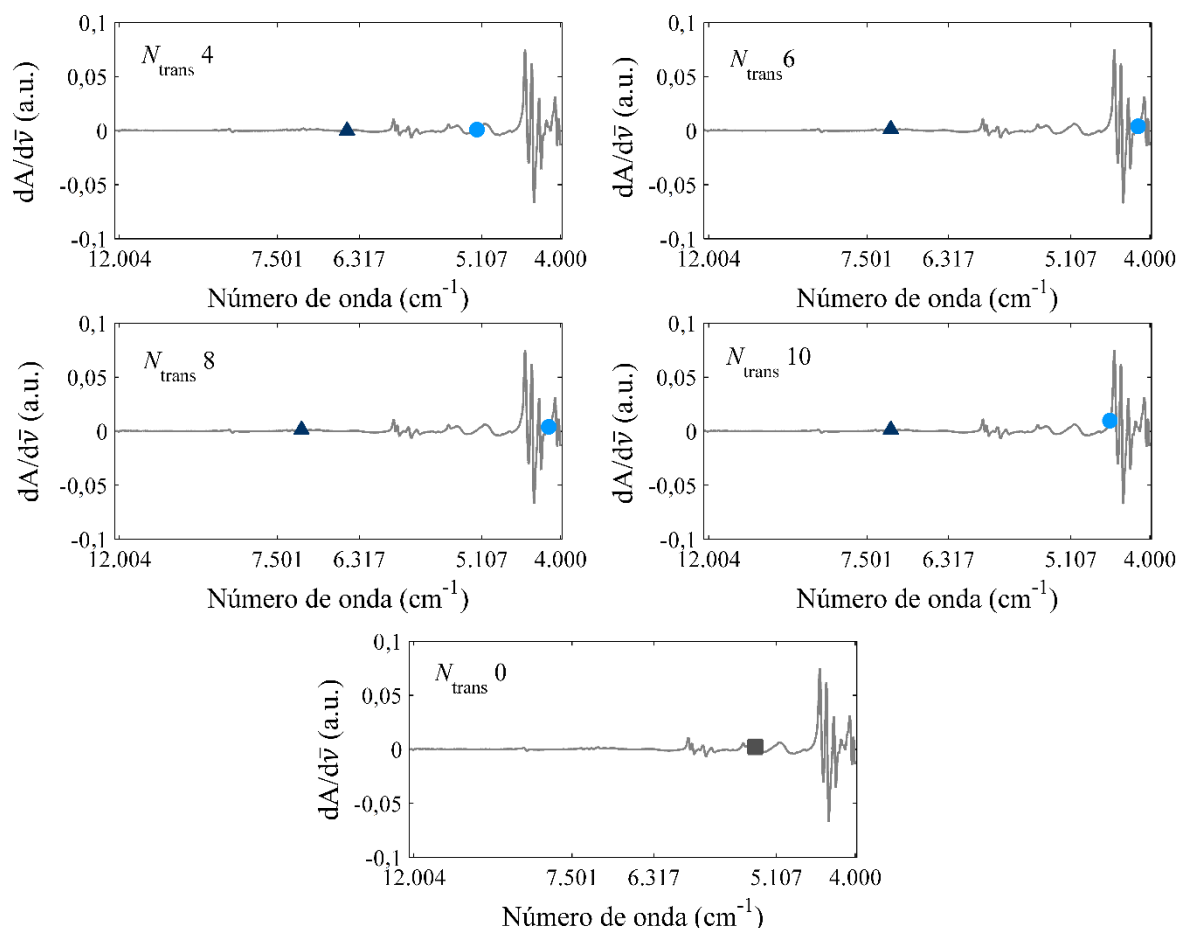


Figura 5.6 - Espectros NIR médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios $C1$: ▲, $C2$: ● e SPA na forma usual: ■.

Enquanto o critério $C1$ na maioria dos casos selecionou variáveis na região onde ocorrem bandas de primeiro sobretom de O - H, o critério $C2$, para diferentes N_{trans} , selecionou variáveis nas regiões do primeiro sobretom ($6.000 - 5.000 \text{ cm}^{-1}$) e de combinação ($4.600 - 4.000 \text{ cm}^{-1}$) de estiramento C - H. Na forma usual, o SPA-LDA selecionou por volta de 5.128 cm^{-1} , onde aparecem bandas de combinação de O - H em amostras de álcool.

Na **Figura 5.7a-c** são apresentados os gráficos contendo os valores de *threshold* para os três modelos aplicados ao conjunto de teste medido no instrumento secundário. As **Figuras 5.7b** e **5.7c** correspondem aos critérios $C1$ e $C2$ com $N_{trans} = 4$ e $N_{trans} = 10$, respectivamente, que forneceram os melhores resultados de classificação.

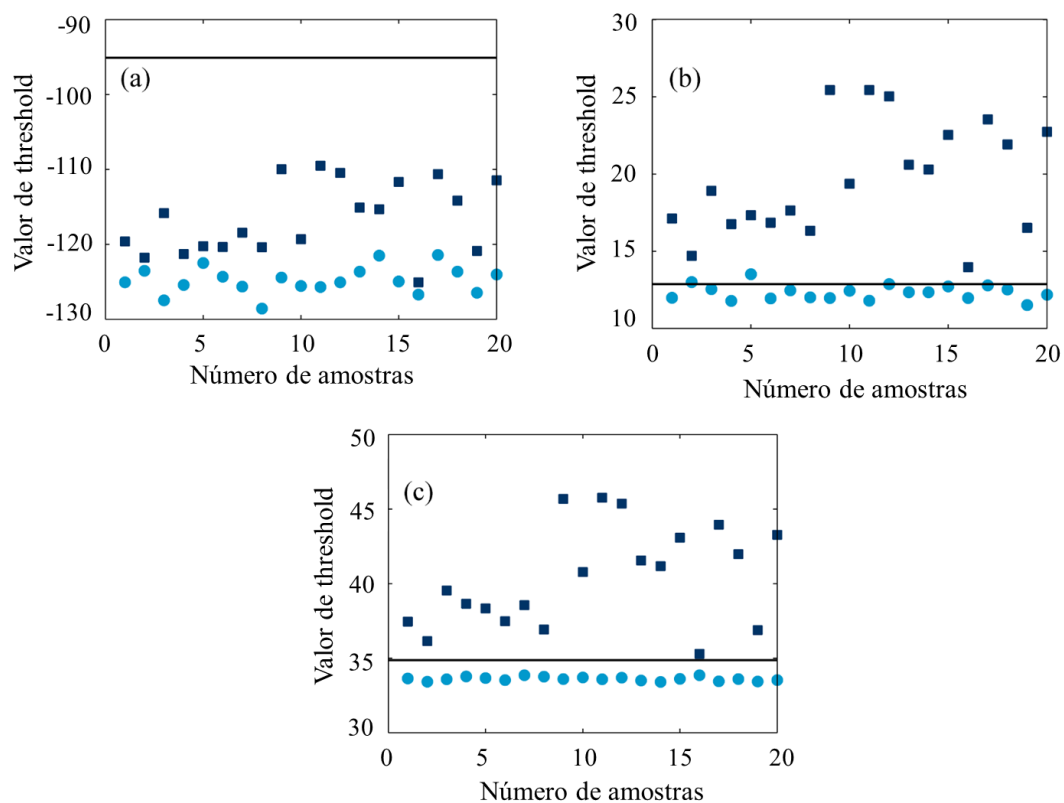


Figura 5.7 - Valores de *threshold* obtidos pelos modelos SPA-LDA usual (a), critério *C1* (b) e critério *C2* (c) na classificação das amostras de EHC medidas no instrumento secundário. Amostras não-adulteradas (●) e adulteradas (■).

Como pode ser visto, os critérios *C1* (**Figura 5.7b**) e *C2* (**Figura 5.7c**) alcançam uma boa discriminação entre as duas classes de amostras. Em ambos os casos, as amostras adulteradas foram encontradas acima do valor de *threshold*.

O número de erros obtidos pelo modelo SPA-LDA na forma usual após padronização DS e PDS do conjunto de teste (S) é apresentado na **Tabela 5.5**, onde foi observado que os procedimentos de padronização reduziram substancialmente os erros de classificação. O menor erro foi encontrado após DS com $N_{trans} = 4$ e PDS com $N_{trans} = 10$ (janela de três pontos), onde apenas duas amostras foram incorretamente classificadas.

Tabela 5.5 – Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização direta (DS) e padronização direta em etapas (PDS). A largura da janela utilizada no modelo PDS é apresentada entre parênteses.

Método	$N_{trans} 4$	$N_{trans} 6$	$N_{trans} 8$	$N_{trans} 10$
DS	2	3	3	3
PDS (3)	4	3	3	2
PDS (5)	4	3	3	3
PDS (7)	4	3	3	3
PDS (11)	3	3	3	3

Os resultados da transferência realizada por meio dos critérios propostos são comparados aos resultados alcançados com os métodos de padronização na **Figura 5.8**. É possível observar que os modelos SPA-LDA adaptados com os critérios *C1* e *C2* forneceram resultados na maioria das vezes similares àqueles obtidos pelos métodos de padronização. O destaque, contudo, é a classificação correta de todas as amostras de teste quando o critério *C2* foi aplicado com $N_{trans} = 10$, o que pode ser associado à quantidade significativa de amostras de transferência empregadas na construção do modelo, cujo cálculo do custo é baseado unicamente nessas amostras.

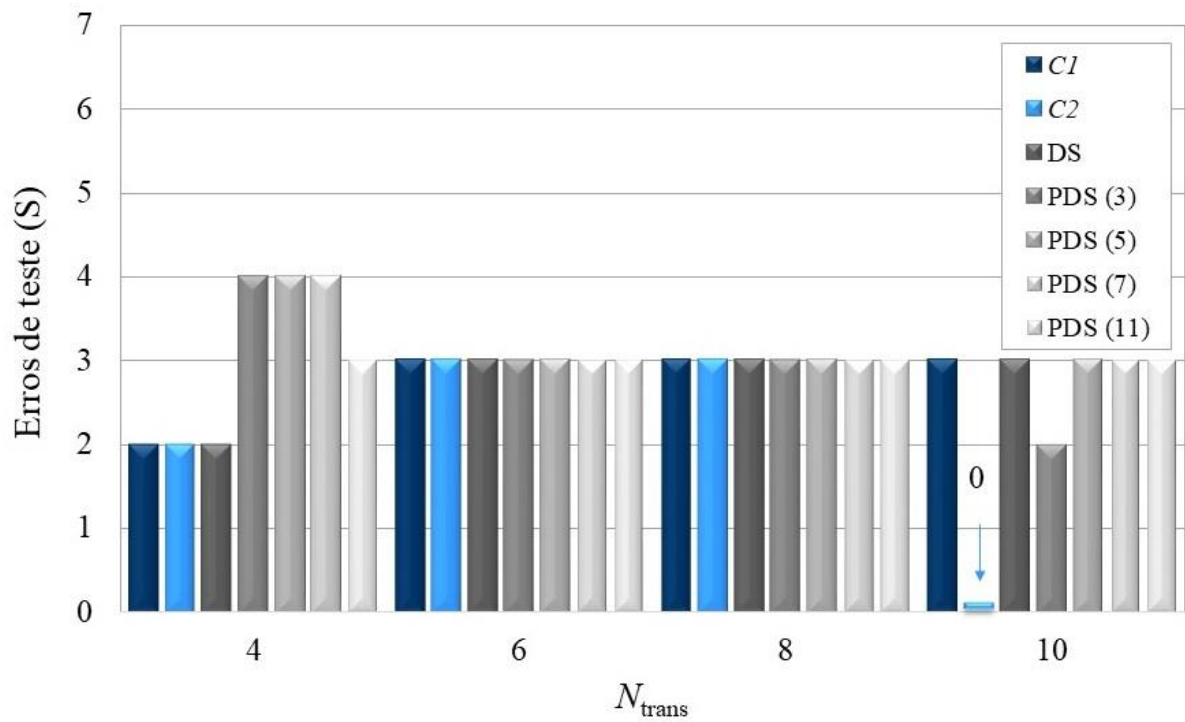


Figura 5.8 - Número de erros obtidos pelo SPA-LDA com os critérios *C1* e *C2* e SPA-LDA na forma usual após procedimentos de padronização, aplicados às amostras de teste (S) de EHC. A largura da janela empregada na padronização PDS é indicada entre parênteses.

Além dos erros de classificação, também foram utilizados parâmetros como sensibilidade, especificidade e acurácia para comparar o desempenho dos critérios propostos em relação aos métodos de padronização.

A **Tabela 5.6** apresenta os valores dos parâmetros de desempenho avaliados em termos do conjunto de teste (S). Para a padronização, apenas os melhores resultados da **Tabela 5.5** são mostrados, que correspondem à padronização DS para 4, 6 e 8 amostras de transferência e padronização PDS para $N_{trans} = 10$ usando janela de 3 pontos.

Tabela 5.6 - Valores de sensibilidade, especificidade e acurácia obtidos pelos modelos aplicados ao conjunto de teste (S).

Sensibilidade (%)				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
<i>C1</i>	100	100	100	100
<i>C2</i>	100	100	100	100
Padronização	90	85	85	90
Especificidade (%)				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
<i>C1</i>	90	85	85	85
<i>C2</i>	90	85	85	100
Padronização	100	100	100	100
Acurácia (%)				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
<i>C1</i>	95	92,5	92,5	92,5
<i>C2</i>	95	92,5	92,5	100
Padronização	95	92,5	92,5	95

Conforme apresentado, os critérios propostos apresentaram valores de sensibilidade, especificidade e acurácia compatíveis com aqueles obtidos pelo SPA-LDA usual após procedimentos de padronização, chegando a mostrar melhor desempenho quando 10 amostras de transferência foram utilizadas no critério *C2*.

5.1.5 Considerações finais

Os resultados obtidos neste estudo de caso demonstram que os critérios propostos, baseados na inclusão de amostras de transferência no cálculo do custo do SPA-LDA, reduziram substancialmente o número de erros obtidos na classificação das amostras de teste medidas no instrumento secundário. Além disso, observou-se também que a variável selecionada e o número de erros variaram de acordo com o número de amostras de transferência.

Considerando o número de erros obtidos para os conjuntos de treinamento e teste (P), foi verificado que a inclusão de amostras de transferência no cálculo do custo levou a um aumento desses valores. Para *C1*, isso pode ser atribuído à quantidade insuficiente de amostras de transferência para guiar o procedimento de seleção de variáveis. Para *C2*, a quantidade maior de erros é associada ao cálculo do custo baseado unicamente nas amostras de transferência.

Quando os critérios foram comparados aos métodos de padronização DS e PDS, foram obtidos resultados iguais e, em alguns casos, superiores aos métodos de padronização.

5.2 Determinação da extinção específica pela absorção na região do ultravioleta

A **Tabela 5.7** mostra os valores medidos e de referência para K_{232} , K_{270} e ΔK . Todas as amostras de AOEV não-adulteradas apresentaram valores de extinção específica nos comprimentos de onda de 232 e 270 nm (K_{232} e K_{270}) e variação específica de extinção (ΔK) igual ou inferior aos limites estabelecidos pelo Conselho Oleícola Internacional (IOOC: *International Olive Council*)^[99], reforçando os resultados obtidos pela pesquisa^[128].

Tabela 5.7 - Valores medidos e de referência para K_{232} , K_{270} e ΔK obtidas para as sete amostras de AOEV não-adulteradas. Os valores são obtidos após primeira (medida em 270 nm) e segunda (medida em 232 nm) diluição com ciclohexano. Cada lote é representado por um número.

Amostras (1ª diluição)	Valores de K_{270} , K_{232} e ΔK na região ultravioleta ($K_{1\text{cm}}^{1\%}$)			
	K_{270}		ΔK	
	medido	referência	medido	referência
L1	0,13		0,00	
L2	0,07		0,00	
L3	-0,02		0,00	
L4	0,05	$\leq 0,22$	0,01	$\leq 0,01$
L5	0,05		0,01	
L6	0,12		0,01	
L7	0,12		0,00	
Amostras (2ª diluição)	K_{232}		ΔK	
	medido	referência	medido	referência
	medido	referência	medido	referência
L1	0,25		-0,01	
L2	0,18		0,00	
L3	0,05		0,00	
L4	0,16	$\leq 2,50$	0,00	$\leq 0,01$
L5	0,14		0,00	
L6	0,17		0,00	
L7	0,18		0,00	

5.3 Espectrometria UV-Vis

5.3.1 Análise espectral

A **Figura 5.9a** apresenta os espectros de absorção molecular no UV-Vis das amostras de AOEV medidas nos dois instrumentos. Os espectros correspondem à região entre 364 e 706 nm e como poder ser observado, existe uma variação sistemática da linha de base ao longo da faixa de trabalho. Para corrigir esse problema, foi utilizado o pré-processamento *baseline offset* e os espectros resultantes são apresentados na **Figura 5.9b**.

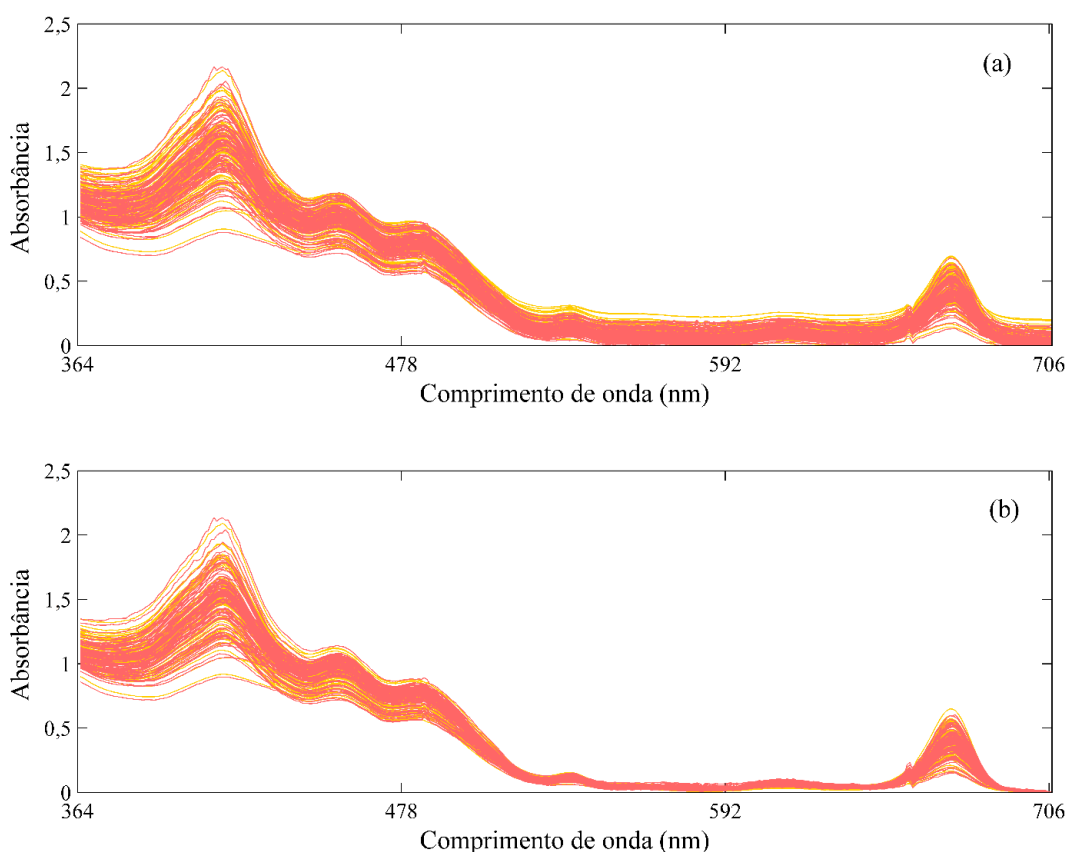


Figura 5.9 - Espectros de absorção molecular no UV-Vis das amostras de AOEV antes (a) e após pré-processamento (b). Espectros adquiridos nos instrumentos A (—) e B (—).

Comparando os espectros coletados nos dois instrumentos na **Figura 5.9b**, verifica-se, a olho nu, um perfil muito semelhante. Em ambos os casos, é possível observar três picos claramente definidos (em aproximadamente 410, 450 e 480 nm) correspondentes à absorção de carotenoides na região azul do espectro ^[133]. Essas absorções estão associadas a transições eletrônicas do estado fundamental S_0 para o estado S_2 , com energia do estado S_2 dependendo da extensão da conjugação dos π -elétrons dos carotenoides ^[134]. O pico a cerca de 670 nm, próximo à região do vermelho, corresponde à absorção da clorofila e envolve as transições eletrônicas entre os estados S_0 e S_1 ^[133].

A **Figura 5.10** traz os espectros medidos nos instrumentos A (**Figura 5.10a**) e B (**Figura 5.10b**) com discriminação das classes de AOEV analisadas. Em ambos os casos, observa-se que as amostras das duas classes apresentam os mesmos picos de absorção mencionados anteriormente, com diferença na absorbância registrada, que é maior para as amostras não-adulteradas e vai diminuindo à medida que a quantidade de óleo de soja na mistura aumenta. Esse aumento pode estar provocando uma diminuição da concentração dos carotenoides e clorofila na mistura, resultando em menores valores de absorbância.

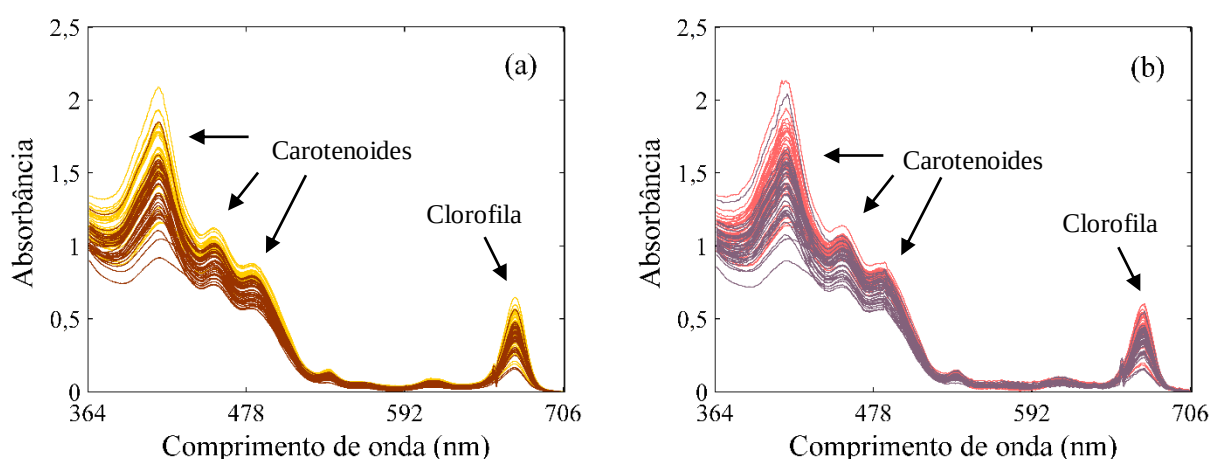


Figura 5.10 - Espectros UV-Vis pré-processados registrados nos instrumentos A (a) e B (b) com indicação das duas classes de AOEV. Amostras não-adulteradas (—) e adulteradas (—) no instrumento A; amostras não-adulteradas (—) e adulteradas (—) no instrumento B.

5.3.2 Análise exploratória

A **Figura 5.11** apresenta os gráficos dos escores obtidos por PC 1 $\times$ PC 2, PC 1 $\times$ PC 3 e PC 1 $\times$ PC 4 para as amostras do conjunto de teste medidas nos dois instrumentos.

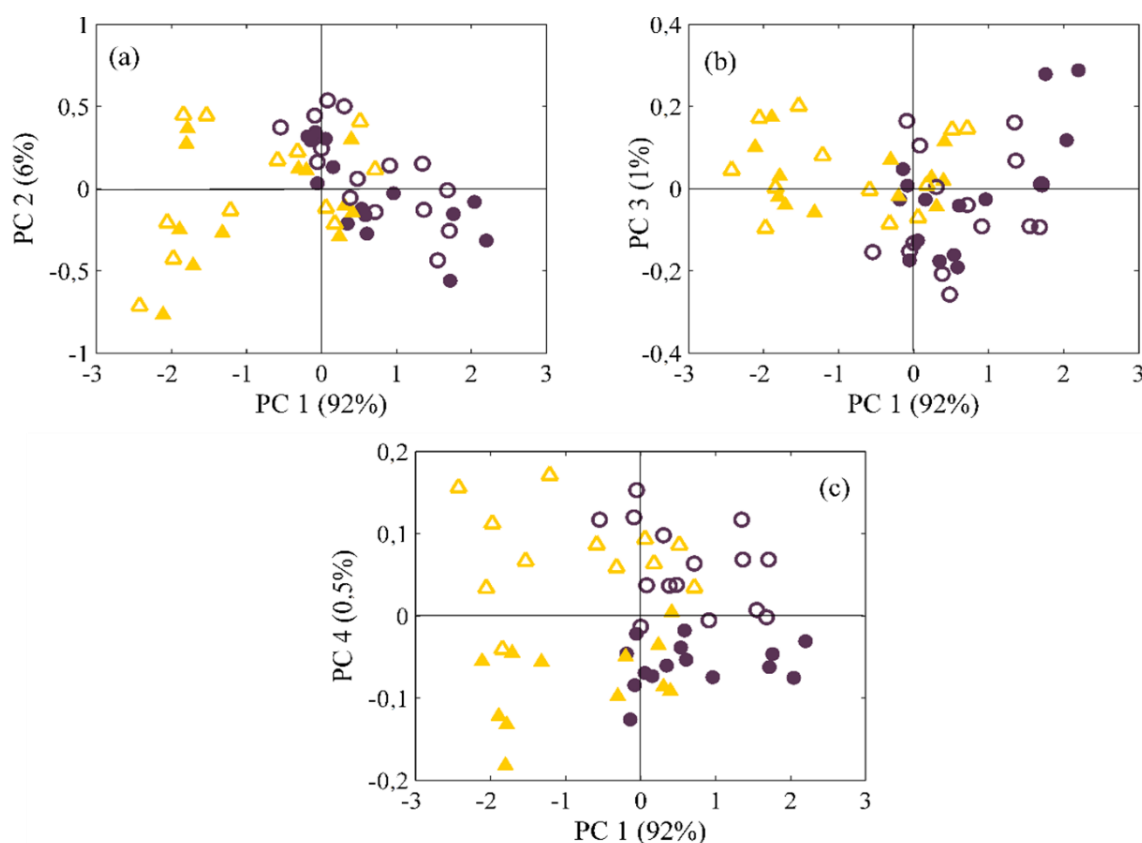


Figura 5.11 - Gráficos dos escores nas quatro primeiras PCs para as amostras de AOEV do conjunto de teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento A; Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento B.

Os gráficos dos escores em PC 1 $\times$ PC 2 (**Figura 5.11a**) e PC 1 $\times$ PC 3 (**Figura 5.11b**) apontam uma tendência de discriminação entre as classes de AOEV adulterados e não-adulterados ao longo da PC 1, que apresentou 92% da variância explicada dos dados. Em ambos os casos, as amostras não-adulteradas, representadas pelos círculos, aparecem com valores positivos de escores.

Embora a PC 4 explique apenas 0,5% da variância dos dados (**Figura 5.11c**), é possível observar que nessa PC ocorre uma separação das amostras em função do instrumento com o qual foram medidas, o que indica que mesmo com a semelhança entre os espectros dos dois instrumentos, existe alguma diferença no conjunto de dados.

A **Figura 5.12** mostra o gráfico dos *loadings* para as quatro primeiras PCs. É possível observar que para a PC 1, responsável por discriminar as duas classes de objetos com 92% da variância explicada, as variáveis com maior peso se encontram nas regiões de absorção dos carotenoides (principalmente no primeiro pico em aproximadamente 410 nm) e da clorofila (aproximadamente 670 nm). Para a PC 4, que separa as amostras quanto aos instrumentos de medida, apesar do sinal ruidoso é possível verificar que as variáveis com maior peso estão em aproximadamente 480 nm (terceiro pico de absorção dos carotenoides), num ponto próximo ao pico de absorção da clorofila e na região em torno de 592 nm, onde as duas classes de azeite se sobrepõem.

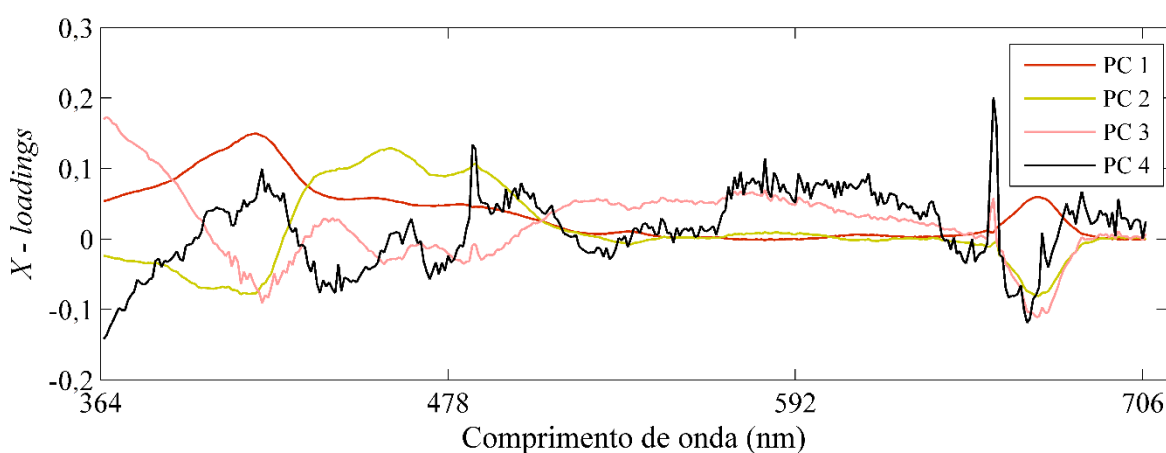


Figura 5.12 - Gráfico dos *loadings* para as quatro primeiras PCs.

Uma vez que as PCs 2 e 3 não realizam separação significativa, elas não foram discutidas no gráfico dos *loadings*.

5.3.3 Modelos SPA-LDA

As **Tabelas 5.8** e **5.9** apresentam os resultados de classificação obtidos pelos modelos SPA-LDA na sua forma usual construídos para os dois instrumentos.

Tabela 5.8 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento A. O número de variáveis empregadas no modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Instrumento A (5)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	34	32	2	94 ^b
Adulterada	28	4	24	86 ^c
Conjunto de Teste - Instrumento A (5)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	15	15	-	100 ^b
Adulterada	12	4	8	67 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Tabela 5.9 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento B. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Instrumento B (3)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	34	33	1	97 ^b
Adulterada	28	9	19	68 ^c
Conjunto de Teste - Instrumento B (3)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	15	15	-	100 ^b
Adulterada	12	3	9	75 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Como pode ser observado, os modelos construídos com os dados obtidos em cada instrumento apresentaram desempenhos diferentes nos conjuntos de treinamento e teste. Dessa forma, a escolha do instrumento primário foi guiada com base na quantidade de erros obtidos para o conjunto de treinamento que por sua vez, foi menor para o instrumento A (Brasil). Embora a acurácia para o conjunto de teste brasileiro tenha sido menor, a diferença entre os modelos dos dois países é de apenas um erro de classificação, o que não impede a escolha do instrumento brasileiro como primário.

Nas **Tabelas 5.8 e 5.9**, verifica-se também que os erros mais frequentes estão associados às amostras adulteradas sendo classificadas como não-adulteradas. Para ambos os modelos, esse comportamento foi observado principalmente para as amostras adulteradas contendo 1% e 5% de óleo de soja.

A **Tabela 5.10** apresenta os resultados obtidos com a utilização do modelo com o instrumento primário (instrumento A do Brasil) na previsão das amostras de teste medidas no secundário (instrumento B da Argentina).

Tabela 5.10 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com base no instrumento primário aplicado as amostras de teste medidas no instrumento secundário. *N* indica o número de amostras de teste em cada classe de AOEV.

Conjunto de teste secundário				
Classe	<i>N</i>	Classe Prevista		Desempenho (%)
Verdadeira		Não-adulterada	Adulterada	
Não-adulterada	15	14	1	93 ^b
Adulterada	12	11	1	8,3 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Como mostrado na tabela, mais de 50% das amostras foram incorretamente classificadas quando o modelo primário foi aplicado às amostras de teste (S), resultado que se contrapõe à semelhança espectral e aponta a necessidade de transferência dos modelos.

5.3.4 Transferência dos modelos SPA-LDA

Na **Tabela 5.11** são apresentados os erros para os conjuntos de treinamento e teste (P) obtidos modelos SPA-LDA, construídos para o instrumento primário, na sua forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} .

Tabela 5.11 – Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário.

Parâmetros	<i>N</i>	Resultados de classificação dos modelos SPA-LDA					
		Usual	Critérios	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
Erros de treinamento	62	6	<i>C1</i>	10	9	10	10
			<i>C2</i>	22	9	10	10
Erros de teste (P)	27	4	<i>C1</i>	3	3	3	3
			<i>C2</i>	15	5	4	3

Para os dois critérios, os erros de treinamento foram maiores quando comparados ao SPA-LDA usual, fato que pode estar relacionado à participação das amostras de transferência na seleção do melhor subconjunto de variáveis. Os erros de teste (P) para *C1* foram ligeiramente inferiores enquanto para *C2* variaram aleatoriamente.

A **Figura 5.13** apresenta os erros obtidos após aplicação dos modelos ao conjunto de amostras de teste (S). Os resultados para *C1* e *C2* são apresentados para diferentes valores de N_{trans} . O modelo SPA-LDA na forma usual corresponde à $N_{\text{trans}} = 0$ e as barras indicam o número de erros.

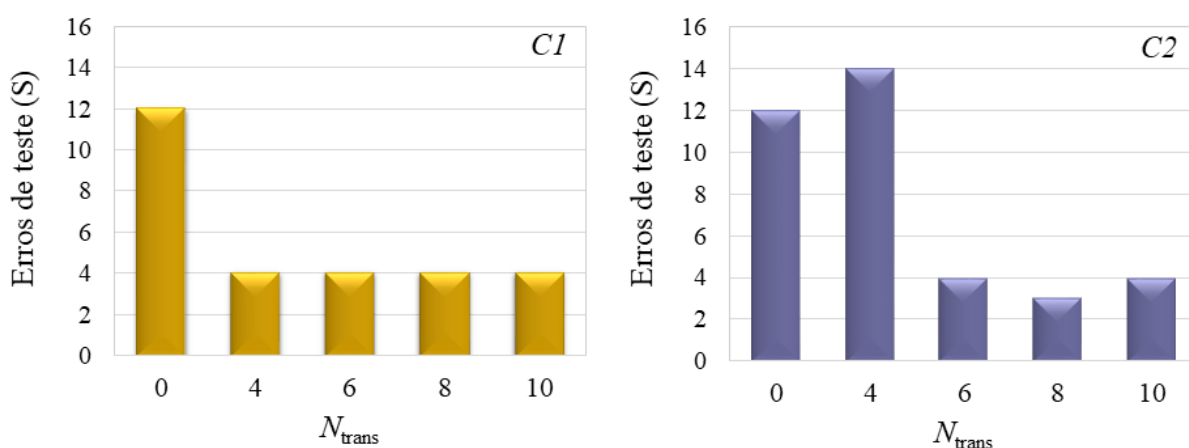


Figura 5.13 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de AOEV analisadas por espectrometria UV-Vis. O modelo SPA-LDA na forma usual corresponde à $N_{\text{trans}} = 0$ e as barras indicam o número de erros.

Novamente, a aplicação dos critérios *C1* e *C2* aos dados do instrumento secundário diminuiu consideravelmente o número de erros. De fato, se forem utilizadas seis ou mais amostras de transferência, ambos os critérios conduzem a erros de teste (S) compatíveis com os erros de teste (P).

Os tipos de erros de classificação encontrados para cada critério são mostrados na **Tabela 5.12**. Para o critério *C1*, a maioria das amostras adulteradas que foram incorretamente classificadas como não-adulteradas apresentam nível de adulteração de 1%. No caso do critério *C2*, os erros oscilam entre as amostras com 1% e 5% de adulteração. A exceção foi encontrada para o critério *C2* com $N_{\text{trans}} = 4$, onde para a classe adulterada apenas uma amostra com 15% de adulteração foi corretamente classificada.

Tabela 5.12 - Erros de classificação obtidos pelos critérios *C1* e *C2* aplicados às amostras de teste (S) medidas no instrumento secundário. *N* indica o número de amostras de teste em cada classe de AOEV.

Critério <i>C1</i>									
Classe Verdadeira	<i>N</i>	<i>N_{trans} 4</i>		<i>N_{trans} 6</i>		<i>N_{trans} 8</i>		<i>N_{trans} 10</i>	
		Classe Prevista		Classe Prevista		Classe Prevista		Classe Prevista	
		NA	A	NA	A	NA	A	NA	A
NA	15	13	2	14	1	13	2	13	2
A	12	2	10	3	9	2	10	2	10

Critério <i>C2</i>									
Classe Verdadeira	<i>N</i>	<i>N_{trans} 4</i>		<i>N_{trans} 6</i>		<i>N_{trans} 8</i>		<i>N_{trans} 10</i>	
		Classe Prevista		Classe Prevista		Classe Prevista		Classe Prevista	
		NA	A	NA	A	NA	A	NA	A
NA	15	12	3	15	-	15	-	13	2
A	12	11	1	4	8	3	9	2	10

NA: Amostras não-adulteradas

A: Amostras adulteradas

A **Figura 5.14** mostra os subconjuntos de variáveis selecionadas pelo algoritmo SPA nos modelos de classificação.

Na maioria dos casos, variáveis foram selecionadas nas proximidades das bandas de absorção dos carotenoides e clorofila em todos os modelos. A exceção foi obtida para o critério *C2* com $N_{trans} = 4$ amostras, onde apenas uma variável foi selecionada em uma região onde os espectros das duas classes estão sobrepostos (aproximadamente 592 nm).

Como consequência, o número resultante de erros de teste (S) foi ainda maior comparado com o modelo SPA-LDA na forma usual, conforme apresentado na **Figura 5.13**. Neste caso particular, o resultado insatisfatório pode ser atribuído ao uso de um número insuficiente de amostras para guiar o processo de seleção de variáveis. De fato, o critério *C2* utiliza apenas as amostras de transferência na avaliação dos subconjuntos candidatos de variáveis, enquanto que o critério *C1* usa tanto as amostras de transferência quanto as de treinamento.

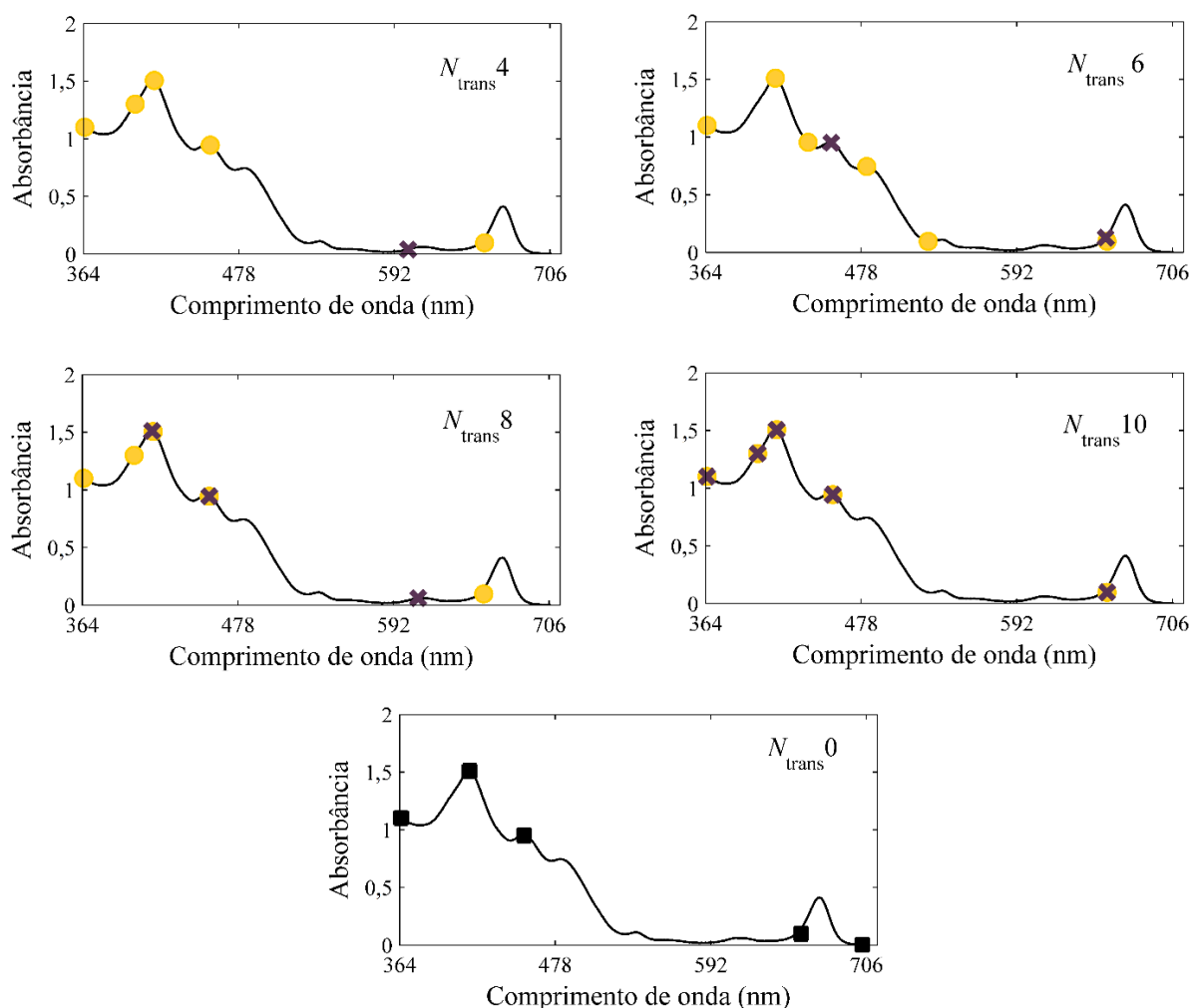


Figura 5.14 - Espectros UV-Vis médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios C1: ●, C2: ✕ e SPA na forma usual: ■.

Na **Figura 5.15a-c** são apresentados os gráficos dos escores da função discriminante 1 (FD 1) obtidos para os três modelos, aplicados ao conjunto de teste (S). As **Figuras 5.15b** e **5.15c** correspondem aos critérios C1 e C2 com $N_{trans} = 8$.

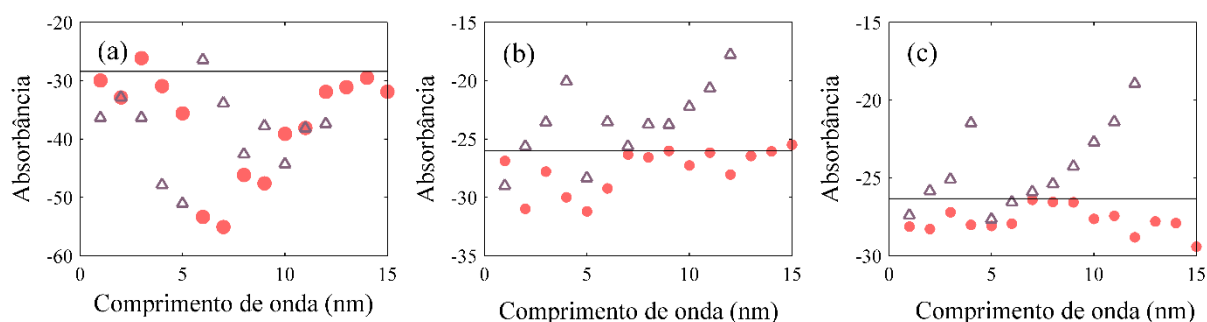


Figura 5.15 - Gráficos dos escores da FD 1 obtidos para os modelos SPA-LDA usual (a), critério C1 (b) e critério C2 (c). Amostras não-adulteradas (●) e adulteradas (▲) medidas no instrumento secundário.

O número de erros obtidos pelo modelo SPA-LDA na forma usual após padronização DS e PDS do conjunto de teste (S) é apresentado na **Tabela 5.13**. A padronização reduziu significativamente os erros de classificação, sobretudo quando 8 (para DS e PDS usando janela de 11 pontos) e 10 (para DS) amostras de transferência são utilizadas.

Tabela 5.13 - Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização DS e PDS. A largura da janela utilizada no modelo PDS é apresentada entre parênteses.

Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	7	5	4	4
PDS (3)	7	8	7	6
PDS (5)	6	7	7	7
PDS (7)	5	5	6	7
PDS (11)	7	7	4	6

Os resultados da transferência realizada por meio dos critérios propostos são comparados aos resultados alcançados pelos métodos de padronização na **Figura 5.16**.

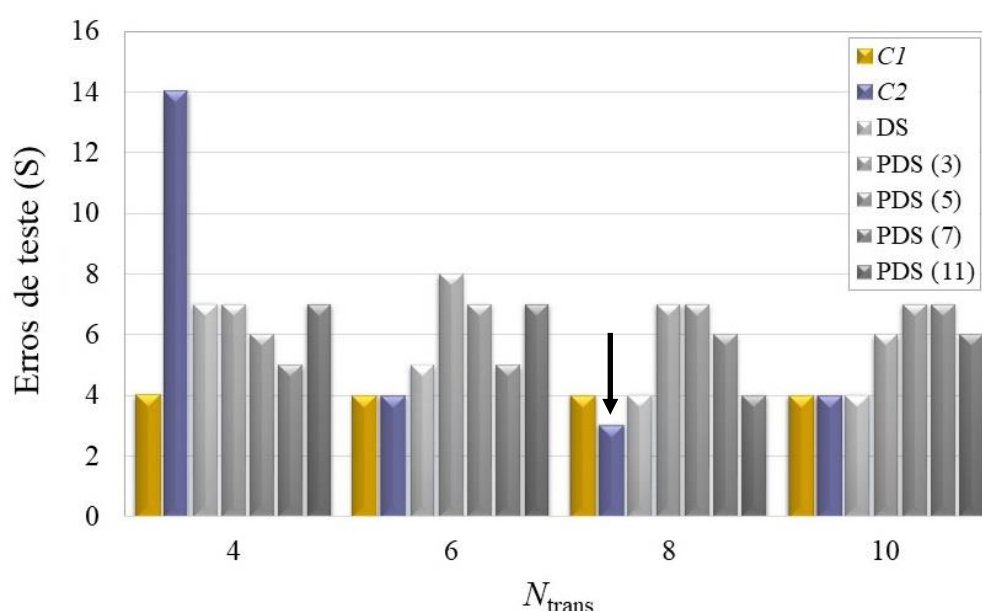


Figura 5.16 - Número de erros obtidos pelo SPA-LDA com os critérios C1 e C2 e SPA-LDA na forma usual após procedimentos de padronização, aplicados às amostras de teste (S). A largura da janela empregada na padronização PDS é indicada entre parênteses.

Com exceção do critério C2 para $N_{trans} = 4$, os resultados dos critérios foram compatíveis ou superiores aos resultados obtidos pelo SPA-LDA após a padronização. Isto é particularmente verdade para $N_{trans} = 8$.

A **Tabela 5.14** apresenta os valores de sensibilidade, especificidade e acurácia, avaliados em termos do conjunto de teste (S) para comparação do desempenho dos critérios propostos em relação aos métodos de padronização. Para a padronização, apenas os melhores resultados da **Tabela 5.13** são mostrados, que correspondem a padronização DS empregando 6, 8 e 10 amostras de transferência e padronização PDS empregando $N_{\text{trans}} = 4$ com janela de 7 pontos.

Tabela 5.14 - Valores de sensibilidade, especificidade e acurácia obtidos pelos modelos aplicados ao conjunto de teste (S).

Sensibilidade (%)				
Método	$N_{\text{trans}} 4$	$N_{\text{trans}} 6$	$N_{\text{trans}} 8$	$N_{\text{trans}} 10$
<i>C1</i>	83	75	83	83
<i>C2</i>	8,3	67	75	83
Padronização	75	67	75	67
Especificidade (%)				
Método	$N_{\text{trans}} 4$	$N_{\text{trans}} 6$	$N_{\text{trans}} 8$	$N_{\text{trans}} 10$
<i>C1</i>	87	93	87	83
<i>C2</i>	80	100	100	87
Padronização	87	93	93	100
Acurácia (%)				
Método	$N_{\text{trans}} 4$	$N_{\text{trans}} 6$	$N_{\text{trans}} 8$	$N_{\text{trans}} 10$
<i>C1</i>	85	85	85	85
<i>C2</i>	48	85	89	85
Padronização	81	81	85	85

Conforme apresentado, os critérios propostos apresentaram valores para os parâmetros compatíveis com aqueles obtidos pelo SPA-LDA usual após procedimentos de padronização, chegando a mostrar melhor desempenho quando 8 amostras de transferência foram utilizadas.

5.3.5 Considerações finais

Para a espectrometria UV-Vis, embora os dados registrados nos dois instrumentos não tenham revelado diferença significativa no perfil espectral, um exame cuidadoso detectou pequenas variações que foram constatadas quando o modelo primário foi aplicado ao conjunto de amostras medidas no instrumento secundário. Mais uma vez, os critérios foram aplicados como método de transferência e os resultados foram favoráveis, de modo que os erros de classificação das amostras do conjunto de teste secundário foram significativamente reduzidos.

As variáveis selecionadas e o número de erros variaram com o número de amostras de transferência.

Nesse estudo foi detectado um resultado insatisfatório fornecido pelo critério C2 quando foram utilizadas quatro amostras de transferência, o que pode ser atribuído à utilização de um número insuficiente de amostras para guiar o processo de seleção de variáveis.

Quando comparados aos métodos de padronização direta (DS) e padronização direta em etapas (PDS), os critérios mostraram resultados similares ou superiores aos resultados obtidos pelo SPA-LDA após a padronização. Isto é particularmente verdade para $N_{\text{trans}} = 8$.

5.4 Espectrometria de fluorescência molecular

5.4.1 Análise espectral

A **Figura 5.17** apresenta os espectros de emissão de fluorescência das amostras de AOEV na região entre 350 e 600 nm, coletados nos instrumentos A (**Figura 5.17a**) e B (**Figura 5.17b**). Comparando os dois espectros, podemos observar uma grande diferença na qualidade do sinal, que para o instrumento B, é ruidoso e apresenta baixa relação sinal/ruído. Ainda é possível observar que os espectros das amostras adulteradas apresentaram maior intensidade de fluorescência nos dois instrumentos.

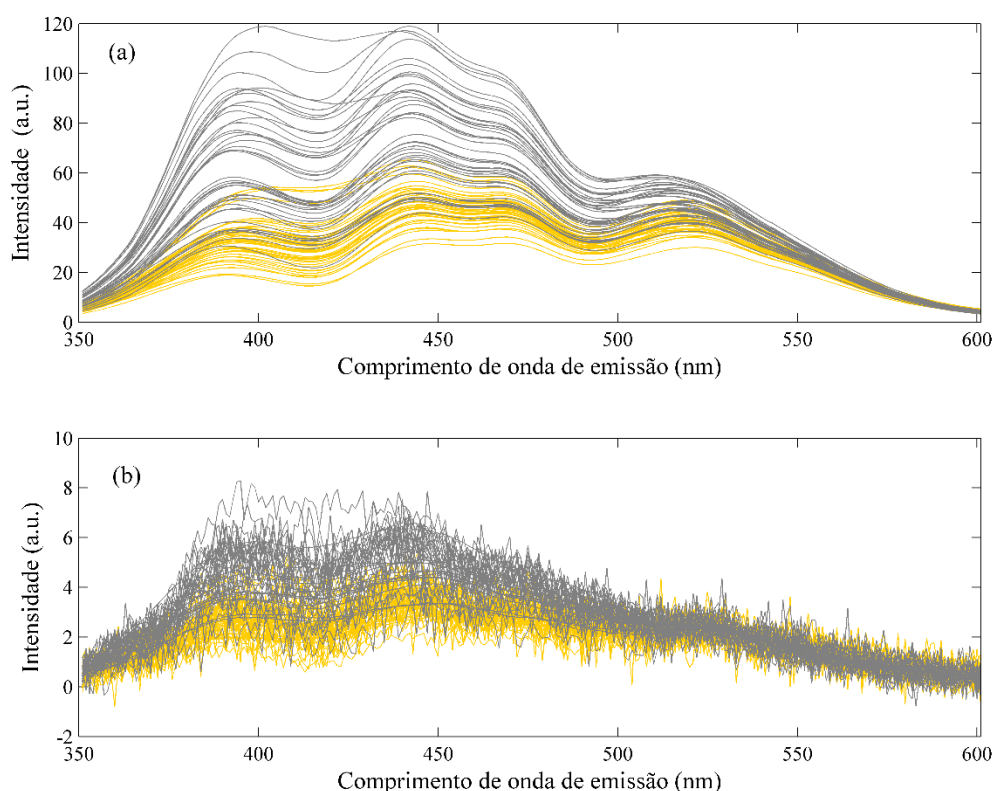


Figura 5.17 - Espectros de emissão de fluorescência molecular das amostras de AOEV medidas nos instrumentos A (a) e B (b). Amostras não-adulteradas (—) e adulteradas (—).

Na faixa espectral investigada, a fluorescência dos AOEVs tem sido atribuída principalmente aos produtos de oxidação de ácidos graxos (dois picos lisos a 445 e 475 nm) e a vitamina E (525 nm) ^[135]. Estudos indicam que os principais fluoróforos encontrados nos AOEVs também estão presentes na composição de óleos vegetais ^[136]. Contudo, os óleos vegetais apresentam maiores concentrações de produtos de oxidação de ácidos graxos, provavelmente em decorrência do processo de refino, resultando em intensidades de fluorescência mais elevadas entre 400 e 500 nm ^[137]. Isso explica o aumento da intensidade de

fluorescência nessa região para as amostras adulteradas, considerando ainda que a intensidade será maior quanto maior a quantidade de óleo de soja adicionada na mistura.

5.4.2 Análise exploratória

A **Figura 5.18** mostra o gráfico dos escores obtidos por PC 1 $\times$ PC 2 para as amostras de teste medidas de AOEV em ambos os instrumentos.

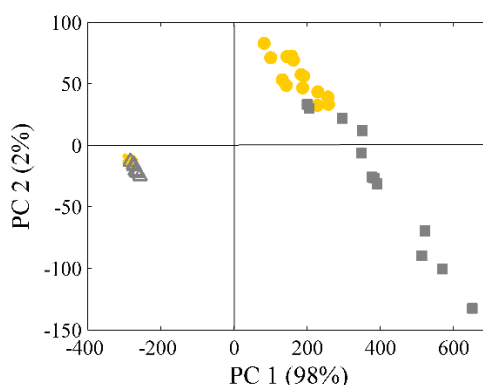


Figura 5.18 - Gráficos dos escores em PC 1 $\times$ PC 2 para as amostras de AOEV do conjunto teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (■) medidas no instrumento A; Amostras não-adulteradas (✕) e adulteradas (▲) medidas no instrumento B.

Ao longo da PC 1, que porta 98% da variância explicada, ocorre a separação dos grupos de AOEV em função do instrumento de medida. Na PC 2, as classes não-adulterada e adulterada analisadas no instrumento A tendem a se separar. Neste caso, verifica-se uma dispersão das amostras adulteradas, provavelmente devido à diferença nos níveis de adulteração. No instrumento B as duas classes de amostras se sobrepõem.

A **Figura 5.19** mostra o gráfico dos *loadings* usados para verificar a contribuição das variáveis para as duas PCs.

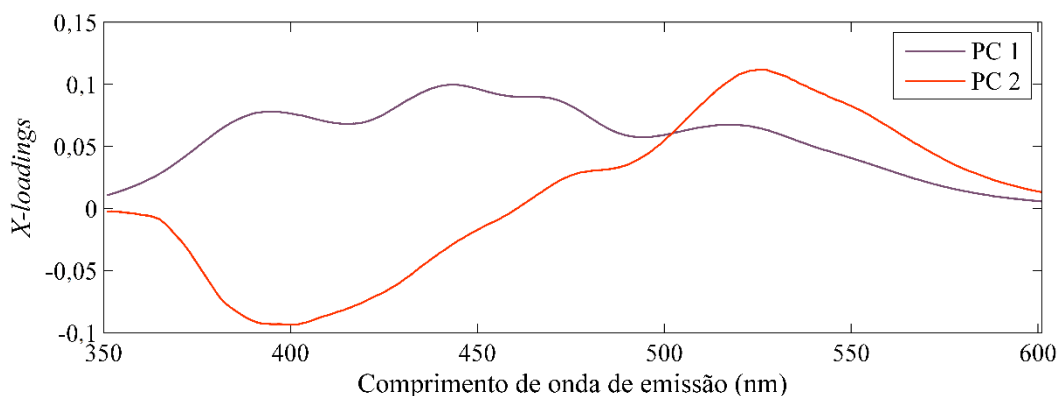


Figura 5.19 - Gráfico dos *loadings* para as duas primeiras PCs.

Para a PC 1, responsável por separar grupos em função dos instrumentos de medida, todas as variáveis entre 400 e 525 nm contribuem. No caso da PC 2, que separa as duas classes de AOEV analisadas no instrumento A, o maior peso é atribuído a variável em torno de 525 nm, onde acontece a fluorescência da vitamina E.

5.4.3 Modelos SPA-LDA

As **Tabelas 5.15** e **5.16** apresentam os resultados de classificação obtidos pelos modelos SPA-LDA na sua forma usual construídos para os dois instrumentos.

Tabela 5.15 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento A. O número de variáveis empregadas no modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Instrumento A (8)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	34	34	-	100 ^b
Adulterada	28	4	24	86 ^c

Conjunto de Teste - Instrumento A (8)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	14	-	100 ^b
Adulterada	12	3	9	75 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Tabela 5.16 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com os dados registrados no instrumento B. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Instrumento B (3)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	34	32	2	94 ^b
Adulterada	28	6	22	79 ^c

Conjunto de Teste - Instrumento B (3)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	14	-	100 ^b
Adulterada	12	8	4	33 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Como pode ser observado, os melhores resultados de treinamento e teste foram obtidos para o instrumento A. A diferença é ainda mais desfavorável para o conjunto de teste que, no instrumento B, teve a maioria de suas amostras adulteradas incorretamente classificadas.

Novamente, foi verificado que os erros mais frequentes estão associados às amostras adulteradas sendo classificadas como não-adulteradas. Para o modelo baseado no instrumento A, tanto no conjunto de treinamento como no de teste, esse comportamento foi observado principalmente para as amostras contendo 1% e 5% de óleo de soja. Os modelos construídos com base no instrumento B apresentaram erros que variaram para diferentes níveis de adulteração nos dois subconjuntos.

Considerando a qualidade do sinal espectral e os resultados dos modelos, o instrumento A foi considerado como primário. A **Tabela 5.17** apresenta os resultados da aplicação do modelo construído com base no instrumento primário na previsão das amostras de teste medidas no instrumento secundário.

Tabela 5.17 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído para o instrumento primário (A) aplicado às amostras de teste medida no instrumento secundário. *N* indica o número de amostras de teste em cada classe de AOEV.

Conjunto de teste secundário				
Classe Verdadeira	<i>N</i>	Classe Prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	14	-	100 ^b
Adulterada	12	12	-	0 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Neste caso, todas as amostras adulteradas foram incorretamente classificadas quando o modelo primário foi aplicado na previsão de amostras medidas no instrumento secundário, resultado que corrobora com a significativa diferença espectral entre os dois instrumentos.

5.4.4 Transferência dos modelos SPA-LDA

Na **Tabela 5.18** são apresentados os erros para os conjuntos de treinamento e teste (P) obtidos modelos SPA-LDA na sua forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . Neste caso, os erros de classificação do conjunto de treinamento obtidos por *C1* se equiparam aos do SPA-LDA usual, enquanto para *C2*, menores valores foram encontrados. No conjunto de teste (P), com exceção de *C2* para $N_{trans} = 4$, todos os resultados foram iguais.

Tabela 5.18 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário.

Parâmetros	N	Resultados de classificação dos modelos SPA-LDA					
		Usual	Crítérios	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
Erros de treinamento	62	4	<i>C1</i>	4	4	4	4
			<i>C2</i>	0	1	2	2
Erros de teste (P)	26	3	<i>C1</i>	3	3	3	3
			<i>C2</i>	5	3	3	3

Os modelos construídos com base nas medidas realizadas no instrumento primário foram aplicados aos espectros das amostras de teste medidas no instrumento secundário antes e após suavização dessas amostras pelo filtro Savitzky-Golay, com polinômio de segunda ordem e janela de 25 pontos. A **Figura 5.20** compara o perfil espectral em cada caso.

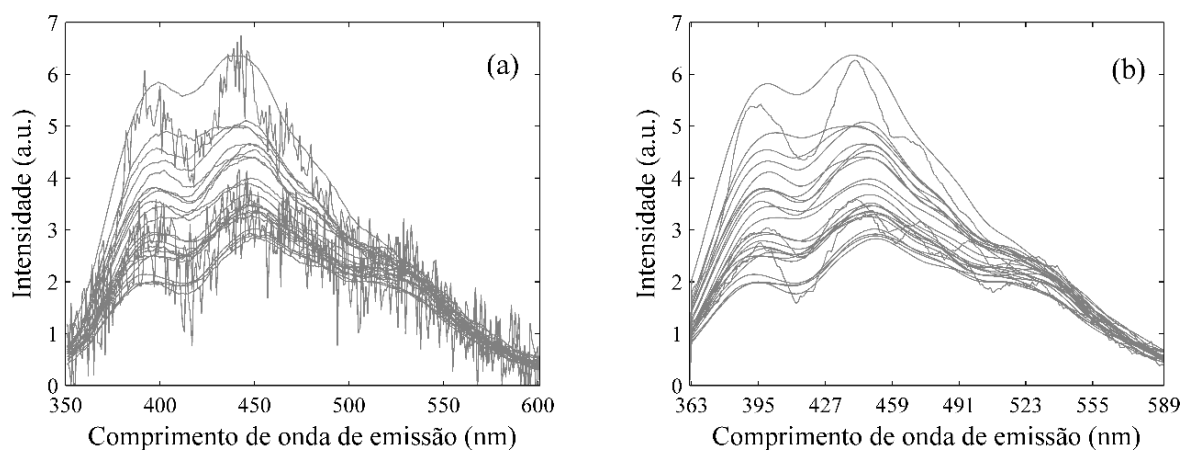


Figura 5.20 - Espectros de emissão de fluorescência molecular das amostras de AOEV do conjunto de teste medido no instrumento secundário antes (a) e após (b) suavização.

Como pode ser observado, os ruídos diminuem significativamente com a suavização. A **Figura 5.21** apresenta os erros obtidos pelos três modelos (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) antes (**Figura 5.21a e 5.21b**) e após (**Figura 5.21c e 5.21d**) a suavização.

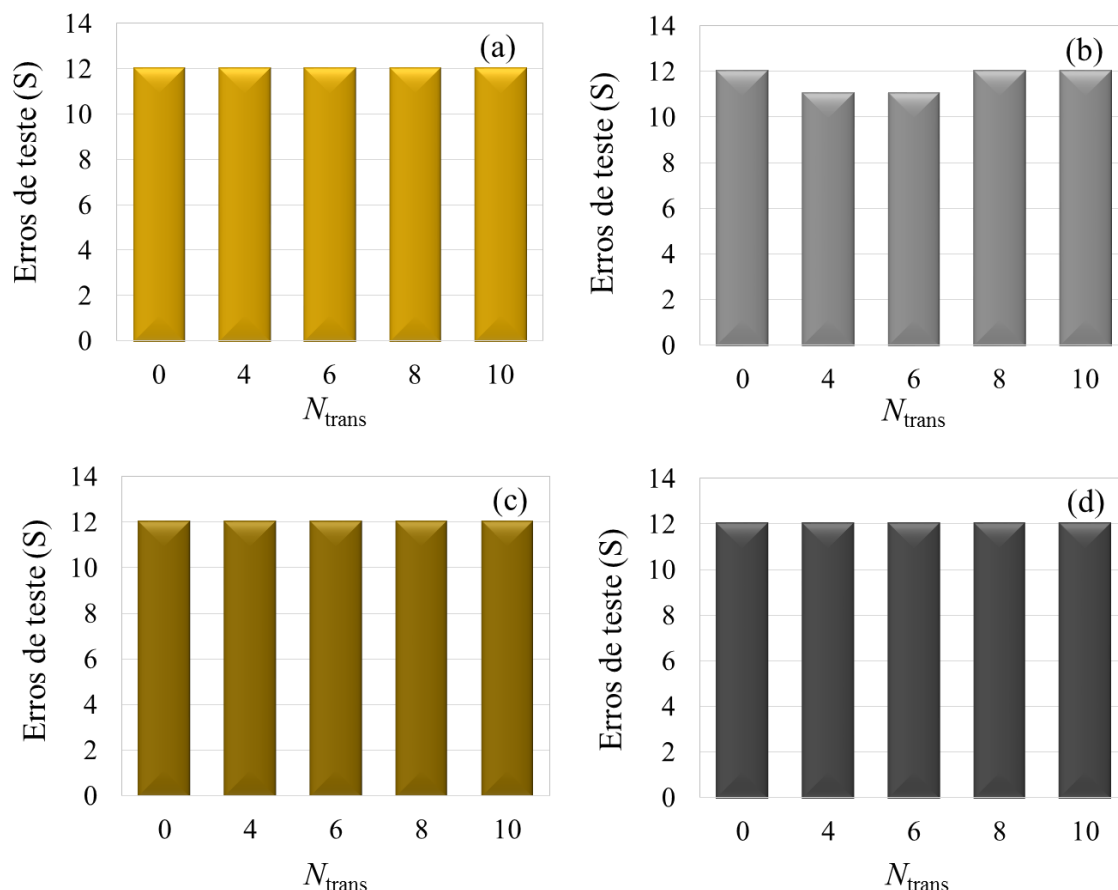


Figura 5.21 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios $C1$ e $C2$) aplicados às amostras de teste (S) de AOEV analisadas por espectrometria de fluorescência molecular. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.

Os resultados para $C1$ e $C2$ são apresentados para diferentes valores de N_{trans} . O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam os erros. Neste caso, as diferenças espectrais não permitiram que os erros de teste fossem significativamente reduzidos.

Antes do pré-processamento, usando o critério $C1$ todas as amostras adulteradas continuaram a ser incorretamente classificadas como não-adulteradas. Com o critério $C2$, mesmo com toda a diferença espectral, quando seis ou quatro amostras de transferência foram empregadas, os erros diminuíram em uma unidade. Neste caso, para $N_{trans} = 4$, onze das quatorze amostras não-adulteradas foram classificadas como adulteradas, enquanto para $N_{trans} = 6$, onze das doze amostras adulteradas foram classificadas como não-adulteradas. Após a suavização, todas as amostras adulteradas continuaram a ser incorretamente classificadas como não-adulteradas pelos dois critérios.

A **Figura 5.22** mostra os subconjuntos de variáveis selecionadas pelo algoritmo SPA nos modelos de classificação. É possível observar que, para que os erros de classificação fossem

reduzidos em pelo menos uma unidade, uma quantidade elevada de variáveis foi utilizada, como apontam as variáveis selecionadas para $N_{\text{trans}} = 4$ e $N_{\text{trans}} = 6$ usando o critério C2.

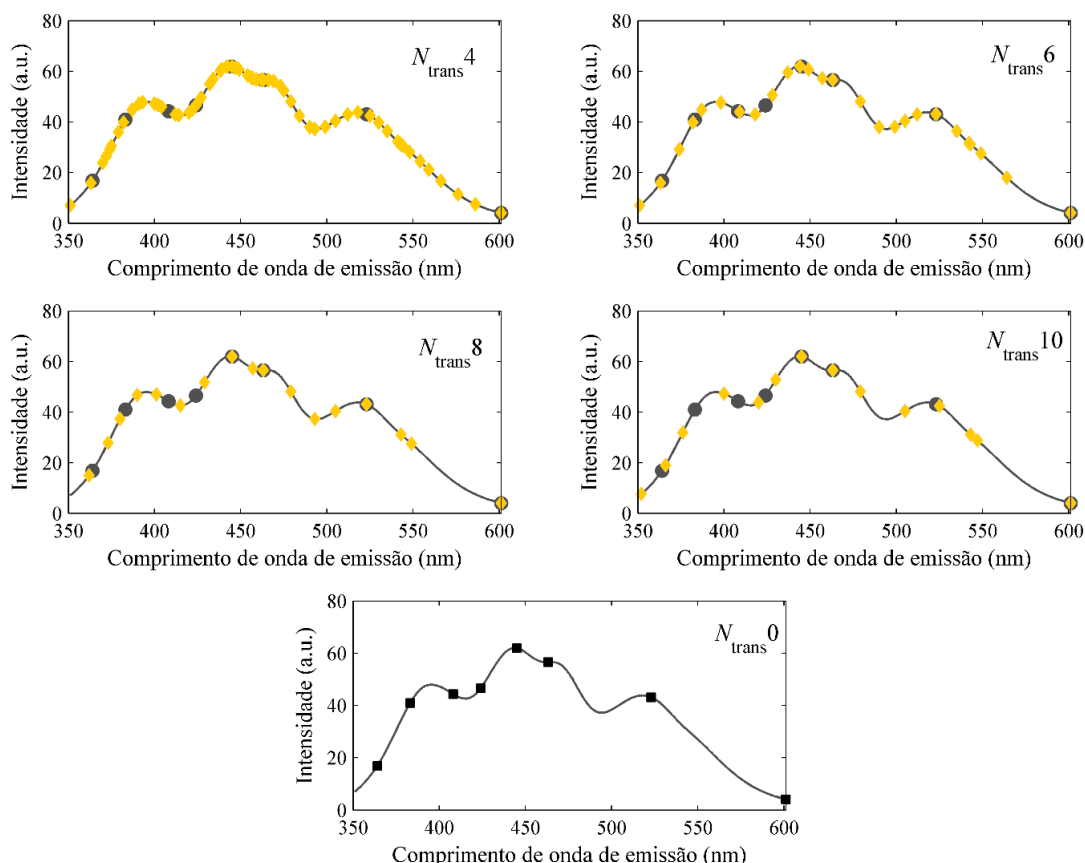


Figura 5.22 - Espectros médios de fluorescência molecular do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA em modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios C1: ●, C2: ◆ e SPA na forma usual: ■.

Como os espectros do instrumento secundário são muito diferentes dos espectros do instrumento primário, espera-se que os procedimentos de padronização forneçam melhores resultados de classificação. Os resultados da aplicação de diferentes estratégias de padronização para os dados de fluorescência estão detalhados em Milanez *et al.* (2017) ^[138].

5.4.5 Transferência de classificação usando métodos de padronização

Os métodos de padronização DS e PDS também foram aplicados aos espectros sem pré-processamento e pré-processados das amostras de teste medidas no instrumento secundário.

Para o instrumento primário, além do modelo SPA-LDA na forma usual (resultados na **Tabela 5.15**), também foi construído o modelo baseado no PLS-DA. A **Tabela 5.19** detalha os resultados encontrados. Comparando os dois modelos, observa-se que a diferença está em mais uma amostra de treinamento adulterada incorretamente classificada pelo PLS-DA.

Tabela 5.19 - Resultados de classificação obtidos com pelo modelo PLS-DA construído para os dados registrados no instrumento primário A. O número de variáveis latentes empregadas no modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Instrumento A (4)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	34	34	-	100 ^b
Adulterada	28	5	23	82 ^c
Conjunto de Teste - Instrumento A (4)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	14	-	100 ^b
Adulterada	12	3	9	75 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

A **Tabela 5.20** mostra os erros de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. Assim como aconteceu com o SPA-LDA, todas as amostras adulteradas foram incorretamente classificadas.

Tabela 5.20 - Resultados de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. *N* indica o número de amostras de teste em cada classe de AOEV.

Conjunto de teste secundário				
Classe Verdadeira	<i>N</i>	Classe Prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	14	-	100 ^b
Adulterada	12	12	-	0 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

O número de erros obtidos pelos modelos SPA-LDA e PLS-DA após padronização DS e PDS do conjunto de teste (S) é apresentado nas **Tabelas 5.21** e **5.22**, respectivamente, para os espectros sem e com pré-processamento das amostras de teste (S).

Tabela 5.21 - Erros de classificação obtidos pelo modelo SPA-LDA aplicado às amostras de teste (S) após padronização dos dados sem e com pré-processamento. A largura da janela empregada na padronização PDS é indicada entre parênteses.

SPA-LDA - Conjunto de teste (S) sem pré-processamento				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	10	6	5	5
PDS (3)	11	10	11	11
PDS (5)	10	10	9	11
PDS (7)	11	11	11	11
PDS (9)	12	11	11	11

SPA-LDA - Conjunto de teste (S) pré-processado				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	3	8	5	6
PDS (3)	13	11	12	11
PDS (5)	12	12	12	12
PDS (7)	12	12	11	12
PDS (9)	12	12	11	12

Tabela 5.22 - Erros de classificação obtidos pelo modelo PLS-DA aplicado às amostras de teste (S) após padronização dos dados sem e com pré-processamento. A largura da janela empregada na padronização PDS é indicada entre parênteses.

PLS-DA - Conjunto de teste (S) sem pré-processamento				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	10	2	2	3
PDS (3)	10	10	10	10
PDS (5)	11	10	9	10
PDS (7)	10	10	9	10
PDS (9)	10	10	9	10

PLS-DA - Conjunto de teste (S) pré-processado				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	6	4	4	5
PDS (3)	10	10	9	10
PDS (5)	10	10	9	10
PDS (7)	10	10	9	10
PDS (9)	10	9	8	10

Como pode ser observado, para o modelo SPA-LDA (**Tabela 5.21**) o melhor resultado de padronização foi alcançado com o conjunto de teste pré-processado, com apenas três erros obtidos, enquanto o PLS-DA apresenta melhor desempenho com os dados sem pré-

processamento, com dois erros de classificação. Além disso, verifica-se que a menor quantidade de erros foi obtida após a padronização DS, no qual o ajuste é realizado relacionando-se cada comprimento de onda do espectro primário a todos os comprimentos de onda do espectro secundário. Esses resultados corroboram com a informação fornecida pela **Figura 5.22**, que aponta a necessidade de um grande conjunto de variáveis para realizar a transferência.

Na **Figura 5.23**, pode-se comparar os espectros das amostras de teste antes e após padronização DS e PDS (9) usando 8 amostras de transferência, que para a maior parte dos procedimentos apresentou os melhores resultados. Mais uma vez fica em evidência que o ajuste desse conjunto de dados é melhor realizado quando se aplica a padronização DS.

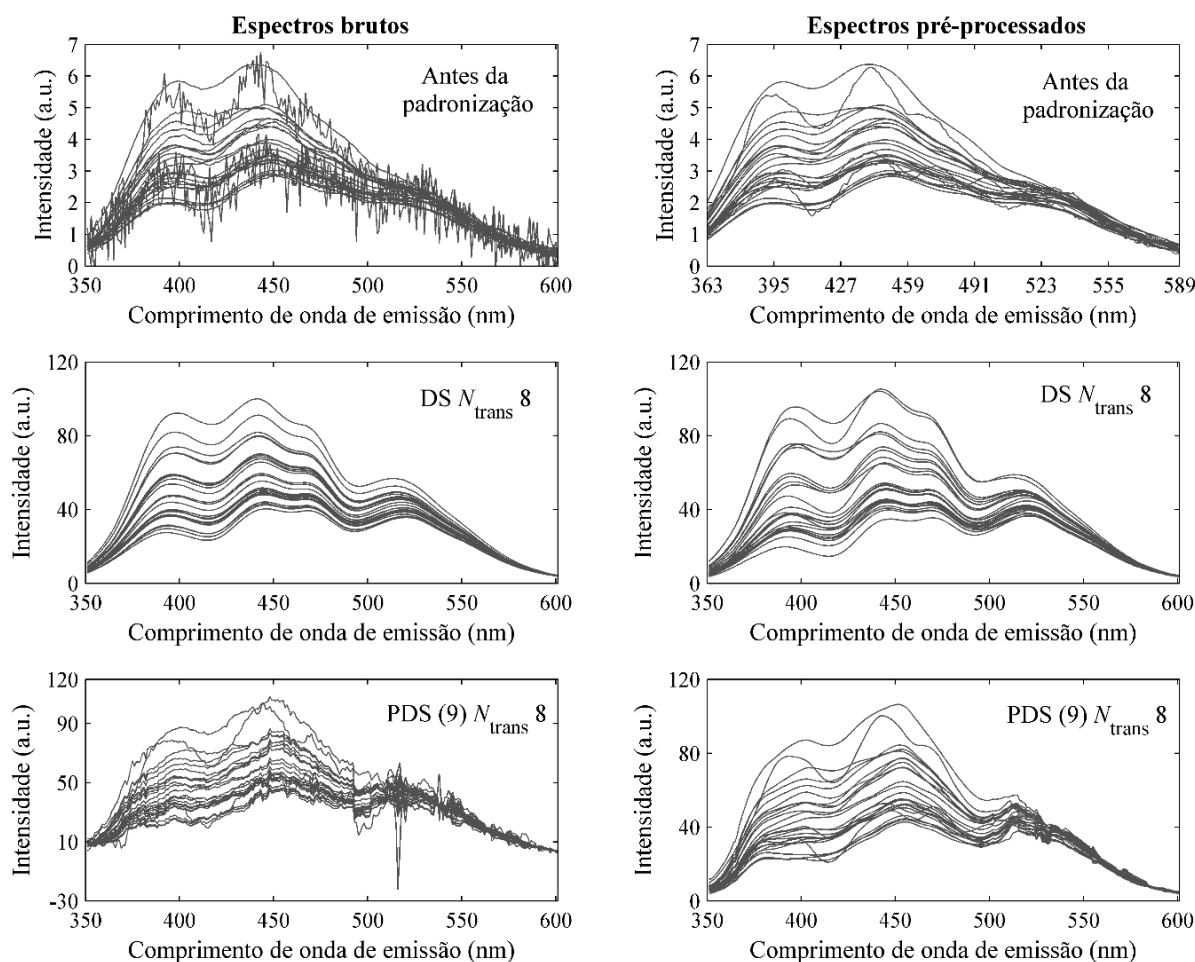


Figura 5.23 - Espectros de emissão de fluorescência molecular das amostras de teste (S). São apresentados os espectros sem e com pré-processamento antes e após procedimentos de padronização usando $N_{trans} = 8$.

A **Figura 5.24** compara os gráficos dos escores obtidos pelo modelo SPA-LDA (**Figura 5.24a** e **5.24b**) e dos valores estimados pelo modelo PLS-DA (**Figura 5.24c** e **5.24d**) aplicados ao conjunto de teste (S) antes e depois da padronização. As **Figuras 5.24b** e **5.24d**

correspondem a padronização DS dos dados pré-processados (para SPA-LDA) e sem pré-processamento (para PLS-DA) utilizando $N_{trans} = 4$ e $N_{trans} = 6$, respectivamente.

Como pode ser visto, os modelos de classificação obtiveram melhores resultados quando os espectros padronizados foram empregados.

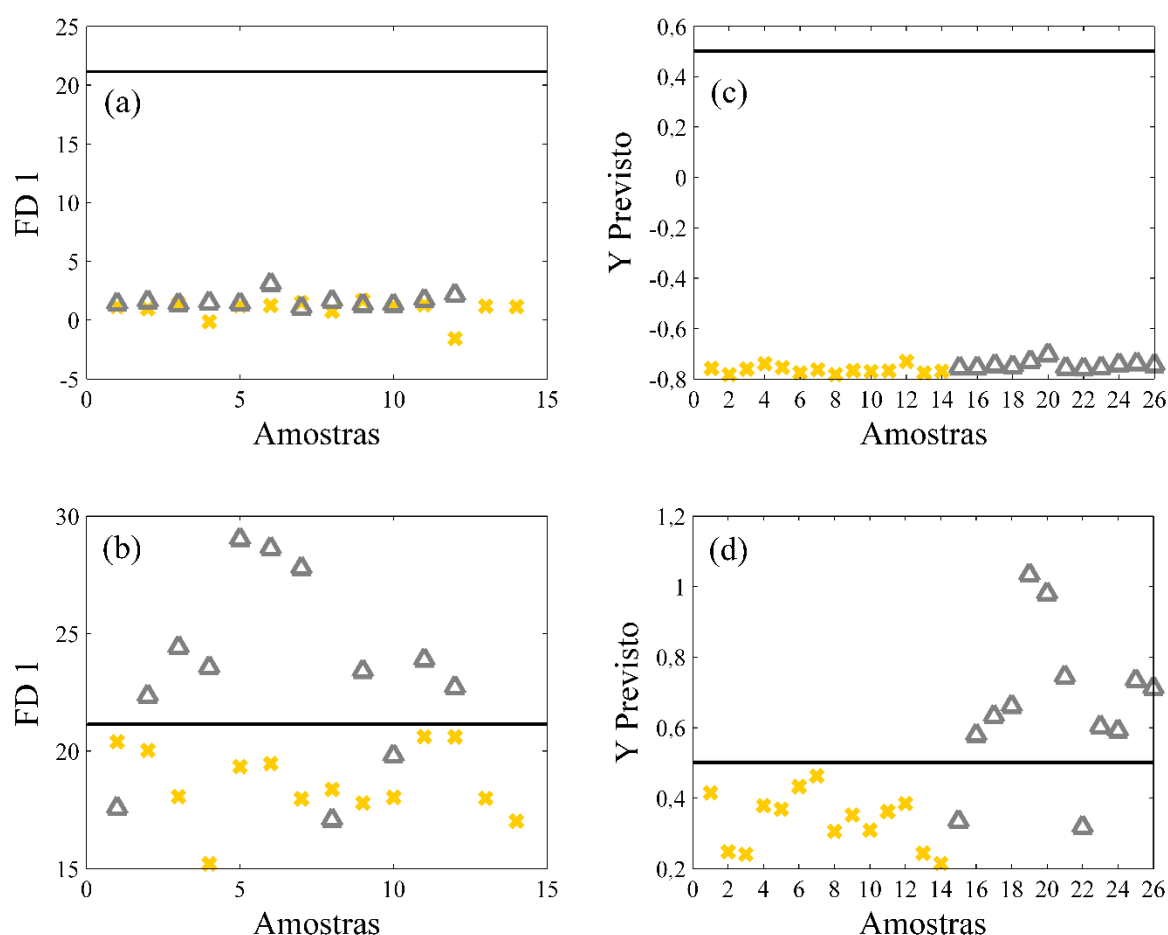


Figura 5.24 - Gráficos dos escores obtidos pelo modelo SPA-LDA para as amostras de teste (S) antes (a) e depois (b) da padronização DS com $N_{trans} = 4$. Valores estimados pelo modelo PLS-DA para as amostras de teste (S) antes (c) e após (d) a padronização DS dos dados pré-processados usando $N_{trans} = 6$. Amostras não-adulteradas (x) e adulteradas (Δ). As linhas horizontais são os limites de decisão estabelecidos pelos modelos SPA-LDA e PLS-DA.

5.4.6 Considerações finais

Neste estudo de caso, os critérios propostos foram avaliados em um problema que envolve uma grande diferença entre as respostas instrumentais. Embora os erros de treinamento e teste (P) obtidos pelos critérios tenham sido similares aos obtidos pelo SPA-LDA na forma usual, eles não conseguiram superar a magnitude das diferenças e realizar a transferência de classificação. Ainda assim, quando usado o critério C2 com 4 e 6 amostras de transferência foi aplicado ao conjunto de teste (S), os erros diminuíram em uma unidade.

Com o propósito de verificar se a transferência de classificação seria possível nesse conjunto de dados, estratégias de padronização DS e PDS foram utilizadas e comparadas para os modelos SPA-LDA e PLS-DA. Os resultados demonstraram que a transferência pode ser eficientemente realizada relacionando cada comprimento de onda do instrumento primário a todos os comprimentos de onda do instrumento secundário, como acontece na padronização DS. De certa forma, esses resultados corroboram com a informação fornecida no gráfico das variáveis selecionadas pelos critérios, onde é apontada a necessidade de um conjunto grande de variáveis para estabelecer uma relação entre os espectros.

5.5 Imagens digitais

5.5.1 Análise dos histogramas

A **Figura 5.25** apresenta os histogramas RGB obtidos para as 54 amostras de AOEV analisadas em cada dispositivo de captura de imagens. Para cada componente de cor, 256 níveis de cores podem ser obtidos variando de 0 a 255. No entanto, muitos desses níveis apresentaram resposta igual a zero, não participando da composição de cor da imagem. Estes níveis foram removidos do conjunto de dados com o cuidado de manter o mesmo número de variáveis (níveis de cores) em cada instrumento. Esses histogramas foram normalizados, de modo que apenas os histogramas resultantes são apresentados e utilizados no procedimento quimiométrico.

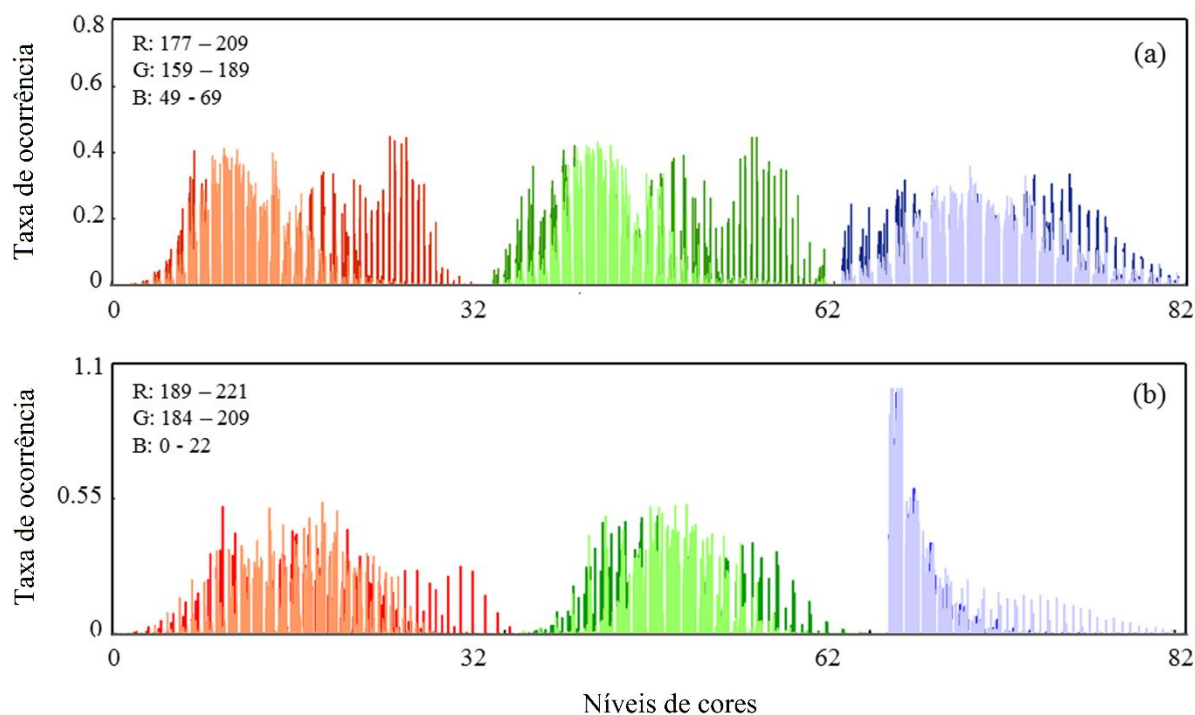


Figura 5.25 - Histogramas RGB das amostras de AOEV adquiridos na webcam (a) e no smartphone (b). As amostras não-adulteradas estão representadas pelas cores mais escuras e as adulteradas pelas cores mais claras. Os valores indicados para as componentes RGB correspondem à faixa de níveis de cores presentes nos histogramas.

Na **Figura 5.25**, o número de níveis de cores após a remoção é apresentado no eixo x, ao passo que os valores correspondentes na escala de 0 a 255 são listados para os componentes RGB nos histogramas. Como pode ser visto, diferentes números de níveis de cores foram obtidos em cada instrumento, sobretudo para a componente B. Além disso, é possível observar que a diferença entre as duas classes de amostras é mais evidente nos histogramas gerados pela webcam, especialmente para as componentes R e G.

5.5.2 Análise exploratória

A **Figura 5.26** apresenta os gráficos dos escores em PC 1 $\times$ PC 2, PC 1 $\times$ PC 3, PC 1 $\times$ PC 4, PC 1 $\times$ PC 5 e PC 1 $\times$ PC 6 para as amostras de teste medidas nos dois dispositivos. As seis PCs são apresentadas em virtude do espalhamento da variância explicada para esse conjunto de dados.

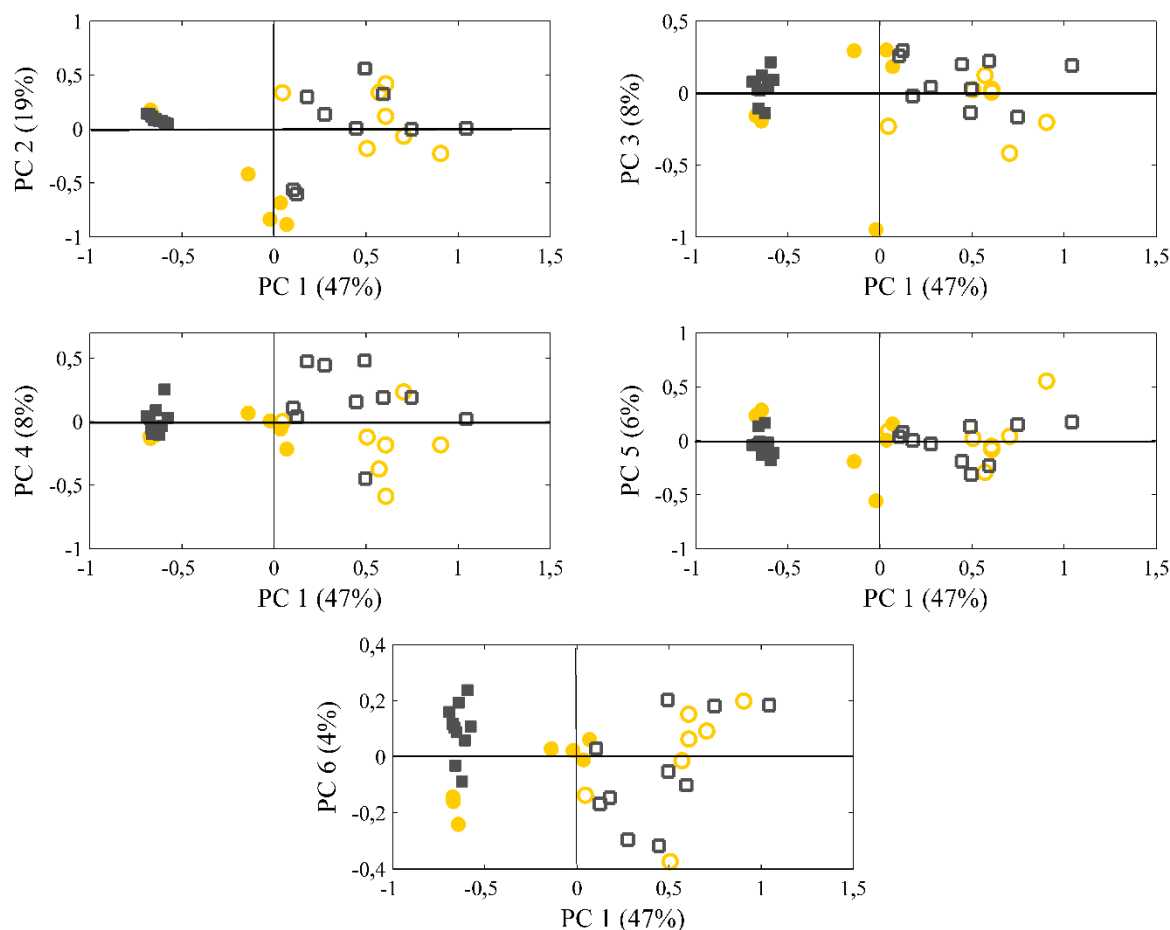


Figura 5.26 - Gráficos dos escores nas seis primeiras PCs para as amostras de AOEV do conjunto teste medidas em ambos os instrumentos. Amostras não-adulteradas (●) e adulteradas (■) medidas na webcam; Amostras não-adulteradas (○) e adulteradas (□) medidas no smartphone.

Em todos os casos, é possível observar que as amostras tendem à separação em função do instrumento de medida ao longo da PC 1, embora exista um grupo de amostras situado na região central dessa PC. Além disso, verifica-se que os dados adquiridos com o smartphone estão sempre distribuídos no lado positivo da PC 1. As PCs 2, 3 e 5 não apresentaram separação significativa, enquanto a PC 4 tende a separar as duas classes de AOEV medidas no smartphone e a PC 6 separa as duas classes de AOEV medidas na webcam.

A **Figura 5.27** mostra o gráfico dos *loadings* usados para verificar a contribuição das variáveis para cada PC.

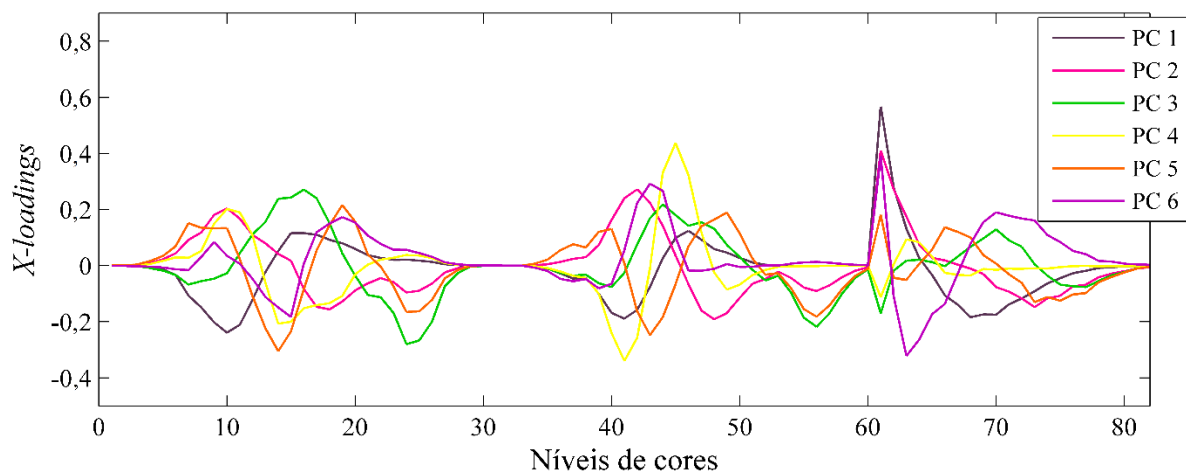


Figura 5.27 - Gráfico dos *loadings* para as seis primeiras PCs.

Conforme apresentado, as variáveis que compõem a componente R contribuem basicamente da mesma forma para todas as PCs analisadas. Na componente G, algumas variáveis aparecem com maior peso na PC 4, onde ocorre a discriminação das duas classes de AOEV medidas com o smartphone. No caso da componente B, a variável de maior peso contribui significativamente para a PC 1, que separa as amostras de acordo com o instrumento com o qual foram medidas, seguida de PC 2 e PC3.

5.5.3 Modelos SPA-LDA

As **Tabelas 5.23** e **5.24** apresentam os resultados de classificação obtidos pelos modelos SPA-LDA na sua forma usual construídos para os dois instrumentos.

Tabela 5.23 - Resultados de classificação obtidos com os modelos SPA-LDA na forma usual construídos para as imagens registradas na webcam. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Webcam (5)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Classe verdadeira	<i>N</i>			100 ^a
Não-adulterada	14	14	-	100 ^b
Adulterada	23	-	23	100 ^c
Conjunto de Teste – Webcam (5)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Classe verdadeira	<i>N</i>			88 ^a
Não-adulterada	7	5	2	71 ^b
Adulterada	10	-	10	100 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Tabela 5.24 - Resultados de classificação obtidos com os modelos SPA-LDA na forma usual construídos para as imagens registradas no smartphone. O número de variáveis empregadas em cada modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento - Smartphone (4)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	14	12	2	86 ^b
Adulterada	23	6	17	74 ^c
Conjunto de Teste – Smartphone (4)				
Classe verdadeira	<i>N</i>	Classe prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	7	7	-	100 ^b
Adulterada	10	1	9	90 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Comparando os resultados, é possível observar que os modelos de cada instrumento apresentaram desempenhos diferentes para os conjuntos de treinamento e teste. Com isso, a escolha do instrumento primário foi guiada com base na quantidade de erros obtidos para o conjunto de treinamento, que por sua vez, foi igual à zero para webcam.

Nas **Tabelas 5.23 e 5.24**, verifica-se também que os erros de treinamento do smartphone mais frequentes estão associados às amostras adulteradas sendo classificadas como não-adulteradas. Neste caso, o teor de óleo de soja adicionado às amostras adulteradas incorretamente classificadas variou aleatoriamente.

A **Tabela 5.25** apresenta os resultados obtidos com a utilização do modelo primário (webcam) na previsão das amostras de teste medidas no instrumento secundário (smartphone).

Tabela 5.25 - Resultados de classificação obtidos com o modelo SPA-LDA na forma usual construído com base nas imagens registradas no instrumento primário (webcam) aplicado às amostras de teste medidas no instrumento secundário (smartphone). *N* indica o número de amostras de teste em cada classe de AOEV.

Conjunto de teste secundário				
Classe Verdadeira	<i>N</i>	Classe Prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Não-adulterada	7	5	2	71 ^b
Adulterada	10	7	3	30 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

Conforme apresentado na **Tabela 5.25**, apenas 47% das amostras foram corretamente classificadas quando o modelo primário foi usado na previsão de amostras medidas no instrumento secundário.

5.5.4 Transferência dos modelos SPA-LDA

Na **Tabela 5.26** são apresentados os resultados obtidos pelos modelos SPA-LDA, obtidos com base nas imagens da webcam, na forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . Para o critério *C1*, quando 4 e 6 amostras de transferência foram utilizadas, erros maiores de treinamento e teste foram obtidos, quando comparado ao modelo SPA-LDA na forma usual. No caso do critério *C2*, os erros foram ainda maiores nos dois subconjuntos, em relação ao SPA-LDA usual.

Tabela 5.26 - Erros de classificação obtidos pelos modelos SPA-LDA na forma usual e com os critérios *C1* e *C2* empregando diferentes N_{trans} . O conjunto de teste (P) corresponde às amostras medidas no instrumento primário.

Parâmetros	N	Resultados de classificação dos modelos SPA-LDA					
		Usual	Critérios	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
Erros de treinamento	37	0	<i>C1</i>	5	5	2	1
			<i>C2</i>	15	15	15	15
Erros de teste (P)	17	2	<i>C1</i>	6	6	4	4
			<i>C2</i>	13	13	13	13

A **Figura 5.28** apresenta o número de erros obtidos após aplicação dos modelos ao conjunto de amostras de teste medido no instrumento secundário. Os resultados para *C1* e *C2* são apresentados para diferentes valores de N_{trans} . O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam os erros.

É possível observar que o critério *C1*, usando $N_{trans} = 8$, conseguiu reduzir os erros de classificação em duas unidades, enquanto *C2* reduziu os erros em apenas uma unidade usando 4, 6 e 8 amostras de transferência. Embora esses resultados não sejam significativos, ainda sim, demonstram que os critérios conseguem diminuir o número de erros.

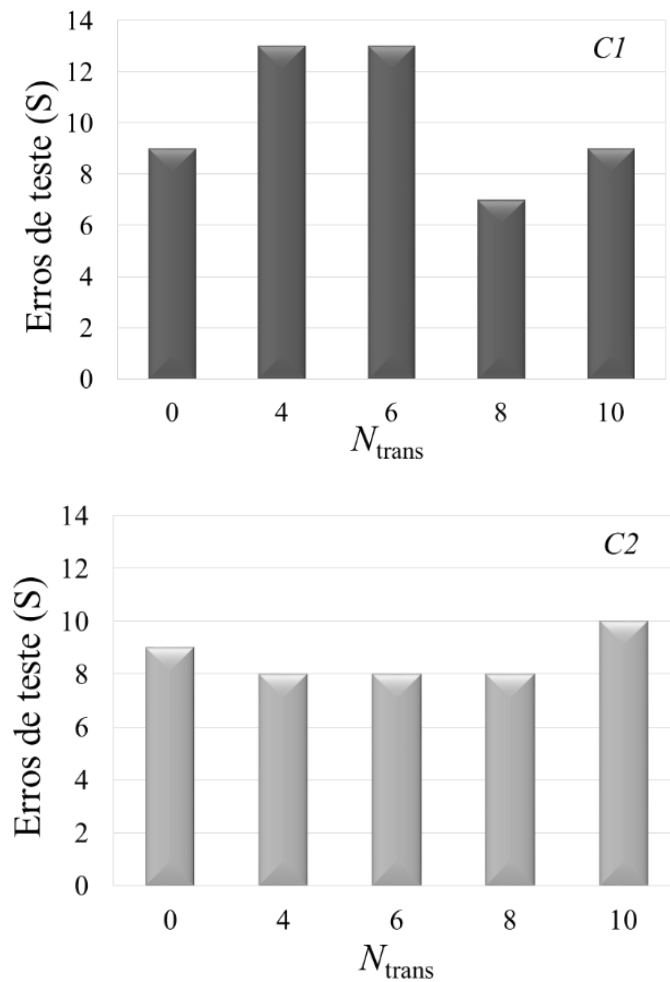


Figura 5.28 - Erros obtidos pelos três modelos de classificação (SPA-LDA na forma usual, critérios *C1* e *C2*) aplicados às amostras de teste (S) de AOEV medidas por imagens digitais a partir do smartphone. O modelo SPA-LDA na forma usual corresponde à $N_{trans} = 0$ e as barras indicam o número de erros.

A **Figura 5.28** revela que o critério *C1*, usando $N_{trans} = 8$, conseguiu reduzir os erros de classificação em duas unidades, enquanto *C2* reduziu os erros em apenas uma unidade usando 4, 6 e 8 amostras de transferência. Embora esses resultados não sejam significativos, ainda sim, demonstram que os critérios conseguem diminuir o número de erros.

A **Figura 5.29** mostra os subconjuntos de variáveis selecionadas pelo algoritmo SPA nos modelos de classificação. É possível observar que, para diferentes N_{trans} , o critério *C2* não selecionou variáveis relacionadas a componente de cor azul. Além disso, o total de variáveis selecionadas por *C2* foi sempre menor quando comparado a *C1* e ao SPA-LDA na forma usual. Com relação a *C1* usando $N_{trans} = 8$, que obteve o melhor resultado de classificação, variáveis foram selecionadas em todas as componentes de cores.

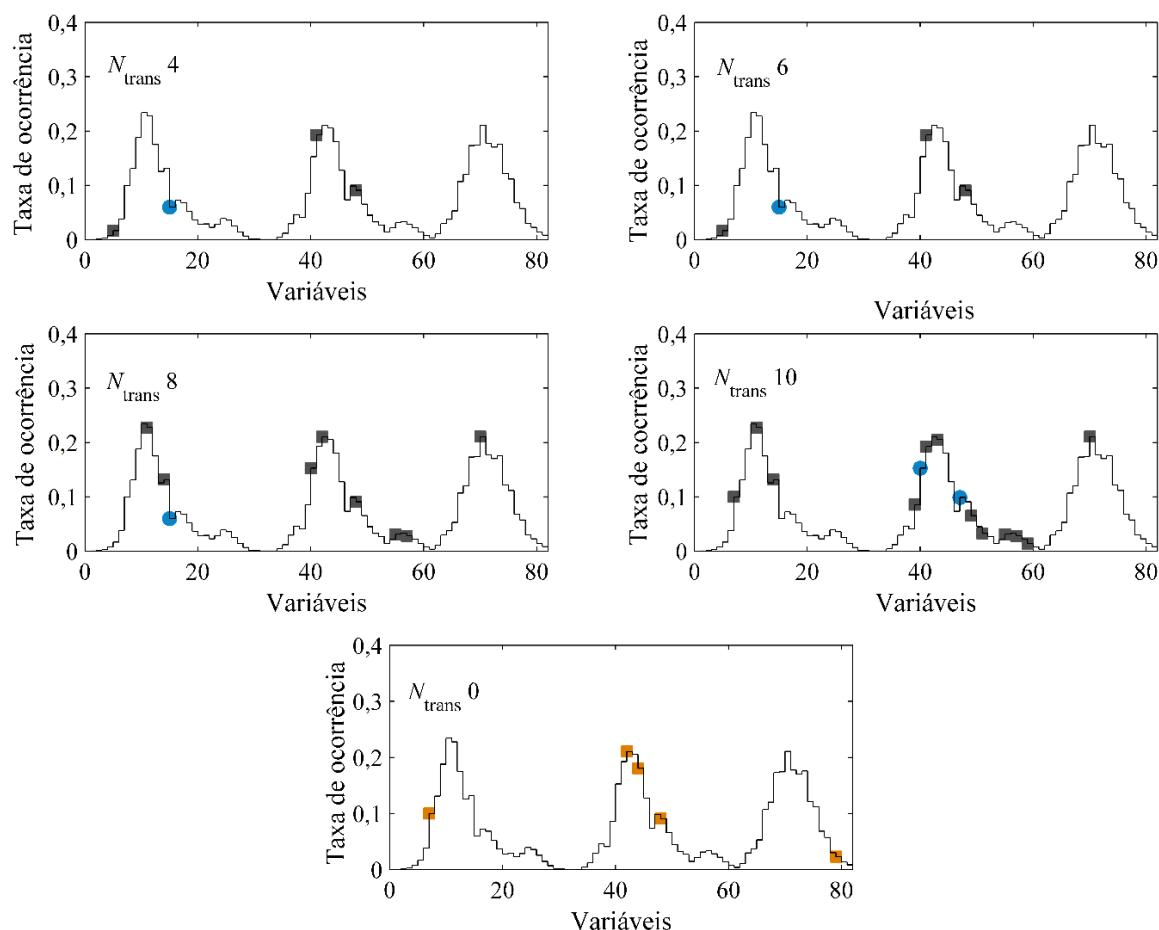


Figura 5.29 - Histogramas médios do conjunto de dados com indicação das variáveis selecionadas pelo algoritmo SPA nos modelos de classificação envolvendo diferentes números de amostras de transferência. Variáveis selecionadas pelos critérios C1: ■, C2: ● e SPA na forma usual: ■.

Como os histogramas do instrumento secundário são muito diferentes dos histogramas do instrumento primário, sobretudo na componente de cor B, espera-se que os procedimentos de padronização forneçam melhores resultados de classificação. Os resultados das aplicações de diferentes estratégias de padronização para os dados de imagens digitais também estão detalhados em Milanez *et al.* (2017) ^[138].

5.5.5 Transferência dos modelos SPA-LDA usando procedimentos de padronização

Para o instrumento primário, além do modelo SPA-LDA na forma usual (resultados na **Tabela 5.23**) também foi construído o modelo baseado na PLS-DA. A **Tabela 5.27** detalha os resultados encontrados. Comparando os modelos SPA-LDA (**Tabela 5.23**) e PLS-DA (**Tabela 5.27**), observa-se que a diferença está no erro de treinamento obtido pelo PLS-DA.

Tabela 5.27 - Resultados de classificação obtidos com o modelo PLS-DA construído para as imagens registradas no instrumento primário. O número de variáveis latentes empregadas no modelo é apresentado entre parênteses. *N* indica o número de amostras de treinamento e teste em cada classe de AOEV.

Conjunto de Treinamento				
Classe	N	Instrumento primário (6)		Desempenho (%)
		Classe Prevista		
		Não-adulterada	Adulterada	
Verdadeira				97 ^a
Não-adulterada	14	13	1	93 ^b
Adulterada	23	-	23	100 ^c
Conjunto de Teste				
Classe	N	Instrumento primário (6)		Desempenho (%)
		Classe Prevista		
		Não-adulterada	Adulterada	
Verdadeira				88 ^a
Não-adulterada	7	5	2	71 ^b
Adulterada	10	-	10	100 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

A **Tabela 5.28** mostra os erros de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. Os mesmos erros obtidos com o modelo SPA-LDA foram encontrados com o PLS-DA.

Tabela 5.28 - Resultados de classificação decorrentes da aplicação do modelo PLS-DA ao conjunto de amostras de teste medido no instrumento secundário. *N* indica o número de amostras de teste em cada classe de AOEV.

Conjunto de teste secundário				
Classe	N	Classe Prevista		Desempenho (%)
		Não-adulterada	Adulterada	
Verdadeira				47 ^a
Não-adulterada	7	5	2	71 ^b
Adulterada	10	7	3	30 ^c

^a Acurácia, ^b Especificidade, ^c Sensibilidade.

O número de erros obtidos pelos modelos SPA-LDA e PLS-DA após padronização DS e PDS do conjunto de teste (S) é apresentado na **Tabela 5.29** para as amostras de teste medidas no instrumento secundário, o smartphone. Neste caso, janelas de 3, 5, 7, 9 e 11 pontos foram empregadas na padronização PDS. Diferente dos conjuntos de dados anteriores, os melhores resultados de classificação envolvendo imagens digitais foram obtidos após padronização direta em etapas (PDS), especificamente para $N_{\text{trans}} = 6$ e janela de 5 pontos para o modelo SPA-LDA. O menor número de erros obtidos após a padronização PDS para as imagens digitais pode estar

relacionado ao fato do conjunto de dados ser formado por variáveis discretas, para o qual a padronização em janelas de variáveis mostrou-se mais eficiente.

Tabela 5.29 – Erros de classificação obtidos pelos modelos SPA-LDA e PLS-DA aplicados às amostras de teste (S) analisadas no instrumento secundário após padronização. A largura da janela empregada na padronização PDS é indicada entre parênteses.

SPA-LDA (Conjunto de teste secundário)				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	10	12	11	9
PDS (3)	12	9	9	6
PDS (5)	11	3	6	4
PDS (7)	11	5	7	5
PDS (9)	9	7	9	4
PDS (11)	9	6	9	5

PLS-DA (Conjunto de teste secundário)				
Método	N_{trans} 4	N_{trans} 6	N_{trans} 8	N_{trans} 10
DS	9	10	11	8
PDS (3)	8	10	11	8
PDS (5)	7	8	9	8
PDS (7)	4	9	7	9
PDS (9)	6	7	8	9
PDS (11)	8	6	8	9

A **Figura 5.30** compara os gráficos dos escores obtidos pelo modelo SPA-LDA (**Figura 5.30a** e **Figura 5.30b**) e dos valores estimados pelo modelo PLS-DA (**Figura 5.30c** e **Figura 5.30d**) aplicados ao conjunto de teste medido no instrumento secundário antes e depois padronização. As **Figuras 5.30b** e **5.30d** correspondem a padronização PDS (5) para SPA-LDA e PDS (7) para PLS-DA utilizando $N_{trans} = 6$ e $N_{trans} = 4$, respectivamente. Como pode ser visto, os modelos de classificação obtiveram melhores resultados quando os histogramas padronizados foram empregados.

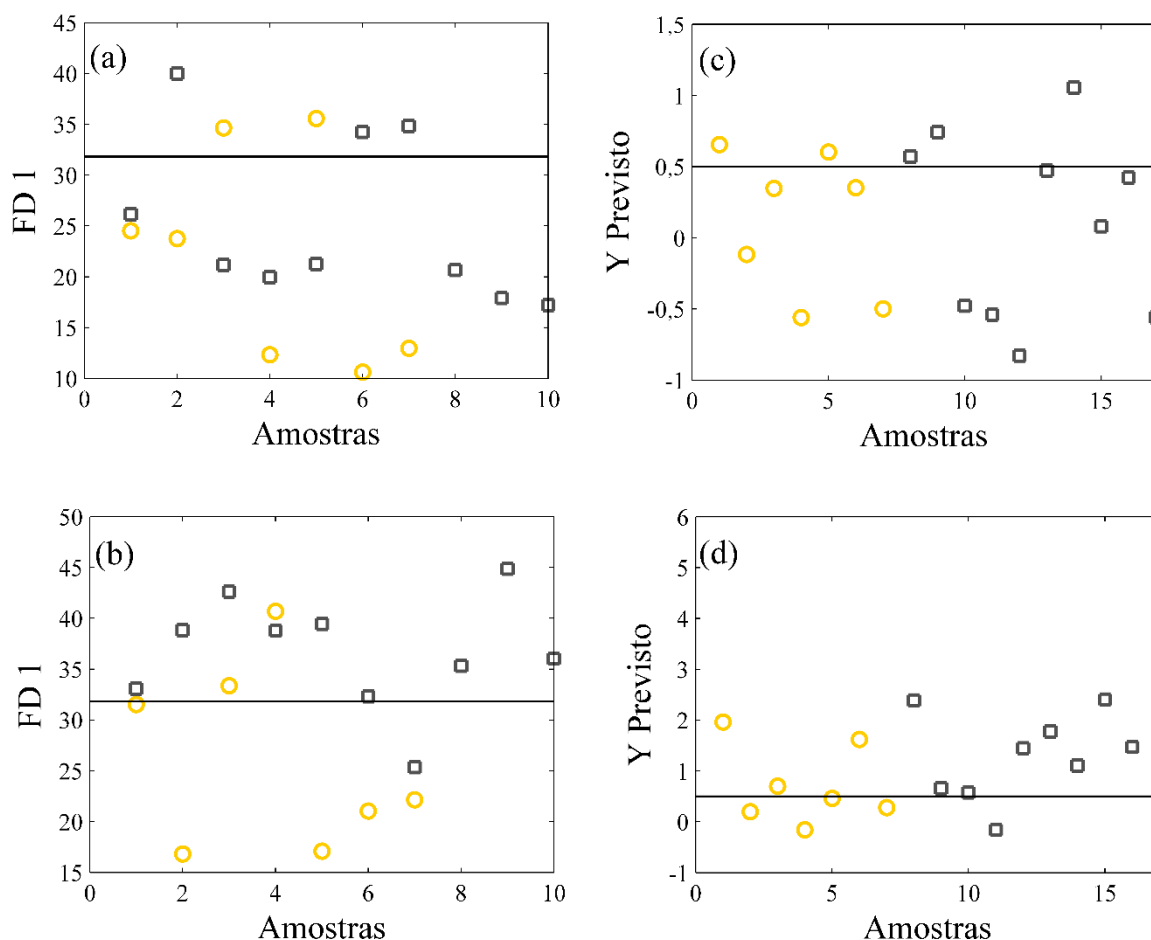


Figura 5.30 - Gráficos dos escores obtidos pelo modelo SPA-LDA para as amostras de teste antes (a) e depois (b) da padronização PDS (5) com $N_{\text{trans}} = 6$. Valores estimados pelo modelo PLS-DA para as amostras de teste antes (c) e após (d) PDS (7) dos dados usando $N_{\text{trans}} = 4$. Amostras não-adulteradas (●) e adulteradas (■). As linhas horizontais são os limites de decisão estabelecidos pelos modelos SPA-LDA e PLS-DA.

5.5.6 Considerações finais

Neste estudo de caso, os critérios foram avaliados em um problema que envolve uma diferença significativa entre as respostas instrumentais. Para o critério *C1*, quando 4 e 6 amostras de transferência foram utilizadas, os erros de treinamento e teste (P) obtidos foram maiores que os do modelo SPA-LDA na forma usual. O critério *C2*, no entanto, apresentou resultados tanto de treinamento com de teste (P) muito discrepantes, em relação ao SPA-LDA usual. Quando aplicados ao conjunto de teste medido no instrumento secundário, os critérios conseguiram reduzir os erros em duas e uma unidade para *C1* ($N_{\text{trans}} = 8$) e *C2* (4, 6 e 8 amostras de transferência), respectivamente.

A fim de verificar se a transferência de classificação seria possível nesse conjunto de dados, estratégias de padronização DS e PDS foram utilizadas e comparadas para os modelos SPA-LDA e PLS-DA. Os resultados demonstraram que a transferência pode ser eficientemente

realizada relacionando cada comprimento de onda do instrumento primário a uma janela em torno desse comprimento de onda no instrumento secundário, por meio do método PDS. Esse resultado pode ser relacionado ao fato do conjunto de dados ser formado por variáveis discretas, fazendo com a padronização usando janelas de variáveis seja mais eficiente.

Capítulo VI

CONCLUSÕES

6 CONCLUSÕES

Nesta tese, foram apresentados dois novos critérios para seleção de variáveis robustas usando o algoritmo das projeções sucessivas (SPA), acoplado à análise discriminante linear, para transferência de modelos de classificação. Os critérios são baseados na inclusão de amostras de transferência, medidas no instrumento secundário, no cálculo do custo de cada subconjunto candidato de variáveis, a fim de minimizar o número de erros obtidos na classificação das amostras medidas no instrumento secundário.

Os critérios foram avaliados para quatro estudos de caso envolvendo diferentes técnicas analíticas e dois problemas de classificação: adulteração de etanol hidratado combustível (espectroscopia NIR) e azeite de oliva extra virgem (espectrometria UV-Vis, espectrometria de fluorescência molecular e imagens digitais). Como parâmetros de desempenho, foram utilizadas sensibilidade, especificidade e acurácia obtidas para o conjunto de amostras de teste medido no instrumento secundário em comparação ao SPA-LDA na forma usual e aos métodos de padronização direta (DS) e padronização direta em etapas (PDS).

Para a espectroscopia NIR e espectrometria UV-Vis, os dois critérios reduziram substancialmente o número de erros obtidos na classificação das amostras medidas no instrumento secundário. Além disso, foi observado que o número de variáveis selecionadas pelo SPA, a posição dessas variáveis e o número de erros de classificação variaram com o número de amostras de transferência empregadas. Quando comparados com o SPA-LDA na forma usual e os métodos DS e PDS, os critérios mostraram resultados similares ou melhores.

Quando os critérios foram aplicados aos dados de espectrometria de fluorescência molecular e imagens digitais, a grande diferença entre as respostas não permitiu que os modelos alcançassem um desempenho satisfatório, de modo que os espectros e histogramas registrados nos respectivos instrumentos secundários precisam ser padronizados antes da classificação. Nesses casos, as adaptações em *C1* e *C2* não se mostraram robustas o suficiente para vencer a magnitude da diferença das respostas analíticas registradas para os dois instrumentos, confirmando as conclusões da literatura.

Estes resultados sugerem que a abordagem proposta, utilizando um dos critérios desenvolvidos, é uma alternativa promissora à recalibração completa do modelo ou padronização das respostas instrumentais, desde que as diferenças entre as respostas instrumentais não ultrapassem certos limites.

Uma das principais vantagens dos critérios propostos é que as amostras de transferência empregadas na avaliação da métrica G_{trans} só precisam ser medidas no instrumento secundário,

bem como acontece com a métrica do RMSET para problemas de calibração. Na prática, esta é uma vantagem importante em relação aos métodos de padronização como DS e PDS, especialmente se os instrumentos primário e secundário não estiverem localizados no mesmo laboratório, se as amostras estiverem deterioradas, quando se tem dificuldade no transporte das amostras ou se o instrumento principal não estiver mais disponível.

REFERÊNCIAS

REFERÊNCIAS

- [1] FORINA, M.; CASOLINO, M. C.; MARTINEZ, C. P.; Multivariate calibration: applications to pharmaceutical analysis, **Journal of Pharmaceutical and Biomedical Analysis**, v. 18, p. 21-33, 1998.
- [2] ESCANDAR, G. M.; DAMIANI, P. C.; GOICOECHEA, H. C.; OLIVIERI, A. C.; A review of multivariate calibration methods applied to biomedical analysis, **Microchemical Journal**, v. 82, p. 29-42, 2006.
- [3] PINTO, L.; DÍAZ NIETO, C. H.; ZÓN, M. A.; FERNÁNDEZ, H.; ARAÚJO, M. C. U.; Handling time misalignment and rank deficiency in liquid chromatography by multivariate curve resolution: Quantitation of five biogenic amines in fish, **Analytica Chimica Acta**, v. 902, p. 59-69, 2016.
- [4] ZHANG, F.; LIU, T.; WANG, X. Z.; LIU, J.; JIANG, X.; Comparative study on ATR-FTIR calibration models for monitoring solution concentration in cooling crystallization, **Journal of Crystal Growth**, v. 459, p. 50-55, 2017.
- [5] TENCATE, A. J.; KALIVAS, J. H.; WHITE, A. J.; Fusion strategies for selecting multiple tuning parameters for multivariate calibration and other penalty based processes: A model updating application for pharmaceutical analysis, **Analytica Chimica Acta**, v. 921, p. 28-37, 2016.
- [6] BERRUETA, L. A.; ALONSO-SALCES, R. M.; HÉBERGER, K.; Supervised pattern recognition in food analysis, **Journal of Chromatography A**, v.1158, p. 196-214, 2007.
- [7] GONZÁLEZ, A. G.; Use and misuse of supervised pattern recognition methods for interpreting compositional data. **Journal of Chromatography A**, v. 1158, p. 215–225, 2007.
- [8] VALDERRAMA, L.; VALDERRAMA, P.; Nondestructive identification of blue pen inks for documentoscopy purpose using iPhone and digital image analysis including an approach for interval confidence estimation in PLS-DA models validation, **Chemometrics and Intelligent Laboratory Systems**, v. 156, p. 188-195, 2016.
- [9] GÓRSKI, Ł.; SORDOŃ, W.; CIEPIELA, F.; KUBIAK, W. W.; JAKUBOWSKA, M.; Voltammetric classification of ciders with PLS-DA, **Talanta**, v. 146, p. 231-236, 2016.
- [10] GEANA, E. I.; POPESCU, R.; COSTINEL, D.; DINCA, O. R.; IONETE, R. E.; STEFANESCU, I., ARTEM, V.; BALA, C.; Classification of red wines using suitable markers coupled with multivariate statistic analysis, **Food Chemistry**, v. 192, p. 1015-1024, 2016.
- [11] PEREIRA, H. V.; AMADOR, V. S.; SENA, M. M.; AUGUSTI, R.; PICCIN, E.; Paper spray mass spectrometry and PLS-DA improved by variable selection for the forensic discrimination of beers, **Analytica Chimica Acta**, v. 940, p. 104-112, 2016.
- [12] HONORATO, F. A.; GALVÃO, M. F. P.; BARROS NETO, B.; ARAÚJO, M. C. U.; DE CARVALHO F. R.; Robust modeling for multivariate calibration transfer by the successive

projections algorithm, **Chemometrics and Intelligent Laboratory Systems**, v. 76, p. 65-72, 2005.

[13] FEUDALE, R. N.; WOODY, N. A.; TAN, H.; MYLES, A. J.; BROWN, S. D.; FERRÉ, J.; Transfer of multivariate calibration models: a review, **Chemometrics and Intelligent Laboratory Systems**, v. 64, p. 181-192, 2002.

[14] HONORATO, F. A.; BARROS NETO, B.; MARTINS, M. N.; GALVÃO, R. K. H.; PIMENTEL, M. F.; Transferência de calibração em métodos multivariados, **Química Nova**, v. 30, p. 1301-1312, 2007.

[15] R.K.H. GALVÃO, S.F.C. SOARES, M.N. MARTINS, M.F. PIMENTEL, M.C.U. ARAÚJO, Calibration transfer employing univariate correction and robust regression, **Analytica Chimica Acta**, v. 864, p. 1-8, 2015.

[16] WANG, Y.; VELTKAMP, D. J.; KOWALSKI, B. R.; Multivariate Instrument Standardization, **Analytical Chemistry**, v. 63, p. 2750-2756, 1991.

[17] GONDIM, C. S.; JUNQUEIRA, R. G.; SOUZA, S. V. C.; Tendências em validação de métodos de ensaios qualitativos, **Revista do Instituto Adolfo Lutz**, v. 70, p. 433-447, 2011.

[18] NOORD, O. E.; Multivariate calibration standardization, **Chemometrics and Intelligent Laboratory Systems**, v. 25, p. 85-97, 1994.

[19] BOUVERESSE, E.; MASSART, D. L.; Standardisation of near-infrared spectrometric instruments: A review, **Vibrational Spectroscopy**, v. 11, p. 3-15, 1996.

[20] VASCOCELOS, F.V. C. **Uso da região espectral de sobretons para determinação do teor de biodiesel e classificação de misturas diesel/biodiesel adulteradas com óleo vegetal**. 2011. 93 f. Dissertação (Mestrado em Química) – Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2011.

[21] BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics A Practical Guide**. New York: John Wiley & Sons, 1998.

[22] PONTES, M. J. C. **Algoritmo das projeções sucessivas para seleção de variáveis espectrais em problemas de classificação**. 2009. 144 f. Tese (Doutorado em Química) – Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2009.

[23] HONORATO, F. A. **Previsão de propriedades das gasolinas do Nordeste empregando espectroscopia NIR/MIR e transferência de calibração**. 2006. 106 f. Tese (Doutorado em Química) – Departamento de Química Fundamental, Universidade Federal de Pernambuco, Recife, 2006.

[24] SAVITZKY, A.; GOLAY, M. J. E.; Smoothing and differentiation of data by simplified least squares procedures, **Analytical Chemistry**, v. 36, p. 1627-1639, 1964.

[25] BRERETON, R. G. **Chemometrics: data analysis for the laboratory and chemical plant**, John Wiley & Sons Ltd, Bristol, 2003.

- [26] CORREIA, P. R. M.; FERREIRA, M. M. C.; Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos, **Química Nova**, v. 30, p. 481-487, 2007.
- [27] FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPE, P. L. O.; Quimiometria I: calibração multivariada, um tutorial, **Química Nova**, v. 22, p. 724-731, 1999.
- [28] MOITA NETO, J.M.; MOITA, G.C; Uma introdução à análise exploratória de dados multivariados, **Química Nova**, v. 21, p. 467-469, 1998.
- [29] LENHARDT, L.; BRO, R.; ZEKOVIĆ, I.; DRAMIĆANIN, T.; DRAMIĆANIN, M. D.; Fluorescence spectroscopy coupled with PARAFAC and PLS DA for characterization and classification of honey, **Food Chemistry**, v. 175, p. 284-291, 2015.
- [30] BOTELHO, B. G.; REIS, N.; OLIVEIRA, L. S.; SENA, M. M.; Development and analytical validation of a screening method for simultaneous detection of five adulterants in raw milk using mid-infrared spectroscopy and PLS-DA, **Food Chemistry**, v. 181, p. 31-37, 2015.
- [31] ALMEIDA, M. R.; FIDELIS, C. H. V.; BARATA, L. E. S.; POPPI, R. J.; Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation, **Talanta**, v. 117, p. 305-311, 2013.
- [32] SOUTO, U. T. C. P.; BARBOSA, M. F.; DANTAS, H. V.; PONTES, A. S.; LYRA, W. S.; DINIZ, P. H. G. D.; ARAÚJO, M. C. U.; SILVA, E. C.; Identification of adulteration in ground roasted coffees using UV-Vis spectroscopy and SPA-LDA, **LWT - Food Science and Technology**, v. 63, p. 1037-1041, 2015.
- [33] VILAR, W. T. S.; ARANHA, R. M.; MEDEIROS, E. P.; PONTES, M. J. C.; Classification of Individual Castor Seeds Using Digital Imaging and Multivariate Analysis, **Journal of Brazilian Chemical Society**, v. 26, p. 102-109, 2015.
- [34] MARQUES, A. S.; MORAES, E. P.; JÚNIOR, M. A. A.; MOURA, A. D.; NETO, V. F. A.; NETO, R. M.; LIMA, K. M. G.; Rapid discrimination of *klebsiella pneumoniae* carbapenemase 2 – producing and non-producing *klebsiella pneumoniae* strains using near-infrared spectroscopy (NIRS) and multivariate analysis, **Talanta**, v. 134, p. 126-131, 2015.
- [35] PEERBHAY, K. Y.; MUTANGA, O.; ISMAIL, R.; Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu-Natal, South Africa, **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 79, p. 19-28, 2013.
- [36] ALMEIDA, M. R.; CORREA, D. N.; ROCHA, W. F. C.; SCAFI, F. J. O.; POPPI, R. J.; Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation, **Microchemical Journal**, v. 109, p. 170-177, 2013.
- [37] PONTES, M. J. C.; PEREIRA, C. F.; PIMENTEL, M. F.; VASCONCELOS, F. V. C.; SILVA, A. G. B.; Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification, **Talanta**, v. 85, p. 2159-2165, 2011.

- [38] FISHER, R. A. The use of multiple measurements in taxonomic problems, **Annals of Eugenics**, v.7, p. 179–188, 1936.
- [39] BRUNS, R. E.; FAIGLE, J. F. G.; Quimiometria, **Química Nova**, v. 8, p. 84-99, 1985.
- [40] DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D. L.; The Mahalanobis distance, **Chemometrics and Intelligent Laboratory Systems**, v. 50, 1-18, 2000.
- [41] MASSART, D. L.; VANDEGINSTE, B. G. M.; BUYDENS, S. J.; Lewi, P. J.; Smeyers-Verbeke. Journal Handbook of Chemometrics and Qualimetrics: Parte B, Amsterdam: Elsevier, 1997.
- [42] SIEDLECKI, W.; SKLANSKY, J.; On automatic feature selection, **International Journal of Pattern Recognition and Artificial Intelligent**, v. 2, p. 197-220, 1988.
- [43] DASH, M.; LIU, H.; Feature selection for classification, **Intelligent Data for Analysis**, v. 1, p. 131-156, 1997.
- [44] HIBBERT, D. B.; Genetic algorithms in chemistry, **Chemometrics and Intelligent Laboratory Systems**, v. 19, p. 277-293, 1993.
- [45] EFRON, B. An introduction to the Bootstrap (Hardcover). Publisher: Chapman & Hall/CRC, May 15, 1994.
- [46] FORINA, M.; LANTERI, S.; CASALE, M.; OLIVEROS, M. C. C.; Stepwise orthogonalization of predictors in classification and regression techniques: An “old” technique revisited, **Chemometrics and Intelligent Laboratory Systems**, v. 87, p. 252-261, 2007.
- [47] CERQUEIRA, E. O.; ANDRADE, J. C.; POPPI, R. J.; Redes Neurais e suas aplicações em calibração multivariada, **Química Nova**, v. 24, p. 864-873, 2001.
- [48] ARAÚJO, M. C. U.; SALDANHA, T. C. B.; GALVÃO, R. K. H.; YONEYAMA, T.; CHAME, H. C.; VISANI, V.; The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, **Chemometrics and Intelligent Laboratory Systems**, v. 57, p. 65-73, 2001.
- [49] PONTES, M. J. C.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; MOREIRA, P. N. T.; PESSOA NETO, O. D.; JOSÉ, G. E.; SALDANHA, T. C. B. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems**, v. 78, p. 11-18, 2005.
- [50] SOARES, S. F. C.; GOMES, A. A.; FILHO, A. R. G.; ARAÚJO, M. C. U.; GALVÃO, R. K. H.; The successive projections algorithm, **Trends in Analytical Chemistry**, v. 42, p. 84-98, 2013.
- [51] SOARES, S. F. C.; GALVÃO, R. K. H.; PONTES, M. J. C.; ARAÚJO, M. C. U.; A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems, **Journal of the Brazilian Chemical Society**, v. 25, p. 176-181, 2014.

- [52] WANG, Z. Y.; DEAN, T.; KOWALSKI, B. R.; Additive background correction in multivariate instrument standardization, **Analytical Chemistry**, v. 67, p. 2379-2385, 1995.
- [53] WULFERD, F.; KOK, W. T.; NOORD, O. E.; SMILDE, A. K.; Correction of temperature-induced spectral variation by continuous piecewise direct standardization, **Analytical Chemistry**, v. 72, p. 1639-1644, 2000.
- [54] BOUVERESSE, E.; MASSART, D. L.; Improvement of the piecewise direct standardisation procedure for the transfer of NIR spectra for multivariate calibration, **Chemometrics and Intelligent Laboratory Systems**, v. 32, p. 201-213, 1996.
- [55] LIMA, F. S. G.; BORGES, L. E. P.; Evaluation of standardisation methods of near infrared calibration models, **Journal of Near Infrared Spectroscopy**, 10, 269-278, 2002.
- [56] SALGUERO-CHAPARRO, L.; PALAGOS, B.; PEÑA-RODRÍGUEZ, F.; ROGER, J. M.; Calibration transfer of intact olive NIR spectra between a pre-dispersive instrument and a portable spectrometer, **Computers and Electronics in Agriculture**, v. 96, p. 202-208, 2013.
- [57] DA SILVA, N. C.; CAVALCANTI, C. J.; HONORATO, F. A.; AMIGO, J. M.; PIMENTEL, M. F.; Standardization from a benchtop to a handheld NIR spectrometer using mathematically mixed NIR spectra to determine fuel quality parameters, **Analytica Chimica Acta**, v. 954, p. 32-42, 2017.
- [58] SEICHTER, F.; VOGT, J.; RADERMACHER, P.; MIZAIKOFF, B.; Nonlinear calibration transfer based on hierarchical Bayesian models and Lagrange Multipliers: Error bounds of estimates via Monte Carlo e Markov Chain sampling, **Analytica Chimica Acta**, v. 951, p. 32-45, 2017.
- [59] MALLI, B.; BIRLUTIU, A.; NATSCHLÄGER, T.; Standard-free calibration transfer - An evaluation of different techniques, **Chemometrics and Intelligent Laboratory Systems**, v. 161, p. 49-60, 2017.
- [60] DA SILVA, V. H.; DA SILVA, J. J.; PEREIRA, C. F.; Portable near-infrared instruments: Application for quality control of polymorphs in pharmaceutical raw materials and calibration transfer, **Journal of Pharmaceutical and Biomedical Analysis**, v. 134, p. 287-294, 2017.
- [61] FONOLLOSA, J.; FERNÁNDEZ, L.; GUTIÉRREZ-GÁLVEZ, A.; HUERTA, R.; MARCO, S.; Calibration transfer and drift counteraction in chemical sensor arrays using Direct Standardization, **Sensor Actuator B**, v. 236, p. 1044-1053, 2016.
- [62] LIANG, C.; YUAN, H.; ZHAO, Z.; SONG, C.; WANG, J.; A new multivariate calibration model transfer method of near-infrared spectral analysis, **Chemometrics and Intelligent Laboratory Systems**, v. 153, p. 51-57, 2016.
- [63] DI ANIBAL, C. V.; RUISÁNCHEZ, I.; FERNÁNDEZ, M.; FORTEZA, R.; CERDÀ, V.; CALLAO, M. P.; Standardization of UV-visible data in a food adulteration classification problem, **Food Chemistry**, v. 134, p. 2326-2331, 2012.

- [64] BERGO, M. C. J.; **Transferência de calibração na discriminação de mogno e espécies semelhantes utilizando NIRS e PLS-DA**. 2014. 133 f. Dissertação (Mestrado em Química) - Universidade de Brasília, Brasília, 2014.
- [65] SILVA, N. C.; PIMENTEL, M. F.; HONORATO, R. S.; TALHAVINI, M.; MALDANER, A. O.; HONORATO, F. A.; Classification of Brazilian and foreign gasolines adulterated with alcohol using infrared spectroscopy, **Forensic Science International**, v. 253, p. 33-42, 2015.
- [66] BRERETON, R. G.; Chemometrics for Pattern Recognition, John Wiley & Sons Ltd, Bristol, 2009.
- [67] WOLD, S.; Pattern recognition by means of disjoint principal component models, **Pattern Recognition**, v. 8, p. 127-139, 1976.
- [68] SWIERENGA, H.; HAANSTRA, W. G.; WEIJER, A. P.; BUYDENS, L. M. C.; Comparison of two different approaches toward model transferability in NIR spectroscopy, **Applied Spectroscopy**, v. 52, p. 7 - 16, 1998.
- [69] DE NOORD, O. E.; The influence of data preprocessing on the robustness and parsimony of multivariate calibration models, **Chemometrics and Intelligent Laboratory Systems**, v. 23, p. 65-70, 1994.
- [70] WOLD, S.; ANTTI, H.; LINDGREN, F.; ÖHMAN, J.; Orthogonal signal correction of near-infrared spectra, **Chemometrics and Intelligent Laboratory Systems**, v. 44, p. 175-185, 1998.
- [71] CAI, C.; HARRINGTON, P. B.; Different discrete wavelet transforms applied to denoising analytical data, **Journal of Chemical Information and Computer Sciences**, v.38, p. 1161-1170, 1998.
- [72] ANDREW, A.; FEARN, T.; Transfer by orthogonal projection: making near-infrared calibrations robust to between-instrument variation, **Chemometrics and Intelligent Laboratory Systems**, v. 72, p. 51-56, 2004.
- [73] MYLES, A. J.; ZIMMERMAN, T. A.; BROWN, S. D; Transfer of multivariate classification models between laboratory and process near-infrared spectrometers for the discrimination of green Arabica and Robusta coffee beans, **Applied Spectroscopy**, v. 60, p. 1198-1203, 2006.
- [74] SWIERENGA, H.; GROOT, P. J.; WEIJER, A. P.; DERKSEN, M. W. J.; BUYDENS, L. M. C; Improvement of PLS model transferability by robust wavelength selection, **Chemometrics and Intelligent Laboratory Systems**, v. 41, p. 237 – 248, 1998.
- [75] BOUVERESSE, E.; HARTMANN, C.; MASSART, D. L.; LAST, I. R.; PREBBLE, K. A; Standardization of near-infrared spectrometric instruments, **Analytical Chemistry**, v. 68, p. 982-990, 1996.

- [76] KALIVAS, J. H.; ROBERTS, N.; SUTTER, J. M.; Global optimization by simulated annealing with wavelength selection for ultraviolet-visible spectrophotometry, *Analytical Chemistry*, v. 61, p. 2024 – 2030, 1989.
- [77] KENNARD, R. W.; STONE, L. A. Computer aided design of experiments. *Technometrics*, v. 11, p. 137-148, 1969.
- [78] DANTAS FILHO, H. A.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; SILVA, E. C.; SALDANHA, T. C. B.; JOSÉ, G. E.; PASQUINI, C.; RAIMUNDO JR., I.; ROHWEDDER, J. J. R.; A strategy for selecting calibration samples for multivariate modeling, *Chemometrics and Intelligent Laboratory Systems*, v. 72, p. 83-91, 2004.
- [79] GALVÃO, R. K. H.; ARAÚJO, M. C. U.; JOSÉ, G. E.; PONTES, M. J. C.; SILVA, E. C.; SALDANHA, T. C. B.; A method for calibration and validation subset partitioning, *Talanta*, v.67, p. 736-740, 2005.
- [80] CUADROS-RODRÍGUEZ, L.; PÉREZ-CASTAÑO, E.; RUIZ-SAMBLÁS, C.; Quality performance metrics in multivariate classification methods for qualitative analysis, *Trends in Analytical Chemistry*, v. 80, p. 612-624, 2016.
- [81] SOARES, S. F. C. **Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR**. 2016. 121 f. Tese (Doutorado em Química) – Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2009.
- [82] BALAT, M.; BALAT, H.; Recent trends in global production and utilization of bio-ethanol fuel, *Applied Energy*, v. 86, p. 2273-2282, 2009.
- [83] BASTOS, V. D.; Etanol, Alcoolquímica e Biorrefinarias, Biblioteca Digital do BNDS. Disponível em: < <https://web.bndes.gov.br/bib/jspui/handle/1408/2527> >. Acesso em 23 de março de 2017.
- [84] DA SILVA, A. C. **Metodologias analíticas para a identificação de não conformidades em amostras de álcool combustível**. 2013. 82 F. Dissertação (Mestrado em Química) – Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2013.
- [85] ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Resolução nº 19, de 15.04.2015. Disponível em: <<http://nxt.anp.gov.br/NXT/gateway.dll?f=templates&fn=default.htm&vid=anp:10.1048/enu>>. Acesso em 23 de março de 2017.
- [86] ANP, 10 dicas para fazer valer seus direitos ao abastecer. Disponível em: <<http://www.anp.gov.br/site/extras/defesaConsumidor/10dicas.asp>>. Acesso em 24 de março de 2017.
- [87] GOSSELIN, R. E.; HODGE, H. C.; SMITH, R. P.; GLEASON, M. N. Clinical Toxicology of Commercial Products. Williams & Wilkins, Baltimore, 1976.
- [88] NBR 15531, Álcool etílico-Determinação do teor de água - Método volumétrico de Karl Fischer, 2009.

- [89] NBR 15888, Etanol - Determinação do teor de água - Método coulométrico de Karl Fischer, 2010.
- [90] ASTM E203, Standard Test Method for Water Using Voltametric Karl Fischer Titration, 2008.
- [91] ABNT NBR 16041, Etanol combustível — Determinação dos teores de metanol e etanol por cromatografia gasosa, 2012.
- [92] ASTM D5501, Standard test method for determination of ethanol and methanol content in fuels containing greater than 20% ethanol by gas chromatography, 2013.
- [93] DANTAS, H. V.; BARBOSA, M. F.; NASCIMENTO, E. C. L.; MOREIRA, P. N. T.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; Screening analysis of natural gas with respect to methane content by near-infrared spectrometry, **Microchemical Journal**, v. 114, p. 210-215, 2014.
- [94] LUTZ, O. M. D.; BONN, G. K.; RODE, B. M.; HUCK, C. W.; Reproducible quantification of ethanol in gasoline via a customized mobile near-infrared spectrometer, **Analytica Chimica Acta**, v. 826, p. 61-68, 2014.
- [95] SILVA, A. C.; PONTES, L. F. B. L.; PIMENTEL, M. F.; PONTES, M. J. C.; Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, **Talanta**, v. 93, p. 129-134, 2012.
- [96] LERMA-GARCÍA, M. J.; RAMIS-RAMOS, G.; HERRERO-MARTÍNEZ, J. M.; SIMÓ-ALFONSO, E. F.; Authentication of extra virgin olive oils by Fourier-transform infrared spectroscopy, **Food Chemistry**, v. 118, p. 78-83, 2010.
- [97] BRANDÃO, P. A.; COSTA, F. G. P.; BARROS, L. R.; DO NASCIMENTO, G. A. J.; Ácidos graxos e colesterol na alimentação humana, **Revista Agropecuária Técnica**, v. 26, p. 5-14, 2005.
- [98] GÓMEZ-CARAVACA, A. M.; MAGGIO, R. M.; CERRETANI, L.; Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review, **Analytica Chimica Acta**, v. 913, p. 1-21, 2016.
- [99] International Olive Council, Trade standard applying to olive oils and olive-pomace oils, COI/T.15/NC N°3/Rev. 8. <<http://www.internationaloliveoil.org/estaticos/view/222-standards>>. Acesso em março de 2015.
- [100] CCFO - Codex Committee on Fats and Oils. FAO/WHO ID N°: CX-709. Disponível em: <<http://www.fao.org/fao-who-codexalimentarius/committees/committee-detail/en/?committee=CCFO>>. Acesso em 19 de março de 2017.
- [101] ANVISA - Agência Nacional de Vigilância Sanitária. Resolução RDC nº 270, de 22 de setembro de 2005. Disponível em: <<http://portal.anvisa.gov.br/legislacao#/visualizar/27630>>. Acesso em 19 de março de 2017.

- [102] MAPA – Ministério de Agricultura, Pecuária e Abastecimento. Instrução normativa 1/2012. Disponível em: <http://sistemasweb.agricultura.gov.br/sislegis/action/detalhaAto.do?method=visualizarAtoPortalMapa&chave=629707739>> Acesso e 19 de março de 2017.
- [103] GOUVINHAS, I.; DE ALMEIDA, J. M. M. M.; CARVALHO, T.; MACHADO, N.; BARROS, BARROS, A. I. R. N. A; Discrimination and characterisation of extra virgin olive oils from three cultivars in different maturation stages using Fourier transform infrared spectroscopy in tandem with chemometrics, **Food Chemistry**, v. 174, p. 226-232, 2015.
- [104] JORGE, R. O. **Caracterização de azeites virgem extra “Gourmet” varietais e “Blends” comercializados no mercado do Rio Grande do Sul**. 2010. 103 f. Tese (Doutorado em Ciências) – Departamento de Ciência e Tecnologia Agroindustrial, Universidade Federal de Pelotas, Pelotas, 2010.
- [105] JOHNSON, R. Food Fraud and “Economically Motivated Adulteration” of Food and Food Ingredients, Congressional Research Service, 2014. Disponível em: <https://www.fas.org/sgp/crs/misc/R43358.pdf>>, acesso em maio de 2015.
- [106] JIMÉNEZ-CARVELO, A. M.; PÉREZ-CASTAÑO, E.; GONZÁLEZ-CASADO, A.; CUADROS-RODRÍGUEZ, L.; One input-class and two input-class classifications for differentiating olive oil from other edible vegetable oils by use of the normal-phase liquid chromatography fingerprint of the methyl-transesterified fraction, **Food Chemistry**, v. 221, p. 1784-1791, 2017.
- [107] RUIZ-SAMBLÁS, C.; MARINI, F.; CUADROS-RODRÍGUEZ, L.; GONZÁLEZ-CASADO, A.; Quantification of blending of olive oils and edible vegetable oils by triacylglycerol fingerprint gas chromatography and chemometric tools, **Journal of Chromatography B**, v. 910, p. 71-77, 2012.
- [108] APETREI, I. M.; APETREI, C.; Detection of virgin olive oil adulteration using a voltammetric e-tongue, **Computers and Electronics in Agriculture**, v. 108, p. 148-154, 2014.
- [109] POULLI, K. I.; MOUSDIS, G. A.; GEORGIOU, C. A.; Rapid synchronous fluorescence method for virgin olive oil adulteration assessment, **Food Chemistry**, v. 105, p. 369-375, 2007.
- [110] MABOOD, F.; BOQUE, R.; FOLCARELLI, R.; BUSTO, O.; AL-HARRASI, A.; HUSSAIN, J.; Thermal oxidation process accelerates degradation of the olive oil mixed with sunflower oil and enables its discrimination using synchronous fluorescence spectroscopy and chemometric analysis, **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 143, p. 298-303, 2015.
- [111] DE SOUZA, L. M.; DE SANTANA, F. B.; GONTIJO, L. C.; MAZIVILA, S. J.; BORGES NETO, W.; Quantification of adulterations in extra virgin flaxseed oil using MIR and PLS, **Food Chemistry**, v. 182, p. 35-40, 2015.
- [112] GEORGOULI, K.; DEL RINCON, J. M.; KOIDIS, A.; Continuous statistical modelling for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman spectroscopic data, **Food Chemistry**, v. 217, p. 735-742, 2017.

- [113] VAN WETTEN, I. A.; VAN HERWAARDEN, A. W.; SPLINTER, R.; BOERRIGTER-EENLING, R.; VAN RUTH, S. M.; Detection of sunflower oil in extra virgin olive oil by fast differential scanning calorimetry, **Thermochimica Acta**, v. 603, p. 237-243, 2015.
- [114] ŠMEJKALOVÁ, D.; PICCOLO, A.; High-power gradient diffusion NMR spectroscopy for the rapid assessment of extra-virgin olive oil adulteration, **Food Chemistry**, v. 118, p. 153-158, 2010.
- [115] MIGNANI, A. G.; CIACCHERI, L.; OTTEVAERE, H.; THIENPONT, H.; CONTE, L.; MAREGA, M.; CICHELLI, A.; ATTILIO, C.; CIMATO, A.; Visible and near-infrared absorption spectroscopy by an integrating sphere and optical fibers for quantifying and discriminating the adulteration of extra virgin olive oil from Tuscany, **Analytical and Bioanalytical Chemistry**, v. 399, p. 1315-1324, 2011.
- [116] TORRECILLA, J. S.; CANCELLA, J. C.; MATUTE, G.; DÍAZ-RODRÍGUEZ, P.; FLORES, A. I.; Self-organizing maps based on chaotic parameters to detect adulterations of extra virgin olive oil with inferior edible oils, **Journal of Food Engineering**, v. 118, p.400-405, 2013.
- [117] SKOOG, D. A.; HOLLER, F. J.; NEIMAN, T. A. Princípios de Análise Instrumental. 5. Ed. Porto Alegre: Bookman, 2006.
- [118] XIAOBO, Z.; JIEWEN, Z.; POVEY, M. J. W.; HOLMES, M.; HANPIN, M.; Variables selection methods in near-infrared spectroscopy, **Analytica Chimica Acta**, v. 667, p. 14-32, 2010.
- [119] SOLOMONS, G.; FRYHLER, G. Química Orgânica. 8ª ed. Rio de Janeiro: LTC, editora LTDA, 2001.
- [120] PASQUINI, C.; Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications, **Journal of the Brazilian Chemical Society**, v. 14, p. 198-219, 2003.
- [121] DA SILVA, E. C. Métodos Espectroanalíticos, João Pessoa, 2010.
- [122] SIERRA, M. M. S.; GIOVANELA, M.; DONARD, O. F. X.; BELIN, C.; A utilização da espectroscopia de fluorescência no estudo da matéria orgânica dissolvida nas águas naturais: evolução e perspectivas, **Química Nova**, v. 19, p. 294-301, 1996.
- [123] MARQUES FILHO, O.; VIEIRA NETO, H. Processamento Digital de Imagens, Rio de Janeiro: Brasport, 1999.
- [124] ANTONELLI, A.; COCCHI, M.; FAVA, P.; FOCA, G.; FRANCHINI, G. C.; MANZINI, D.; ULRICI, A.; Automated evaluation of food colour by means of multivariate image analysis coupled to a wavelet-based classification algorithm, **Analytica Chimica Acta**, v. 515, p. 3-13, 2004.
- [125] DU, C. J; SUN, D. W.; Recent developments in the applications of image processing techniques for food quality evaluation, **Trends in Food Science & Technology**, v. 15, p. 230-249, 2004.

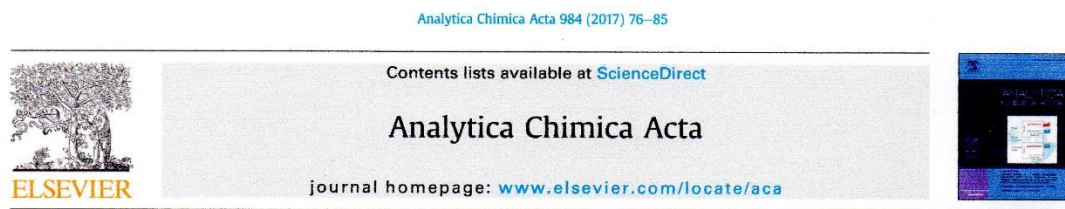
- [126] QUEIROZ, J. E. R.; GOMES, H. M. Introdução ao Processamento Digital de Imagens, **Revista RITA: instruções para preparação de documentos em Word**, 8 (1): 1, 2001.
- [127] SOLOMOM, C.; BRECKON, T. Fundamentals of Digital Image Processing - A Practical Approach with Examples in Matlab. 1st Edition. United States of America: John Wiley & Sons Ltd, 2011.
- [128] PROTESTE - Associação Brasileira de Defesa do Consumidor. Disponível em: <<http://www.proteste.org.br/azeite>>. Acessado em setembro de 2016.
- [129] O. Zenebon, N.S. Pascuet, P. Tiglea, Métodos físico-químicos para análise de alimentos, fourth ed., Instituto Adolfo Lutz, São Paulo, 2008.
- [130] GONZALEZ, R.C.; WOODS, R.E. Processamento de imagens digitais. Tradução Roberto Marcondes Cesar Júnior, Luciano da Fontoura Costa. 1^a Ed. São Paulo: Editora Blucher, 2000.
- [131] DE LIRA, L. F. B.; DE ALBUQUERQUE, M. S.; PACHECO, J. G. A.; FONSECA, T. M.; DE SIQUEIRA CAVALVANTI, E. H.; STRAGEVITCH, L.; PIMENTEL, M. F.; Infrared spectroscopy and multivariate calibration to monitor stability quality parameters of biodiesel, **Microchemical Journal**, v. 96, p. 126-131, 2010.
- [132] SILVA, A. C.; PONTES, L. F. B. L.; PIMENTEL, M. F.; PONTES, M. J. C.; Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, **Talanta**, v. 93, p. 129-134, 2012.
- [133] TARAKOWSKI, R.; MALANOWSKI, A.; KOŚCIESZA, R.; SIEGOCZYŃSK, R. M.; VIS spectroscopy and pressure induced phase transitions – Chasing the olive oils quality, **Journal of Food Engineering**, v. 122, p. 28-32, 2014.
- [134] YOUNG, A. J.; FRANK, H. A.; Energy transfer reactions involving carotenoids: quenching of chlorophyll fluorescence, **Journal of Photochemistry and Photobiology B**, v. 36, p. 3-15, 1996.
- [135] GUZMAN, E.; BAETEN, V.; PIERNA, J. A. F.; GARCIA-MESA, J. A.; Evaluation of the overall quality of olive oil using fluorescence spectroscopy, **Food Chemistry**, v. 173, p. 927-934, 2015.
- [136] MAGALHÃES, K. F.; CAIRES, A. R. L.; SILVA, M. S.; ALCANTARA, G. B.; OLIVEIRA, S. L.; Endogenous fluorescence of biodiesel and products thereof: Investigation of the molecules responsible for this effect, **Fuel**, v. 119, p. 120-128, 2014.
- [137] KONGBONGA, Y. G. M.; GHALILA, H.; ONANA, M. B.; MAJDI, Y.; LAKHDAR, Z. B.; MEZLINI, H.; SEVESTRE-GHALILA, S.; Characterization of Vegetable Oils by Fluorescence Spectroscopy, **Food and Nutrition Sciences**, v. 2, p. 692-699, 2011.
- [138] MILANEZ, K. D. T. M.; NÓBREGA, T. C. A.; NASCIMENTO, D. S.; INSAUSTI, M.; PONTES, M. J. C.; Transfer of multivariate classification models applied to digital images and fluorescence spectroscopy data, **Microchemical Journal**, v. 133, p. 669-675, 2017.

APÊNDICE

APÊNDICE

Produção científica

Os principais artigos científicos publicados em decorrência do trabalho realizado estão elencados a seguir. Demais produções científicas realizadas neste período podem ser visualizadas na Plataforma Lattes: <http://lattes.cnpq.br/1447222153074682>.



Selection of robust variables for transfer of classification models employing the successive projections algorithm



Karla Danielle Tavares Melo Milanez ^a, Thiago César Araújo Nóbrega ^a,
Danielle Silva Nascimento ^b, Roberto Kawakami Harrop Galvão ^c,
Márcio José Coelho Pontes ^{a,*}

^a Departamento de Química, Universidade Federal da Paraíba, João Pessoa, PB, Brazil

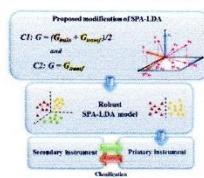
^b Laboratorio FIA, INQUISUR-CONICET, Departamento de Química, Universidad Nacional del Sur, Av. Alem 1253, B8000CPB Bahía Blanca, Buenos Aires, Argentina

^c Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos, São Paulo 12228-900, Brazil

HIGHLIGHTS

- Two new criteria for selection of robust variables for classification transfer.
- Use of linear discriminant analysis coupled to successive projections algorithm.
- Identification of adulteration in extra virgin olive oil and alcohol fuel.
- Comparison of the proposed criteria with DS and PDS methods.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:

Received 17 February 2017

Received in revised form

4 July 2017

Accepted 17 July 2017

Available online 22 July 2017

Keywords:

Multivariate classification transfer

Robust modeling

Successive projections algorithm

Standardization methods

UV–Vis spectroscopy

NIR spectroscopy

ABSTRACT

Multivariate models have been widely used in analytical problems involving quantitative and qualitative analyzes. However, there are cases in which a model is not applicable to spectra of samples obtained under new experimental conditions or in an instrument not involved in the modeling step. A solution to this problem is the transfer of multivariate models, usually performed using standardization of the spectral responses or enhancement of the robustness of the model. This present paper proposes two new criteria for selection of robust variables for classification transfer employing the successive projections algorithm (SPA). These variables are then used to build models based on linear discriminant analysis (LDA) with low sensitivity with respect to the differences between the responses of the instruments involved. For this purpose, transfer samples are included in the calculation of the cost for each subset of variables under consideration. The proposed methods are evaluated for two case studies involving identification of adulteration of extra virgin olive oil (EVOO) and hydrated ethyl alcohol fuel (HEAF) using UV–Vis and NIR spectroscopy, respectively. In both cases, similar or better classification transfer results (obtained for a test set measured on the secondary instrument) employing the two criteria were obtained in comparison with direct standardization (DS) and piecewise direct standardization (PDS). For the UV–Vis data, both proposed criteria achieved the correct classification rate (CCR) of 85%, while the best CCR obtained for the standardization methods was 81% for DS. For the NIR data, 92.5% of CCR was obtained by both criteria as well as DS. The results demonstrated the possibility of using either of the

* Corresponding author. Universidade Federal da Paraíba, Departamento de Química – Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), CEP 58051-970 João Pessoa, PB, Brazil.

E-mail address: marciocoelho@quimica.ufpb.br (M.J.C. Pontes).

<http://dx.doi.org/10.1016/j.aca.2017.07.037>

0003-2670/© 2017 Elsevier B.V. All rights reserved.

Produção científica

Microchemical Journal 133 (2017) 669–675



Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Transfer of multivariate classification models applied to digital images and fluorescence spectroscopy data



Karla Danielle Tavares Melo Milanez^a, Thiago César Araújo Nóbrega^a, Danielle Silva Nascimento^b,
Matias Insausti^b, Márcio José Coelho Pontes^{a,*}

^a Departamento de Química, Universidade Federal da Paraíba, João Pessoa, PB, Brazil

^b Laboratorio FIA, INQUISUR-CONICET, Departamento de Química, Universidad Nacional del Sur, Av. Alem 1253, B8000CPB Bahía Blanca, Buenos Aires, Argentina

ARTICLE INFO

Article history:

Received 15 December 2016

Received in revised form 23 February 2017

Accepted 4 March 2017

Available online 7 March 2017

Keywords:

Extra virgin olive oil

Adulteration

Fluorescence spectroscopy

Digital images

Classification transfer

ABSTRACT

This work evaluates the use of transfer of classification models for identifying adulteration of extra virgin olive oil (EVOO) samples involving, separately, two analytical techniques: fluorescence spectroscopy and digital imaging. The chemometric procedures, including development of classification models and application of classification transfer methods, were performed individually for each analytical technique. Methods of direct standardization (DS) and piecewise direct standardization (PDS) were applied to transfer samples sets in order to estimate an adjustment function and apply it to a samples set measured by the secondary instrument. For purposes of comparison, classification models were built based on linear discriminant analysis (LDA) with previous selection of variables by the successive projections algorithm (SPA), and partial least squares discriminant analysis (PLS-DA). The performance of the classification models was evaluated according to the number of errors and correct classification rate (CCR) for the prediction set measured by the secondary instrument. Before standardization, SPA-LDA and PLS-DA models achieved the same CCR using two analytical techniques: 54% for fluorescence emission spectra and 47% for histograms of digital images. After the standardization, a substantial increase of the CCR was observed. For the SPA-LDA models, a CCR of 88% was obtained for the fluorescence emission spectra and 82% for the histograms of the digital images. The PLS-DA classification models reached 85% and 76% of CCR for the fluorescence and imaging data, respectively, after standardization. These results demonstrate the efficiency of standardization procedures applied to multivariate classification models developed from fluorescence spectroscopy and digital images.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Multivariate classification models have been widely used in qualitative analysis involving conformity assessment and identification of adulteration in a variety of matrices [1–6]. Usually, the information obtained is used to discriminate samples into categories, according to similarity standards [7].

A number of supervised classification methods are reported in literature, including: soft independent modeling of class analogy (SIMCA) [8], k-nearest neighbors (k-NN) [9], partial least squares discriminant analysis (PLS-DA) [10] and linear discriminant analysis (LDA) [11]. Among these, LDA and PLS-DA have been constantly employed in classification problems [12–15]. Complications emerge when the classification models built from an instrument or laboratory condition cannot

correctly classify samples analyzed under new conditions or by a different instrument [16]. The differences between the spectral responses may be related to changes in the chemical and/or physical composition of the samples or changes in the instrumental response, normally caused by maintenance irregularities, repairs, changes in the environment of the instrument or even natural wear [17].

A solution to this problem is a full model recalibration from a data set acquired under the new analytical conditions. However, this requires time, money and rigorous control of the experimental conditions [18]. Alternatively, the classification transfer could be undertaken using standardization methods of spectral responses.

Standardization methods include direct standardization (DS) and piecewise direct standardization (PDS) [16]. In both cases, the response from a secondary instrument is adapted to correspond to the response from a primary instrument [19]. For this purpose, spectra of a representative samples set, known as *transfer samples*, are recorded in the two instruments and used in the standardization procedures [20].

The mathematical manipulations involved in transfer procedure can be carried out along the entire spectral range (in direct standardization)

* Corresponding author at: Universidade Federal da Paraíba, Departamento de Química – Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (IAQA), CEP 58051-970 João Pessoa, PB, Brazil.
E-mail address: marciocoelho@quimica.ufpb.br (M.J.C. Pontes).

Produção científica

Analytical
Methods

PAPER

Cite this: *Anal. Methods*, 2015, 7, 8839

Classification of extra virgin olive oil and verification of adulteration using digital images and discriminant analysis

Karla Danielle Tavares de Melo Milanez and Márcio José Coelho Pontes*

This work proposes a new methodology based on digital images and supervised pattern recognition methods for the classification of extra virgin olive oil (EVOO) samples with respect to brand (A, B and C) and verification of adulteration with soybean oil. For this purpose, information in the RGB color space was employed to develop classification models based on linear discriminant analysis (LDA) with previous selection of variables by the successive projections algorithm (SPA), and partial least squares-discriminant analysis (PLS-DA). The performance of classification models was evaluated according to the correct classification rate (CCR) for the prediction set. For the study involving the classification with respect to brand, a correct classification rate of 100% was achieved for both models (LDA and PLS-DA) when applied to the prediction set. The study involving the detection of adulteration with soybean oil in EVOO samples was carried out separately for each brand. For the A EVOO samples, a CCR of 88% and 94% was obtained by SPA-LDA and PLS-DA models for the prediction set, respectively. For the B and C EVOO samples, however, all prediction samples were correctly classified by both models. These results demonstrate the ability of the proposed method and present a promising alternative for the classification of extra virgin olive oil (EVOO) samples with respect to brand and verification of adulteration simply and quickly. Additionally, the method does not use reagents to carry out the analysis and laborious procedures for chemical characterization of the samples are not required.

Received 7th July 2015
Accepted 5th September 2015

DOI: 10.1039/c5ay01765c

www.rsc.org/methods

1. Introduction

Extra virgin olive oils (EVOOs) are often adulterated with vegetable oils of low quality and lower cost.¹ Obtained from the fruit of olive trees only by mechanical means and without the use of solvents, the EVOO has a low yield and high production costs, making adulteration a very attractive practice.² EVOOs have highly regarded nutritional properties mainly attributed to the high levels of some unsaturated fatty acids, which act positively on cholesterol levels and help protect against coronary heart and inflammatory diseases, as well as phenolic compounds known for their antioxidant activity.³ Besides the nutritional properties mentioned above, their sensory aspects are also appreciated, making it essential to verify the authenticity and quality of commercially available EVOOs.

Several studies have been reported in the literature using voltammetric e-tongue,⁴ fluorescence spectroscopy,^{5,6} Fourier transform infrared spectroscopy,^{7,8} differential scanning calorimetry,⁹ and nuclear magnetic resonance,¹⁰ among others^{11,12} for quality control of EVOOs. However, most of these methods

demand high acquisition and maintenance costs and involve laborious analysis.

Given this, techniques based on digital images are a promising alternative for the analysis of food and other products, especially as they are non-invasive, fast and cheap. Some studies have shown the availability of using the technique for quality control in various matrices, including potato chips,¹³ olive oils,^{14,15} cereal grains,¹⁶ edible vegetable oils,¹⁷ castor seeds,¹⁸ and vegetables,¹⁹ among others.^{20–23} Most of these studies use information from color models associated with pattern recognition techniques to cluster, or classify, the samples into categories, according to similarity standards.

Marchal *et al.*¹⁵ proposed a system based on computer vision to perform a fast approximate classification of the impurity content of olive oil samples at three different levels. For this, the authors tested three different initial input parameter vectors derived from the histogram of the channels of the RGB, CIELAB and HSV color spaces, and four different feature extraction methods.

Multivariate classification methods such as linear discriminant analysis (LDA)²⁴ and partial least squares-discriminant analysis (PLS-DA)²⁴ have been widely used in different analytical applications.^{25–27} PLS-DA is based on the standard PLS algorithm and class labels are used as the dependent *y* vector. In classification problems involving only two classes, the PLS

Universidade Federal da Paraíba, Departamento de Química, João Pessoa, CEP 58051-970, PB, Brazil. E-mail: marciocoelho@quimica.ufpb.br; Fax: +55 83 3216 7438; Tel: +55 83 3216 7438