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

TESE DE DOUTORADO

Novas estratégias para seleção de variáveis por intervalos em problemas de classificação

David Douglas de Sousa Fernandes

João Pessoa – PB - Brasil
Agosto/2016

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

TESE DE DOUTORADO

Novas estratégias para seleção de variáveis por intervalos em problemas de classificação

David Douglas de Sousa Fernandes*

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

Orientador: **Prof. Dr. Mário César Ugulino de Araújo**

* Bolsista CAPES

João Pessoa – PB - Brasil
Agosto/2016

F363n Fernandes, David Douglas de Sousa.
Novas estratégias para seleção de variáveis por intervalos em problemas de classificação / David Douglas de Sousa Fernandes.- João Pessoa, 2016.
137 f. : il.-

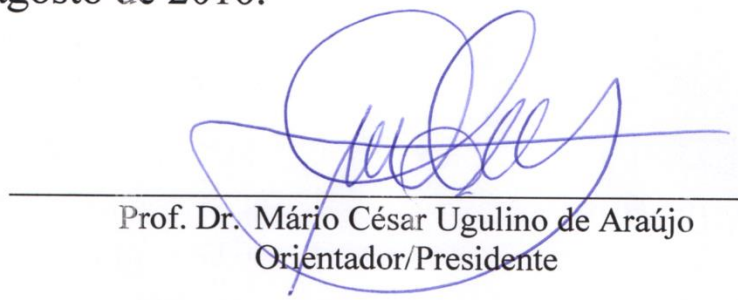
Orientador: Profº. Drº. Mário César Ugulino de Araújo.
Tese (Doutorado) – UFPB/CCEN

1.Química Analítica. 2. Seleção de Intervalos. 3. Algorítimo - Projeções Sucessivas. 4. Mínimos Quadrados Parciais – Análise Discriminante. 5. Modelagem Independente e Flexível – Analogia de Classe. I. Título.

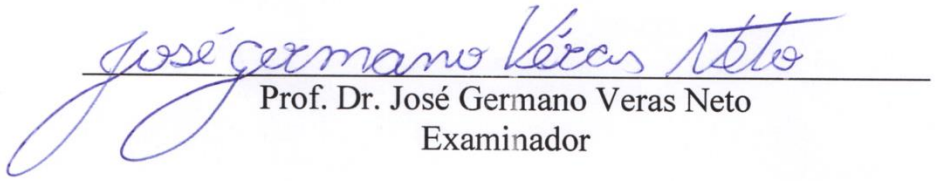
UFPB/BC CDU – 543(043)

Novas estratégias para seleção de variáveis por intervalos em problemas de classificação

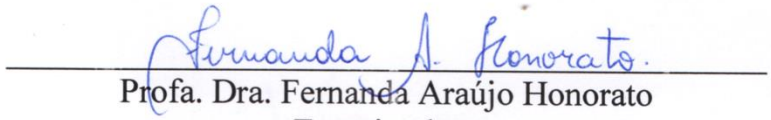
Tese de Doutorado apresentada pelo doutorando David Douglas de Sousa Fernandes e aprovada pela banca examinadora abaixo, em 26 de agosto de 2016.



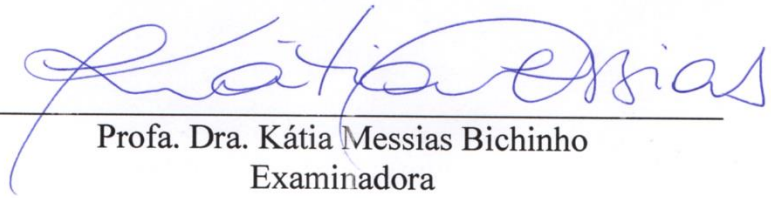
Prof. Dr. Mário César Ugulino de Araújo
Orientador/Presidente



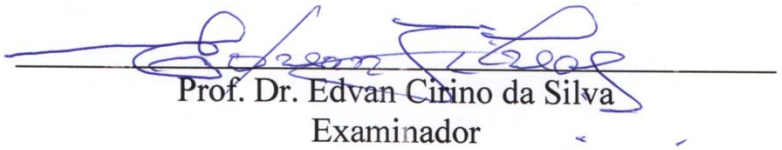
Prof. Dr. José Germano Veras Neto
Examinador



Profa. Dra. Fernanda Araújo Honorato
Examinadora



Profa. Dra. Kátia Messias Bichinho
Examinadora



Prof. Dr. Edvan Cirino da Silva
Examinador

O sucesso nasce do querer, da determinação e persistência em se chegar a um objetivo. Mesmo não atingindo o alvo, quem busca e vence obstáculos, no mínimo fará coisas admiráveis. " (José de Alencar)

A minha família, por sempre terem se dedicado e me ajudado ao máximo,

Dedico...

Agradecimentos

- A Deus, pela vida, saúde e discernimento para fazer este doutorado.
- A minha Mãe/Pai, Mércia Oliveira de Sousa e a minha irmã, Danyhelem de Sousa Fernandes, por sempre se dedicarem ao máximo na hora que sempre precisei, eternamente agradecido.
- A minha esposa Raissa Tavares Estavam Ramalho pela dedicação e apoio nos momentos difíceis.
- Ao Prof. Dr. Mário César Ugulino de Araújo pela confiança depositada, por todos os seus ensinamentos, pelas oportunidades de sempre aprender um pouco mais e principalmente por construir uma amizade verdadeira.
- Ao Prof. Roberto Kawakami Harrop Galvão pelas relevantes contribuições acadêmico-científicas para esta Tese.
- Aos Professores, Dr. Marcelo Fabián Pistonesi, Dra. Maria Centurión e Dra. Susana Di Nezio pela incrível oportunidade de crescimento profissional que tive durante realização do Doutorado-Sanduíche na Universidad Nacional del Sur, Argentina.
- A Adriano Araújo, Amanda Cecilia, Edilene Dantas, Emanuella Santos, Licarion Neto, Mayara Ferreira, Paulo Henrique, Welma Thaise e Wellington Lyra pela amizade, sinceridade, paciência, apoio e participação nos momentos importantes, pelos momentos únicos convividos e pela amizade construída.
- Aos Professores do PPGQ-UFPB pelos ensinamentos durante as disciplinas cursadas.
- A todos os amigos que fiz no LAQA/UFPB e na UNS.
- A Capes pelas bolsas concedidas.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

biPLS - Mínimos quadrados parciais por intervalos “em sentido contrário”, do inglês *Backward Interval Partial Least Squares*.

GA - Algoritmo genético, do inglês *Genetic Algorithm*.

HCA - Análise de agrupamentos hierárquicos, do inglês *Hierarchical Cluster Analysis*.

*i*PLS - Mínimos quadrados parciais por intervalos, do inglês *Interval Partial Least Squares*.

*i*SPA-PLS - Algoritmo das projeções sucessivas para seleção de intervalos em mínimos quadrados parciais, do inglês *Successive Projections Algorithm for Interval Selection in Partial Least-Squares*

LDA - Análise discriminante linear, do inglês Linear *Discriminant Analysis*.

MLR - Regressão linear múltipla, do inglês *Mutiple Linear Regression*.

NIR- Infravermelho próximo, do inglês *Near Infra Red*.

PC - Componentes principais, do inglês *Principal Components*.

PCA - Análise por componentes principais, do inglês Principal *Component Analysis*

PLS - Mínimos quadrados parciais, do inglês *Partial Least Squares*

PLS-DA - Mínimos quadrados parciais para análise discriminante, do inglês *Partial Least Squares Discriminant Analysis.*

RMSECV - Raiz quadrada do erro médio quadrático de validação cruzada, do inglês *Root Mean Square Error of Cross-Validation*

RMSEP - Raiz quadrada do erro médio quadrático de predição, do inglês Root *Mean Square Error of Prediction*

RMSEV - Raiz quadrada do erro médio quadrático de validação, do inglês *Root Mean Square Error of Validation*

SIMCA - Modelagem independente e flexível por analogia de classes, do inglês *Soft Independent Modeling of Class Analogy*

siPLS - Mínimos quadrados parciais por intervalos sinérgicos, do inglês *Synergy Interval-Partial Least-Squares.*

SPA - Algoritmo das projeções sucessivas, do inglês *Successive Projections Algorithm*.

RESUMO

Em Química Analítica tem sido recorrente na literatura o uso de sinais analíticos registrados em múltiplos sensores combinados com posterior modelagem quimiométrica para desenvolvimento de novas metodologias analíticas. Para esta finalidade, geralmente se faz uso de técnicas instrumentais multivariadas como a espectrometrias no ultravioleta-visível ou no infravermelho próximo, voltametria, etc. Neste cenário, o analista se depara com a opção de selecionar variáveis individuais ou intervalos de variáveis de modo de evitar ou diminuir problemas de multicolinearidade. Uma estratégia bem conhecida para seleção de intervalos de variáveis consiste em dividir o conjunto de respostas instrumentais em intervalos de igual largura e selecionar o melhor intervalo com base no critério de desempenho de predição de um único intervalo em regressão por Mínimos Quadrados Parciais (*i*PLS). Por outro lado, o uso da seleção de intervalo para fins de classificação tem recebido relativamente pouca atenção. Uma prática comum consiste em utilizar o método de regressão *i*PLS com os índices de classe codificados como variáveis de resposta a serem preditos, que é a idéia básica por trás da versão da Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) para a classificação. Em outras palavras, a seleção de intervalos para fins de classificação não possui o desenvolvimento de funções nativas (algoritmos). Assim, neste trabalho são propostas duas novas estratégias em problemas de classificação que usam seleção de intervalos de variáveis empregando o Algoritmo das Projeções Sucessivas. A primeira estratégia é denominada de Algoritmo das Projeções Sucessivas para seleção intervalos em Análise Discriminante por Mínimos Quadrados Parciais (*i*SPA-PLS-DA), enquanto a segunda estratégia é denominada de Algoritmo das Projeções Sucessivas para a seleção de intervalos em Modelagem Independente e Flexível por Analogia de Classe (*i*SPA-SIMCA). O desempenho dos algoritmos propostos foi avaliado em três estudos de casos: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade utilizando dados obtidos por voltametria de onda quadrada; classificação de misturas biodiesel/diesel não adulteradas (B5) e adulteradas com óleo de soja (OB5) empregando dados espectrais obtidos na região do ultravioleta-visível; e classificação de óleos vegetais com relação ao prazo de validade usando dados espectrais obtidos na região do infravermelho próximo. Os algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos forneceram bons resultados nos três estudos de caso, com taxas de classificação corretas sempre iguais ou superiores àquelas obtidas pelos modelos PLS-DA e SIMCA utilizando todas as variáveis, *i*PLS-DA e *i*SIMCA com um único intervalo selecionado, bem como SPA-LDA e GA-LDA com seleção de variáveis individuais. Portanto, os algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos podem ser consideradas abordagens promissoras para uso em problemas de classificação empregando seleção de intervalos de variáveis. Num contexto mais geral, a possibilidade de utilização de seleção de intervalos de variáveis sem perda da precisão da classificação pode ser considerada uma ferramenta bastante útil para a construção de instrumentos dedicados (por exemplo, fotômetros a base de LED) para uso em análise de rotina e de campo.

Palavras-chaves: Seleção de intervalos, Algoritmo das Projeções Sucessivas, Análise Discriminantes por Mínimos Quadrados Parciais, Modelagem Independente e Flexível por Analogia de Classe, Classificação.

ABSTRACT

In Analytical Chemistry it has been recurring in the literature the use of analytical signals recorded on multiple sensors combined with subsequent chemometric modeling for developing new analytical methodologies. For this purpose, it uses generally multivariate instrumental techniques as spectrometry ultraviolet-visible or near infrared, voltammetry, etc. In this scenario, the analyst is faced with the option of selecting individual variables or variable intervals so to avoid or reduce multicollinearity problems. A well-known strategy for selection of variable intervals is to divide the set of instrumental responses into equal width intervals and select the best interval based on the performance of the prediction of a unique range in the regression by Partial Least Squares (*i*PLS). On the other hand, the use of interval selection for classification purposes has received relatively little attention. A common practice is to use the *i*PLS regression method with the coded class indices as response variables to be predicted; that is the basic idea behind the release of the Discriminant Analysis by Partial Least Squares (PLS-DA) for classification. In other words, interval selection for classification purposes has no development of native functions (algorithms). Thus, in this work it is proposed two new strategies in classification problems using interval selection by the Successive Projections Algorithm. The first strategy is named Successive Projections Algorithm for selecting intervals in Discriminant Analysis Partial Least Squares (*i*SPA-PLS-DA), while the second strategy is called Successive Projections Algorithm for selecting intervals in Soft and Independent Modeling by Class Analogy (*i*SPA-SIMCA). The performance of the proposed algorithms was evaluated in three case studies: classification of vegetable oils according to the type of raw material and the expiration date using data obtained by square wave voltammetry; classification of unadulterated biodiesel/diesel blends (B5) and adulterated with soybean oil (OB5) using spectral data obtained in the ultraviolet-visible region; and classification of vegetable oils with respect to the expiration date using spectral data obtained in the near infrared region. The proposed *i*SPA-PLS-DA and *i*SPA-SIMCA algorithms provided good results in the three case studies, with correct classification rates always greater than or equal to those obtained by PLS-DA and SIMCA models using all variables, *i*PLS-DA and *i*SIMCA with a single selected interval, as well as SPA-LDA and GA-LDA with selection of individual variables. Therefore, the proposed *i*SPA-PLS-DA and *i*SPA-SIMCA algorithms can be considered as promising approaches for use in classification problems employing interval selection. In a more general point of view, the possibility of using interval selection without loss of the classification accuracy can be considered a very useful tool for the construction of dedicated instruments (e.g. LED-based photometers) for use in routine and *in situ* analysis.

Keywords: Interval selection, Successive Projections Algorithm, Discriminant Analysis Partial Least Squares, Soft Independent Modeling by Class Analogy, Classification.

INTRODUÇÃO

Capítulo 1

1. INTRODUÇÃO

Em Química Analítica, a maioria dos problemas analíticos requer a determinação quantitativa de uma ou mais substâncias presentes em uma amostra. Em outros casos, informações não quantitativas ou semi-quantitativas são requeridas: por exemplo, para autenticar um produto/substância ou verificar se uma substância está presente acima ou abaixo de um nível de concentração pré-estabelecido. Exemplos práticos incluem aplicações em metabolômica, onde dados analíticos provenientes de Ressonância magnética nuclear (NMR, do inglês *Nuclear Magnetic Resonance*) ou Cromatografia a gás acoplada com espectrometria de massas (GC–MS, do inglês *Gas Chromatography - Mass Spectrometry*) são usadas para determinar se uma amostra provém de um tecido saudável ou doente. Outro exemplo é dado em controle de processo estatístico multivariado, onde normalmente um espectro no infravermelho próximo é registrado e o objetivo é ver se o perfil espectral corresponde a um processo que se comporta satisfatoriamamente, ou seja, se uma dada batelada pode ser aceita ou não. Nesses casos, utilizar métodos qualitativos que forneçam uma resposta binária (positivo/negativo ou presença/ausência de um analito) ou ainda uma propriedade categórica pode ser adequada. Eles têm sido comumente usados em sistemas que exigem decisões imediatas a se tomar, uma vez que são uma alternativa atraente para análise quantitativa, o que geralmente fornece mais, mas muitas vezes desnecessárias, informações sobre a amostra e requer um maior investimento de dinheiro e/ou tempo. Em décadas recentes, os métodos qualitativos têm sido cada vez mais desenvolvidos e aplicados em áreas como medicina, biologia e química [1 2].

Os métodos qualitativos têm sido comumente diferenciados de acordo com a estratégia experimental de análise: *análise qualitativa tipo I*, que usa ferramentas que fornecem respostas binárias diretamente e envolvem a detecção (ausência/presença) e

identificação de espécies químicas; e a *análise qualitativa tipo II*, que converte dados brutos de qualquer sistema instrumental em respostas binárias e está relacionado à caracterização (identificação, classificação ou autenticação) da amostra como um todo [3-5]. A aplicação prática desta última requer o uso de Quimiometria, visto que tais sistemas instrumentais modernos são capazes de produzir, em curto tempo, grande quantidade de dados por amostra. Contudo, a influência de dados contendo informações redundantes e/ou não informativas (tais como regiões ruidosas ou saturadas) podem comprometer os resultados, porque nem todas as variáveis ou suas regiões são igualmente importantes para a modelagem. Nesse sentido, a seleção de variáveis é uma etapa importante em análise multivariada, dado que a remoção de variáveis não informativas geralmente produz melhores resultados e modelos mais simples. Tem sido amplamente aceito na literatura que a seleção de variáveis quando bem executada identifica um subconjunto de dados que produz os menores erros possíveis, resultando em modelos com uma maior capacidade preditora em determinações quantitativas ou discriminativa (discriminando entre amostras dissimilares) [6-8].

Do ponto de vista prático, dado um conjunto de sinais (variáveis) registrados em um arranjo de sensores para um grupo de amostras, o desafio é encontrar um subconjunto de variáveis representativo e não redundante capaz de promover melhores resultados quando comparado a um modelo desenvolvido empregando toda a informação registrada ou, ainda, conseguir, ao menos, resultados similares empregando um número menor de variáveis, em concordância com o princípio da parcimônia [9].

No contexto da classificação multivariada, também chamada de reconhecimento de padrões supervisionado, tem aumentando significativamente, na literatura, o número de publicações que combinam o uso de sinais analíticos registrados em múltiplos sensores para desenvolvimento de novas metodologias analíticas de interesse em diversas áreas da

ciência e tecnologia. Classicamente, diversos métodos de classificação multivariada que utilizam toda a informação instrumental, tais como Modelagem Independente e Flexível por Analogia de Classe (SIMCA, do inglês *Soft Independent Modeling by Class Analogy*), *k*-Vizinhos mais Próximos (*k*NN, do inglês *k-Nearest Neighbors*), Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA, do inglês *Partial Least Squares - Discriminant Analysis*) e Análise Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis*), por exemplo, têm sido utilizados para tal finalidade. Por outro lado, esforços consideráveis têm sido recentemente direcionados para o desenvolvimento e avaliação de diferentes procedimentos que possam identificar objetivamente variáveis e/ou regiões informativas, além de eliminar aquelas que contêm principalmente ruídos. São exemplos de técnicas de seleção de variáveis o Algoritmo Genético (GA, do inglês "*Genetic Algorithm*") [10], a Colônia de Formigas (AC, do inglês "*Ant Colony*") [11], o Stepwise (SW) [12] e o Algoritmo das Projeções Sucessivas (SPA, do inglês *Successive Projections Algorithm*) [13, 14]. Já a seleção de intervalos para fins de classificação não possui o desenvolvimento de funções nativas (algoritmos), sendo apenas utilizada como adaptação dos algoritmos de regressão em PLS [15-19].

Uma vez que os métodos de classificação multivariada podem ser divididos em algoritmos com dois principais fundamentos estatísticos relacionados a estratégias de discriminação (como PLS-DA, por exemplo) e modelagem de classe (como SIMCA, por exemplo) [20-22], neste trabalho são propostas duas novas estratégias em problemas de classificação para seleção de variáveis por intervalos baseado no Algoritmo das Projeções Sucessivas:

1) A primeira estratégia é denominada de Algoritmo das Projeções Sucessivas para seleção de intervalos em Análise Discriminante por Mínimos Quadrados Parciais (*i*SPA-PLS-DA). Nesse caso, a seleção dos intervalos ótimos a serem usados nos

modelos *i*SPA-PLS-DA é feita empregando a taxa de erro mínimo como função dos intervalos incluídos no modelo PLS-DA. Em adição, o limiar de cada classe é estabelecido levando-se em conta o melhor compromisso entre a especificidade e sensibilidade, isto é, quando esses parâmetros apresentam o mesmo valor.

2) A segunda estratégia é denominada de Algoritmo das Projeções Sucessivas para a seleção de intervalos em Modelagem Independente e Flexível por Analogia de Classe (*i*SPA-SIMCA), cuja seleção dos intervalos ótimos a serem usados nos modelos *i*SPA-SIMCA também é feita empregando a taxa de erro mínimo como função dos intervalos incluídos no modelo SIMCA.

Para avaliar o desempenho dos algoritmos propostos, três estudos de casos foram avaliados:

(i) classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade utilizando voltametria de onda quadrada;

(ii) identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal empregando espectrometria de absorção molecular Uv-Vis;

(iii) Triagem de óleos de soja com respeito ao estado de conservação utilizando espectroscopia do infravermelho próximo.

Para a avaliação do desempenho dos algoritmos propostos e das técnicas clássicas encontradas na literatura, foram empregadas a sensibilidade, a especificidade e a taxa de classificação correta, cujos fundamentos estão detalhados na **Seção 2.5**.

1.1. OBJETIVO

1.1.1. Objetivo Geral

Desenvolver novos algoritmos em ambiente Matlab utilizando o Algoritmo das Projeções Sucessivas (SPA) como ferramenta de seleção de intervalos para melhorar a capacidade classificatória de modelos de Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) e Modelagem Independente e Flexível por Analogia de Classes (SIMCA).

1.1.2. Objetivos Específicos

- ✓ Desenvolver os algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA;
- ✓ Aplicar os algoritmos propostos em três estudos de caso para: (i) classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade utilizando voltametria de onda quadrada; (ii) identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal empregando espectrometria de absorção molecular Uv-Vis; (iii) Triagem de óleos de soja com respeito ao estado de conservação utilizando espectroscopia do infravermelho próximo.
- ✓ Avaliar o desempenho dos algoritmos propostos em termos de sensibilidade, especificidade e taxa de classificação correta para os três diferentes estudos de casos frente aos métodos tradicionais de classificação multivariada encontrados na literatura, que utilizam todo o sinal analítico (PLS-DA e SIMCA), modelos com seleção de variáveis individuais (GA-LDA e SPA-LDA) e modelos com seleção de intervalos (*i*PLS-DA e *i*SIMCA).

FUNDAMENTAÇÃO TEÓRICA
Capítulo 2

2 MÉTODOS DE CLASSIFICAÇÃO MULTIVARIADA

Os métodos de classificação podem ser distinguidos de acordo com a complexidade dos dados analíticos usados para realizá-los: (i) métodos univariados, que requerem uma única medida específica, como, intensidade de sinal espectrométrico, tempo de retenção ou área de pico em cromatografia; (ii) métodos multivariados, os quais requerem múltiplas medidas ou sinais não específicos, tais como conjunto de resultados analíticos, perfil composicional e impressão digital instrumental; (iii) métodos multidimensionais ou multivias, que podem ser obtidos a partir do uso de técnicas analíticas hifenadas, como cromatografia bidimensional com detecção por espectrometria de massa (GCxGC-MS) ou matrizes excitação-emissão registradas a distintos valores de pH, por exemplo [23-31]. Na **Figura 1** está ilustrada a categorização dos dados analíticos comumente utilizados.

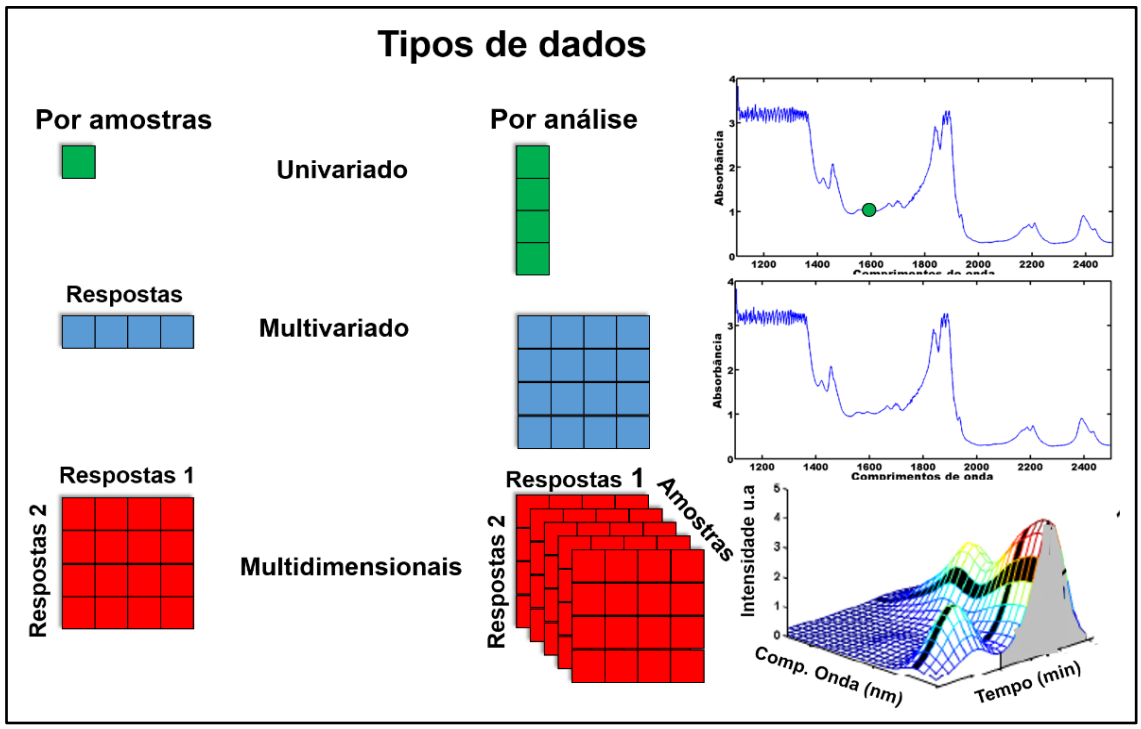


Figura 1 - Categorização dos dados analíticos. (Fonte própria)

O problema da classificação multivariada foi claramente expressado por Kowalsky: "Dada uma coleção de amostras caracterizadas por um conjunto de medidas feitas em

cada uma delas, o objetivo é encontrar e/ou predizer uma propriedade das amostras que não é diretamente mensurável em si ou é muito difícil de medir, mas é pensada para ser indiretamente relacionada com a as medições através de alguma relação desconhecida ou indeterminada". Em outras palavras, o principal objetivo de um método de classificação genericamente considerado é dividir as amostras de tal modo que cada uma delas seja atribuída a uma de uma série de categorias conhecidas como classes. O termo classe se refere a uma divisão de um conjunto de dados com, ao menos, uma característica, atributo, qualidade ou propriedade em comum. Já o termo conjunto de dados se refere aos dados com uma dimensionalidade maior do que um [3,32].

Um único número escalar é um tensor de ordem zero, enquanto um vetor de dados como, uma curva cinética ou um cromatograma é um tensor de primeira ordem. No mesmo sentido, uma matriz de dados (por exemplo, um espectro de fluorescência molecular total ou um espectro–cromatograma) é um tensor de dados de segunda ordem e assim sucessivamente. Em todos os casos, o algoritmo que implementa a classificação é conhecido como classificador, apesar desse termo também se referir algumas vezes ao modelo matemático construído com um algoritmo de classificação. As técnicas de classificação multivariada, portanto, definem as relações matemáticas entre um conjunto de variáveis descritivas um tensor de dados e uma variável qualitativa categórica, isto é, o membro de uma classe definida [3]. Nessas técnicas, o conhecimento prévio do conjunto de amostras ou das classes é requerido para identificar as amostras desconhecidas [33,34,26]. Uma diversidade de métodos de classificação multivariada é frequentemente reportada na literatura, destacando-se a Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA), a Modelagem Independente e Flexível Por Analogia de Classes (SIMCA) e a Análise Discriminante Linear (LDA) [33-36].

2.1 Análise Discriminante por Mínimos Quadrados Parciais

A regressão por mínimos quadrados parciais (PLS) foi originalmente empregada para tratamentos de dados destinados à calibração multivariada. Posteriormente, Barker e Rayens descreveram, pela primeira vez, o formalismo matemático da utilização de PLS para fins de classificação [37]. Brevemente, PLS-DA é um método de classificação onde a matriz ou vetor **y** da propriedade de interesse assume valores correspondentes aos índices de classe 0 e 1, quando existem duas classes. Para mais do que duas classes, pode-se construir vários modelos usando o algoritmo PLS2, onde cada coluna representa uma classe, atribuindo em cada coluna o valor 1 quando pertencer a uma dada classe e 0 quando não pertencer a esta mesma classe [38,39]. O número de variáveis latentes é normalmente escolhido usando validação cruzada para as amostras de treinamento. Na predição do modelo o que indicará se as amostras foram corretamente classificadas será o limiar adotado. Existem várias maneiras de determinar o limiar da classe, tais como o uso de reamostragem empregando intervalos de confianças via *bootstrap* [39], empregando o teorema Bayesiano [37] ou selecionando o ponto onde o número de falsos positivos e falsos negativos é mínimo [40].

2.2 Modelagem Independente e Flexível por Analogia de Classe

O método SIMCA foi proposto por Wold [41,42], sendo utilizado para classificação de amostras em conjuntos de dados com alta dimensionalidade. O espaço multidimensional fica delimitado através da construção de um modelo PCA para cada classe de amostras. Uma amostra será classificada como pertencente a uma dada classe previamente modelada, se possuir características que a permitam ser incluida no espaço multidimensional do modelo construído [26,44]. Matematicamente, a amostra será considerada pertencente à classe se o valor do *F* calculado, razão entre o valor de z^2 (**Equação 1**) e a variância da classe, for menor que o F crítico.

$$z^2 = x^2 + y^2 \quad (1)$$

onde z^2 é igual à soma da distância entre a amostra desconhecida e o eixo da PC (y^2) e a distância entre a projeção da amostra desconhecida na direção da PC e a fronteira da classe (x^2).

2.3 Análise Discriminante Linear

Essa técnica usa combinações lineares entre duas ou mais funções discriminantes. A discriminação ocorre determinando-se os pesos das variáveis independentes do "melhor conjunto de variáveis". Idealmente, para que a classificação de todas as amostras ocorra com sucesso em LDA, devem ocorrer simultaneamente a minimização da variância entre as amostras pertencentes à mesma classe e a maximização entre as amostras pertencentes a classes distintas [26, 45-46]. Contudo, há limitações no uso da LDA, uma vez que sua funcionalidade está restrita a conjuntos de dados de baixa dimensionalidade. Além disso, a capacidade de generalização de modelos LDA pode ser comprometida por problemas de colinearidade. Neste sentido, técnicas de seleção de variáveis são requeridas e têm sido utilizadas com sucesso quando associadas a modelos LDA [22, 47].

2.4 Seleção de variáveis

Em análise multivariada, a dimensionalidade dos dados é normalmente extensa ou apresenta uma relação amostras/variáveis desproporcional, além da presença de variáveis redundantes, não informativas e/ou ruidosas. Com isso, o desempenho de alguns métodos de classificação multivariada é severamente afetado. Para resolver essas limitações, técnicas de seleção de variáveis vêm sendo cada vez mais utilizadas. Essas técnicas baseiam-se no princípio de que um pequeno número de variáveis é capaz de gerar resultados sastifatorios por remover variáveis não informativas, além de minimizar a

multicolinearidade, que comprometem os resultados de métodos de regressão e de classificação multivariada [48,49].

As técnicas de seleção de variáveis podem ser categorizadas de acordo com a forma que o subconjunto de variáveis selecionado assume variáveis individuais ou intervalos de variáveis [48-50]. As técnicas de seleção variáveis individuais são mais comumente encontradas na literatura, seja acoplada a métodos de reconhecimento de padrões ou de calibração multivariada. No contexto da classificação multivariada, as técnicas de seleção de variáveis mais utilizadas têm sido o Algoritmo Genético (GA) e o Algoritmo das Projeções Sucessivas (SPA), descritos posteriormente na **Seção 2.4.1**. Já as técnicas de seleção por intervalos implementadas em funções nativas são inexistentes para fins de classificação, sendo apenas adaptadas por algoritmos de regressão, conforme descrito na **Seção 2.4.2**.

2.4.1 Seleção de variáveis individuais

2.4.1.1 Algoritmo das Projeções Sucessivas

O Algoritmo das Projeções Sucessivas (SPA) é uma técnica *forward* que realiza a seleção partindo de uma variável inicial x_k e, a cada iteração, vai incorporando uma nova variável com a menor multicolinearidade possível em relação aquelas já selecionadas [51]. O SPA foi inicialmente desenvolvido para resolver problemas de multicolinearidade em Regressão Linear Múltipla (MLR, do inglês *Multiple Linear Regression*) [51]. Posteriormente, foi estendido para selecionar intervalos de variáveis em PLS [50, 52].

No contexto da calibração multivariada em MLR, o SPA envolve essencialmente três etapas. Na primeira fase, operações de projeções são empregadas nas colunas da matriz de calibração e um subconjunto de variáveis é formado, levando-se em consideração o critério de minimização da multicolinearidade. Na segunda fase, é

escolhido o subconjunto que obtiver melhor resultado em relação ao critério que avalia a habilidade de previsão de um modelo MLR, a raiz quadrada do erro médio quadrático de validação (RMSEV, do inglês *Root Mean Square Error of Validation*) [51], de acordo com a **Equação 2**.

$$\mathrm{RMSEV}=\sqrt{\frac{1}{\mathrm{Kv}}\sum_{k=1}^{\mathrm{Kv}}(y_v^k-\hat{y}_v^k)^2} \quad (2)$$

onde Kv é número de amostras do conjunto de validação e y_v^k e $\hat{y}_v^k$ são os valores de referência e os valores preditos para o parâmetro de interesse nas amostras de validação. Na terceira fase, o subconjunto escolhido é submetido a um procedimento de eliminação para determinar se alguma variável pode ser removida sem perda significante da capacidade de predição.

O âmbito da classificação, o SPA foi usado para resolver problemas de multicolinearidade em Análise Discriminante Linear (SPA-LDA) [14,49] e tem sido empregado com sucesso em diversas aplicações analíticas [21, 53-54]. A diferença principal está na função de custo utilizada para guiar a classificação ao invés da RMSEV, conforme a **Equação 3**.

$$G=\frac{1}{\mathrm{Kv}}\sum_{k=1}^{\mathrm{Kv}}gK \quad (3)$$

onde gK é o risco de classificação incorreta da amostra de validação, que é definido na **Equação 4**.

$$gK=\frac{r^2(XK,\mu lK)}{\min lj\neq lK\, r^2(XK,\mu lj)} \quad (4)$$

Idealmente, o valor de gK deve assumir valores muito pequenos quando a amostra estiver perto do centro da sua verdadeira classe e distante dos centros das demais classes.

2.4.1.1 Algoritmo Genético

Na década de 60, John Holland propôs o Algoritmo Genético, cujo fundamento matemático simula o mecanismo biológico da Teoria da Evolução das Espécies, de Charles Darwin. Em sua primeira aplicação na área de Química, Lucasius e Kateman [55] utilizaram o GA para selecionar comprimentos de onda na região do ultravioleta para determinação de nucleotídeos. Atualmente, diversos trabalhos reportam a utilização do GA em diversas matrizes químicas e têm sido revisados na literatura [10,48].

No contexto da classificação, a função de custo do algoritmo GA-LDA corresponde ao inverso do risco G (**Equação 3**). Neste caso, o GA diferencia-se do SPA por gerar resultados estocásticos.

2.4.2 Seleção de variáveis por intervalos

Na literatura, o número trabalhos dedicados à seleção de variáveis por intervalos é relativamente pequeno. A motivação para selecionar variáveis em intervalos deve-se ao fato que a remoção de variáveis sem correlação com o parâmetro de interesse produz modelos mais simples, com menor erro sistemático quando comparado ao modelo qu utiliza todas as variáveis espectrais, além possuir maior estabilidade na etapa de predição [56,57].

No contexto das técnicas de reconhecimento de padrões não supervisionadas, a seleção de intervalos em Análise por Componentes Principais (PCA, do inglês *Principal Component Analysis*) tem sido empregada na análise exploratória de dados antes de empregar métodos de classificação, conforme descrito na **Seção 2.4.2.1**. Por outro lado, para fins de classificação multivariada, diversas técnicas de seleção de intervalos em regressão por PLS têm sido adaptadas e utilizadas em algumas aplicações analíticas, conforme descrito na **Seção 2.4.2.2**.

2.4.2.1 Análise por Componentes Principais em Intervalos

Dentre as técnicas de reconhecimento de padrões não supervisionadas, a Análise por Componentes Principais (PCA) destaca-se como uma da mais utilizadas devido à fácil interpretação dos resultados. A PCA reduz a dimensionalidade da matriz de dados empregando combinação linear das variáveis originais. Dessa forma, um novo sistema de eixos ortogonais entre si, denominado de componentes principais (PC, do inglês *Principal Components*), é formado [33,34,41,58]. A maior quantidade de informação possível é conservada nas PCs [35]. Assim, a PCA agrupa variáveis que estão altamente correlacionadas em novas variáveis, criando um conjunto que contém apenas as informações importantes 41,42].

Matematicamente falando, o cosseno do ângulo entre o eixo da variável e o eixo da PC é denominado de pesos (*loadings*) e as coordenadas das amostras no novo sistema de eixos das PCs são chamadas de escores (*scores*) [41,35].

Na seleção de intervalos em PCA o conjunto de dados é dividido em intervalos equidistantes, isto é, contendo o mesmo número de variáveis. Posteriormente, para cada intervalo é realizado um cálculo de PCA e o resultado do comportamento das amostras no intervalo selecionado é apresentado em um gráfico de escores [59].

Diniz e colaboradores realizaram uma análise exploratória aplicando *i*PCA em dados de espectrometria de absorção molecular Uv-Vis (190 a 1.100 nm) para verificação do comportamento de amostras de infusões de chás verdes e pretos de três diferentes origens geográficas. Eles concluíram que o intervalo de 251 a 490 nm era o mais adequado para realizar estudos posteriores de classificação simultânea do tipo e origem geográfica das infusões de chás, o qual alcançou 100% de classificação utilizando PCA-LDA e SPA-LDA [60].

Marder e colaboradores usaram espectroscopia no infravermelho próximo e médio para discriminar sete tipos de óleos comestíveis. *i*PCA foi realizada subdividindo-se os espectros em 8, 16 e 32 intervalos. Os autores afirmam que *i*PCA é uma ferramenta que possibilita encontrar regiões mais seletivas e promove uma melhor distinção dos óleos de acordo com sua composição em termos de cadeias saturadas e insaturadas [61].

2.4.2.2 Seleção de intervalos em PLS

Conforme dito anteriormente, diversas técnicas de seleção de intervalos em regressão por PLS têm sido adaptadas para fins de classificação multivariada [15-19]. Nestes casos, regressão em PLS por intervalos (*i*PLS, do inglês *Interval PLS*) [59] e PLS por intervalos "em sentido contrário" (*bi*PLS, do inglês *Backward Interval PLS*) [62] foram adaptados para realizar PLS-DA. Além dessas, outras estratégias de seleção de intervalos em PLS para calibração multivariada também estão disponíveis: PLS por Intervalos Sinérgicos (*si*PLS, do inglês *Synergy Interval PLS*) [63] e Algoritmo das Projeções Sucessivas para seleção de intervalos em PLS (*i*SPA-PLS, do inglês *Successive Projections Algorithm for Interval Selection in PLS)* [50]. Uma extensão do *i*SPA para seleção de intervalos em regressão PLS, proposto por Gomes e colaboradores, foi adaptado para ser utilizado em calibração multidimensional. Esse novo algoritmo associa o *i*SPA em modelos N-PLS [64] e U-PLS [65]. Em ambos os casos a bilinearização residual (RBL, do inglês *residual bilinearization*) é usado para alcançar a vantagem de segunda ordem. Os algoritmos foram denominados de N-*i*SPA-PLS/RBL e U-*i*SPA-PLS/RBL, respectivamente.

No caso de *i*PLS e *bi*PLS aplicados para fins de classificação, PLS-DA foi empregado usando-se uma variável de resposta codificada para cada categoria de amostras. A variável de resposta y é definida como segue: y = 1 para a amostra de

interesse em uma dada categoria e y = 0 para as demais categorias. Em seguida, *i*PLS e *bi*PLS foram aplicadas para otimizar os modelos PLS-DA. Os modelos de calibração foram desenvolvidos usando regressão PLS com validação cruzada completa. O número ótimo de fatores é então determinado pelos algoritmos baseado no menor número de fatores que fornece o menor valor da soma quadrática residual de predição (PRESS, do inglês *Predictive Residual Sum of Prediction*) na validação cruzada ou o menor valor da raiz do erro médio quadrático de validação cruzada (RMSECV, do inglês *Root Mean Square Error of Cross-Validation*). Contudo, essas abordagens não utilizam uma função de custo apropriada que leva em consideração o número real de amostras classificadas corretamente, além de não adotarem um critério matemático que defina o limiar (*threshold*) de cada classe.

A seguir, são descritas as técnicas de seleção de intervalos em PLS como concebidas para fins de calibração multivariada e multidimensional.

2.4.2.2.1 PLS por intervalos

O método *i*PLS consiste em encontrar uma região no conjunto de dados (faixa), que seja capaz de produzir resultados melhores ou minimamente iguais àqueles encontrados quando o conjumnto de dados originais é considerado. O procedimento consiste em dividir os dados em *n* intervalos de mesma dimensão. Na rotina disponível para aplicar esse método, a quantidade de intervalos é uma opção de entrada definida pelo usuário. Contudo, é preciso ter cuidado ao definir a quantidade de intervalos em que os dados serão divididos, pois um número pequeno de intervalos pode gerar uma faixa muito larga e esta, por sua vez, pode conter quantidades relevantes de informação desnecessária. Caso contrário, dividir os dados em intervalos muito estreitos pode levar a faixas pobres em informações úteis [51].

Após a divisão dos intervalos, o algoritmo *i*PLS calcula um modelo global empregando validação cruzada e um gráfico é gerado, mostrando o RMSECV em cada fator. O usuário decide, então, quantos fatores serão necessários para o cálculo do modelo *i*PLS. Posteriormente, para cada intervalo é calculado um modelo PLS também por validação cruzada e aquele intervalo que apresentar o RMSECV mais baixo é eleito como intervalo selecionado. A desvantagem do algoritmo *i*PLS é que a escolha do intervalo nunca será obtida se a solução que leva ao mínimo global for dois intervalos não sequenciados.

Suhandy e colaboradores utilizaram *i*PLS para selecionar intervalos em espectros terahertz, obtidos em espectrômetro de infravermelho com transformada de Fourier (FT-IR), para determinação de ácido ascórbico-L irradiado com alta e baixa frequência terahertz. Os resultados obtidos por *i*PLS foram superiores quando comparados com PLS completo [66].

Poppi e Borin utilizaram *i*PLS para quantificar contaminantes encontrados em óleos lubrificantes empregando espectroscopia do infravermelho médio. Segundo os autores, o algoritmo *i*PLS foi importante selecionando intervalos extremamente correlacionadas com as propriedades de interesse, levando a modelos mais simples e ajustados quando comparados aos modelos *full* PLS [67].

2.4.2.2.2 PLS por intervalos em sentido contrário

Leardi e Nørgaard empregaram seleção *backward* por intervalos em PLS. Nesse algoritmo, a etapa de divisão dos intervalos e a escolha do número ótimo de fatores PLS é semelhante à mencionada anteriormente para *i*PLS. Conhecendo o número ótimo de fatores PLS, a validação cruzada é usada e, a cada iteração, um conjunto de variáveis é eliminado do modelo de regressão. Ao excluir um conjunto de variáveis, idealmente o

RMSEP deve diminuir, justificando que essas variáveis são prejudiciais e, de fato, deve continuar fora do modelo. Quando a saída de um conjunto de variáveis elevar o RMSEP, o algoritmo é encerrado [62].

Xiaobo e colaboradores empregaram *bi*PLS para avaliar a qualidade de maçãs determinando o teor de sólidos solúveis utilizando espectrometria FT-NIR. O resultado de *bi*PLS obteve um RMSEP menor que o obtido para o modelo *full* PLS [68].

He e colaboradores usaram espectroscopia no infravermelho para determinação de polifenois em folhas de chá. Os resultados por *bi*PLS foram melhores que os obtidos por *full* PLS e *i*PLS [69].

2.4.2.2.3 PLS por intervalos sinérgicos

O s*i*PLS é uma variação do *i*PLS, em que é possível selecionar mais de um intervalo no modelo final otimizado. Semelhantemente ao que ocorre em *i*PLS e *bi*PLS, o número de divisões de intervalos é um critério de escolha do usuário. Por exemplo, o usuário decide dividir o conjunto de dados em vinte intervalos e, dessa forma, o algoritmo *si*PLS inicia as combinações de intervalos dois a dois, depois três a três, quatro a quatro, e assim sucessivamente [63]. Por sua vez, o modelo PLS é calculado para cada combinação de intervalos e o valor de RMSECV é guardado. Todavia, a medida que o número de intervalos e a quantidade de combinações aumentam, aproxima-se cada vez mais de uma busca exaustiva com variáveis individuais, o que demanda maior capacidade computacional.

Comparado com *i*PLS, o *si*PLS pode conduzir a melhores resultados quando o mínimo global menor RMSECV é encontrado com combinações de intervalos. Contudo, essa característica também se torna uma desvantagem, uma vez que se o usuário particiona os dados em vinte intervalos, por exemplo, e somente um dos vinte intervalos

contém informação relacionada com parâmetro de interesse, o algoritmo, ainda assim, selecionará mais de um intervalo.

Wu e colaboradores usaram espectrometria NIR na avaliação de extrato da erva medicinal chinesa Fu-fang Shuanghua. Nesse estudo, PLS, *i*PLS e s*i*PLS foram usados para determinar as concentrações de ácidos clorogênicos, ácidos fenólicos totais, flavonoides totais e sólidos solúveis. O *si*PLS foi responsável pelo modelo quantitativo mais ajustado, com alta correlação e baixo erro de predição, em todos os parâmetros estudados [70].

Ferrão e colaboradores utilizaram métodos de seleção de variáveis por intervalos *si*PLS, *i*PLS e PLS para quantificar simultaneamente o teor de biodiesel em mistura biodiesel/diesel, massa específica e teor de enxofre, usando refletância total atenuada no infravermelho médio. Em todas as propriedades estudadas, o algoritmo *si*PLS produziu modelos melhores, provavelmente por causa da sua capacidade de combinar intervalos que não são necessariamente adjacentes. Contudo, os autores destacaram que ambos os algoritmos *i*PLS e *si*PLS tiveram sucessos na seleção de faixa espectral mais adequada para cada propriedade [71].

2.4.2.2.4 PLS com seleção de intervalos pelo Algoritmo das Projeções Sucessivas

O *i*SPA-PLS é um algoritmo que seleciona intervalos otimizados em PLS e emprega o SPA para guiar a seleção de intervalos. O *i*SPA-PLS está projetado em duas fases. Inicialmente, é calculado o número ótimo de fatores para o modelo *full* PLS empregando o processo de validação por série de teste ou validação cruzada. Essa estratégia serve como estimativa inicial para que o usuário indique o número ótimo de fatores na etapa de seleção intervalos.

Na fase 1, a matriz de repostas instrumentais é centrada na média das colunas e o *i*SPA-PLS divide a matriz em *w* intervalos não sobrepostos, número este escolhido arbitrariamente pelo usuário, com base no número ótimo de fatores determinados previamente. Um requisito é que o número de variáveis *k* contido em cada intervalo tem que ser maior que número ótimo de fatores. Se o número de variáveis da matriz instrumental for divisível por *w*, então cada intervalo terá o mesmo número de variáveis. Caso *k* não seja divisível por *w*, o primeiro intervalo terá um maior número de variáveis. Por exemplo, no caso hipotético se a matriz for composta de 101 variáveis e o usuário definir a divisão da matriz em 5 intervalos, o primeiro intervalo conterá as primeiras 21 variáveis e todos os demais intervalos terão 20 variáveis. Em seguida, os intervalos são submetidos à etapa de projeção via SPA. Ao final da etapa de projeções, a matriz SEL com dimensão $(w - 1) \times w$ é obtida, onde as colunas de SEL contêm os índices das cadeias dos intervalos.

Na fase 2, cada cadeia de intervalos armazenados em SEL é usada na construção de modelos PLS, empregando validação cruzada ou validação por série de teste. Após avaliar todas as cadeias de intervalos, o que apresentar menor RMSECV ou RMSE é selecionado [50].

Diniz e colaboradores empregaram espectroscopia NIR em amostras comerciais de chás para determinação dos teores de umidade e polifenóis totais com PLS, *i*PLS e *i*SPA-PLS. O melhor resultado foi alcançado usando *i*SPA-PLS [52].

Lima e colaboradores utilizam PLS, *i*PLS *i*SPA-PLS e GA-PLS para determinar o teor de antocianinas totais em jabuticaba usando espectroscopia de refletância no infravermelho próximo. O *i*SPA-PLS alcançou os melhores resultados, mesmo comparado com aqueles obtidos para os demais métodos estudados [72].

2.4.2.2.5 PLS com intervalos selecionados pelo Algoritmo das Projeções Sucessivas em dados multidimensionais

As contribuições mais recentes a respeito das técnicas de seleção de variáveis por intervalos encontram-se no âmbito da calibração empregando dados multidimensionais. Gomes e colaboradores propuseram extensões do *i*SPA para seleção de intervalos em regressão PLS para serem aplicados em calibração multidimensional. Os novos algoritmos foram denominados N-*i*SPA-PLS/RBL [64] e U-*i*SPA-PLS/RBL [65].

O *i*SPA combinado com o N-PLS/RBL foi avaliado usando dados simulados e também na quantificação de ofloxacina em amostras de água. Os autores compararam os resultados com N-PLS/RBL sem seleção de variáveis e com GA-N-PLS/RBL. Como resultado, N-*i*SPA-PLS/RBL promoveu melhor acurácia quando comparado ao demais métodos avaliados [64].

Na segunda aplicação, o U-*i*SPA-PLS/RBL também foi submetido à avaliação de dados simulados e na quantificação de fenilefrina em amostras de água. Os resultados obtidos foram comparados com U-PLS/RBL sem seleção de variáveis e PARAFAC. U-*i*SPA-PLS/RBL melhorou o ajuste dos modelos U-PLS/RBL, além de produzir modelos isentos de desvios significativos [65].

2.5 Parâmeros de desempenho em classificação multivariada

Para avaliar o desempenho dos modelos de classificação é recorrente na literatura o uso de alguns parâmetros, como sensibilidade, especificidade e taxa de classificação correta. Normalmente, para se obter os valores referentes a esses parâmetros, os resultados obtidos pelos métodos de classificação multivariada são organizados na forma de tabela de confusão, contingência ou matriz de classificação. Nessa tabela, os resultados das classificações da classe verdadeira *versus* a classe atribuída são representados.

Normalmente, as amostras da classe verdadeira ou classe rotulada são representadas linhas da tabela. Nas colunas da matriz de confusão são representados os resultados das amostras que foram atribuídas pelo classificador do modelo [41-40, 73].

Os parâmetros de desempenho são um conjunto de atributos mensuráveis que têm de ser estabelecidos com as probabilidades que surgem a partir de quatro possíveis cenários: (i) amostras da classe A atribuídas corretamente à classe A (verdadeiro positivo – TP, do inglês *true positive*), (ii) amostras da classe B atribuídas corretamente à classe B (verdadeiro negativo – TN, do inglês *true negative*), (iii) amostras da classe A atribuídas erroneamente à classe B (falso positivo – FP, do inglês *false positive*) e, por fim, (iv) amostras da classe B atribuídas erroneamente à classe A (falso negativo – FN, do inglês *false negative*) [41-40, 73].

Seja n_1, n_2, ..., n_C o número de amostras em cada uma das classes C envolvidas no problema, a taxa de falsos negativos (FNR, do inglês *False Negative Rate*) e a taxa de falsos positivos (FPR, do inglês *False Positive Rate*) para a i-ésima classe são dadas nas **Equações 5** e **6**, como segue:

$$FNR_i = \frac{FN_i}{n_i} \tag{5}$$

$$FPR_i = \frac{FP_i}{\sum_{j \neq i} n_j} \tag{6}$$

onde FN_i e FP_i são o número de falsos negativos e falsos positivos obtidos para a i-ésima classe, respectivamente. Os valores de Sensibilidade (Sn_i) e a Especificidade (Sp_i) correspondentes são calculados de acordo com as **Equações 7** e **8**:

$$Sn_i = 1 - FNR_i \quad (7)$$

$$Sp_i = 1 - FPR_i \quad (8)$$

Já a Taxa de Classificação Correta (TCC) é calculada de acordo com a **Equação 9**.

$$\text{TCC (\%)} = \left(\frac{\text{Número de acertos de classificação}}{\text{Número total de amostras}}\right) \times 100\% \quad (9)$$

DESENVOLVIMENTO DOS ALGORITMOS

Capítulo 3

3.1 Desenvolvimento dos Algoritmos

Diferentemente do que vem sendo encontrado na literatura em aplicações analíticas que empregam adaptações de métodos regressão com seleção de intervalos para fins de classificação multivariada, principalmente utilizando algoritmos de regressão em PLS [15-19], nesta secção é apresentada a teoria detalhada dos dois algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos:

3.1.1 *i*SPA-PLS-DA e iSPA-SIMCA

Os algoritmos propostos nessa seção do trabalho é uma modificação do *i*SPA para seleção de intervalos em regressão PLS proposto por Gomes e colaboradores em 2012 [33]. Contudo, alterações foram realizadas para que o Algoritmo das Projeções Sucessivas possa atuar como ferramenta de seleção de intervalos em um método de classificação multivariada discriminante como o PLS-DA e na modelagem SIMCA. Os novos algoritmos foram nomeados de Algoritmo das Projeções Sucessivas para seleção de intervalos em Mínimos Quadrados Parciais para Análise Discriminante: *i*SPA-PLS-DA e Algoritmo das Projeções Sucessivas para seleção de intervalos em Modelagem Independente e Flexível por Analogia de Classe *i*SPA-SIMCA.

No algoritmo *i*SPA-PLS-DA o cálculo de um modelo PLS-DA empregando todas as variáveis originais é inicialmente usado para indicar ao usuário o número de fatores geralmente menor que o número total de amostras em estudo que será empregado. Da mesma maneira, mas buscando encontar o número de compenentes principais, um modelo SIMCA é feito no algoritmo iSPA-SIMCA. Como respostas, gráficos correspondentes ao número de fatores *versus* a taxa de erro ER, do inglês *Error Rate*, e uma saída gráfica indicando a melhor relação entre o número de componentes principais *versus* a taxa de erro para cada classe contida no banco de dados em estudo é apresentado ao usuário.

A construção dos modelos *i*SPA-PLS-DA e *i*SPA-SIMCA está compreendido, em duas fases. A fase 1 envolve a divisão das respostas instrumentais em intervalos de variáveis e a criação de diferentes combinações de intervalos, de acordo com uma sequência de operações de projeção geométrica. Essa fase é similar à fase 1 do algoritmo *i*SPA-PLS [50]. Na fase 2, a melhor combinação de intervalos é selecionada com base na menor taxa de erro gerada nos modelos PLS-DA e SIMCA resultante. A seguir serão tratados com mais detalhes as operações que ocorrem em cada fase do algoritmo proposto.

Fase 1 - *i*SPA-PLS-DA

Inicialmente, um modelo PLS-DA é aplicado aos dados da matriz de treinamento, utilizando validação cruzada para a determinação do número ótimo de fatores. O número ótimo de fatores é indicado levando em conta a minimização da taxa de erro. Após, o conjunto total de variáveis *k* é então dividido em intervalos não sobrepostos *w*. O número *w* pode ser livremente escolhido pelo usuário, desde que o número de variáveis em cada intervalo seja maior do que número ótimo de fatores.

Se *k* é divisível por *w*, cada intervalo terá o mesmo número de variáveis *k*/*w*. Caso contrário, da divisão *k*/*w* será atribuído a um ou mais intervalos. Adota-se que o restante das variáveis será distribuído igualmente entre os intervalos iniciais. Por exemplo, se $k = 102$ e $w = 5$, o primeiro e segundo intervalos compreenderão 21 variáveis e cada um dos três intervalos restantes compreenderá 20 variáveis.

Dentro de cada intervalo de variáveis, o vetor coluna de Xtrain com a maior norma é tomado como elemento representativo do referido intervalo. Os vetores de coluna *w* obtidos são reunidos em uma matriz denominada de Wtrain com dimensões (Ntrain × *w*), que é então utilizado como entrada para a sequência de operações de projeção envolvidas no algoritmo de SPA padrão. Como resultado, as combinações de intervalos de candidatos

são codificadas em uma matriz de índices SEL, com dimensão (w - 1) × w. Mais especificamente, a combinação de intervalos de m a partir do intervalo de ordem k é definido pelo conjunto de índices {SEL (1, k), SEL (2, k), ..., SEL (m, k)}.

Fase 2 – *i*SPA-PLS-DA

Na segunda fase do *i*SPA-PLS-DA, cada combinação de intervalos codificado na matriz SEL foi empregada na construção de um modelo PLS-DA. A validação cruzada foi novamente empregada, com o número de fatores variando de 1 até o valor obtido em *foptm* utilizado anteriormente no modelo PLS-DA. É considerada ótima a combinação de intervalos que possuir a maior taxa de classificação correta TCC. Caso o número de intervalos selecionados seja igual ao número de variáveis originais k, essa opção leva à resolução de um modelo PLS-DA tradicional, onde todas as variáveis são utilizadas no cálculo. Assim, para o emprego do algoritmo *i*SPA-PLS-DA proposto, k precisa variar entre 1 e (w - 1). Ao final do cálculo, é gerada uma figura correspondente aos intervalos selecionados pelo o algoritmo *i*SPA-PLS-DA, bem como, uma matriz em forma *string* contendo todos os resultados do modelo *i*SPA-PLS-DA, é apresentada na janela de trabalho do MatLab.

Fase 1 – *i*SPA-SIMCA

Após a etapa de escolha do número ótimo de PCs, a fase 1 é iniciada dividindo-se a matriz de respostas instrumentais em intervalos de variáveis, objetivando a criação de diferentes combinações de intervalos, de acordo com uma sequência de operações de projeção de vetores similar à fase 1 do algoritmo *i*SPA-PLS-DA proposto anteriormente. Matematicamente, o conjunto total de variáveis k é dividido em w intervalos não sobrepostos, onde a quantidade de intervalos w pode ser escolhida pelo analista, levando-

se em consideração que o número de variáveis em cada intervalo deve ser maior do que número ótimo de PCs. O critério de divisão das *k* variáveis parte do mesmo princípio empregado em *i*SPA-PLS-DA, ou seja, se *k* variáveis for divisível por *w*, cada intervalo terá o mesmo número de variáveis *k/w*. Caso contrário, o resto da divisão de *k/w* será de ser atribuída a um ou mais intervalos.

Para cada intervalo de variáveis, o vetor coluna na matriz de treinamento Xtrain que possuir a maior norma será escolhido como elemento representativo do referido intervalo. Dessa forma, os vetores coluna *w* obtidos são reunidos em uma matriz Wtrain com dimensões (Ntrain × *w*), que é então utilizado como entrada para a sequência de operações de projeção envolvidas no algoritmo SPA tradicional. Como resultado, as combinações de intervalos de candidatos são codificadas em uma matriz de índices SEL, com dimensão (*w* - 1) × *w*, cuja combinação de intervalos de *m* a partir do intervalo de ordem *k* é definido pelo conjunto de índices {SEL (1, *k*), SEL (2, *k*), ..., SEL (*m*, *k*)}.

Fase 2 – *i*SPA-SIMCA

No algoritmo *i*SPA-SIMCA, cada combinação de intervalos codificado na matriz SEL é empregada na construção de um modelo SIMCA. Para isto, validação cruzada é utilizada e os modelos SIMCA são calculados variando-se o número de componentes principais de 1 até o número ótimos de PCs. Será considerada como ótima a combinação de intervalos que possui a maior TCC. Ao final do cálculo, é gerada uma figura correspondente aos intervalos selecionados pelo o algoritmo *i*SPA-SIMCA, bem como uma matriz em forma *string* contendo todos os resultados do modelo *i*SPA-SIMCA apresentada na janela de comando do MatLab.

De forma resumida, um esquema do funcionamento dos algoritmos propostos *i*SPA-PLS-DA e *i*SPA-SIMCA é apresentado na **Figura 2.**

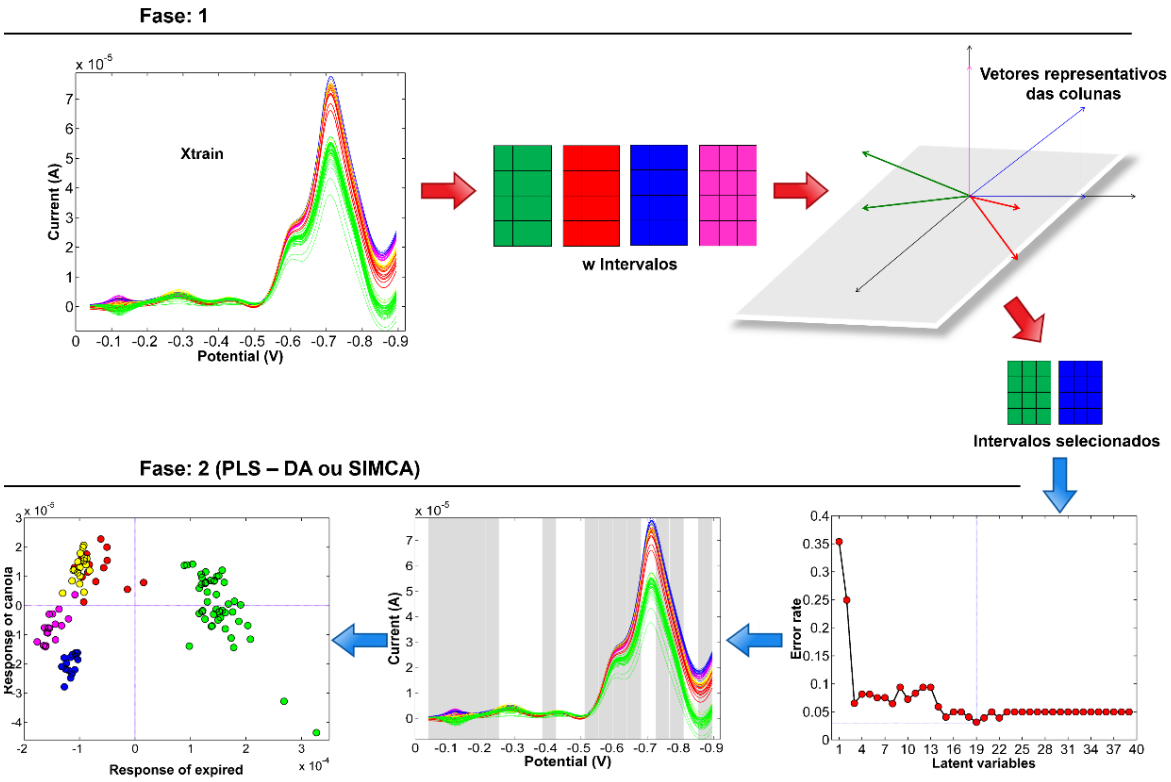


Figura 2 Esquema do funcionamento dos algoritmos propostos iSPA-PLS-DA e *i*SPA-SIMCA. (Fonte própria)

Para demonstrar a aplicabilidade dos algoritmos propostos, o desempenho do *i*SPA-PLS-DA e *i*SPA-SIMCA será avaliado frente aos métodos tradicionais de classificação multivariada que empregam toda informação analítica (PLS-DA e SIMCA), modelos com seleção de variáveis individuais (GA-LDA e SPA-LDA) e modelos com seleção de intervalo (*i*PLS-DA e *i*SIMCA). Para esta finalidade, o desempenho dos métodos foi avaliado em termos de sensibilidade, especificidade e taxa de classificação correta para três diferentes estudos de casos:

(i) Classificação de óleos vegetais com relação ao tipo de matéria-prima utilizando voltametria de onda quadrada;

(ii) Identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal empregando espectrometria de absorção molecular Uv-Vis;

(iii) Triagem de óleos de soja com respeito ao estado de conservação utilizando espectroscopia do infravermelho próximo.

Os cálculos de PLS-DA e SIMCA foram realizados utilizando o pacote *Classification Toolbox 4.0* [40], enquanto GA-LDA e SPA-LDA foram executados empregando as rotinas desenvolvidas por Carneiro et al [74] e Pontes et al [14]. A rotina de GA foi executada para 100 gerações com 200 cromossomos em cada geração e, além disso, cruzamento e mutação foram definidos com probabilidades de 60 e 10%, respectivamente. Devido à natureza estocástica de GA, dez repetições do cálculo foram realizadas e, em seguida, o melhor resultado foi adotado levando em consideração a combinação de variáveis que minimizava a taxa de erro. Já os algoritmos para *i*PLS-DA e *i*SIMCA foram construídos de acordo com os mesmos critérios matemáticos e estatísticos implementados para *i*SPA-PLS-DA e *i*SPA-SIMCA, mas sem a etapa de projeções para seleção de intervalos pelo Algoritmo das Projeções Sucessivas. Para divisão das amostras em conjuntos de treinamento e teste foi utilizado o algoritmo Kennard-Stone (KS) [75]. Todos os métodos utilizam linguagem escrita em ambiente de MatLab®.

Classificação de óleos vegetais

Capítulo 4

4. CLASSIFICAÇÃO DE ÓLEOS VEGETAIS COM RELAÇÃO AO TIPO DE MATÉRIA-PRIMA UTILIZANDO VOLTAMETRIA DE ONDA QUADRADA

4.1 Apresentação

Neste estudo de caso é avaliada a classificação de óleos vegetais de acordo com o tipo de oleoaginosa/estado de conservação (canola, milho, girassol, soja e expirado) utilizando dados de voltametria de onda quadrada. O banco de dados é composto 66 amostras de canola (CA), milho (MI), girassol (GI) e soja (SO), aqui essas amostras são denominadas não expiradas, e 48 amostras de óleos expirados (EX) de diferentes tipos de oloeaginosas, lotes e marcas. A classe de amostras denominadas expiradas foi obitida armazenando as amostras de óleos em seus frascos comerciais originais, sem controle das condições ambientais entre 12 e 18 meses apois a data de validade estipulada no rotulo das embalagens. A quantidade de amostras analisadas por classe é mostrada em detalhes na **Tabela 1**.

Tabela 1: Quantidade de amostras de óleos vegetais analisadas por classe

Classe de óleo	Quantidade de amostras
Canola - CA	15
Girassol -GI	16
Milho - MI	17
Soja - SO	18
Expirado - EX	48

4.2 Experimental

Um processo de extração de ácidos graxos adaptado do método oficial CA 5a-40, da AOCS *The American Oil Chemists' Society* foi realizado em todas as amostras. O processo de extração se deu misturando sob agitação mecânica, alíquotas de 200 µL de óleo com 200 µL de etanol em tubos de propileno de 1,5 mL. Em seguida, a mistura foi

mantida em repouso por trinta minutos para a formação sobrenadante translúcido contendo, majoritariamente ácidos graxos livres. Posteriomente, uma alíquota de 10 µL do sobrenadante foi adicionada à cela eletroquímica juntamente com 10 mL de NaOH 0,1 mol L^{-1}, agitando-se por 60s com o eletrodo rotatório para homogeneizar. Nestas condições, o voltamograma de cada amostra foi registrado em triplicata empregando varredura catódica entre -0.9 a 0 volts com o auxilio de um potenciostato/galvanostato µAutoLab® Tipo II (Eco Chemie), acoplado a um módulo polarográfico 663 VA Stand® (Metrohm) composto com um eletrodo de referência de Ag/AgCl–KCl (3,0 mol L^{-1}) e um fio de platina como contra eletrodo.

4.3 Resultados e discussão

Na **Figura 3a** é mostrado os voltamogramas das 114 amostras de óleos vegetais estudados. Como pode ser observado, os voltamogramas correspondente as amostras da classe expirada, exibem valores de corrente relativamente menor das demais amostras das classes estudas, isto pode está associado à medida que os óleos envelhecem há um aumento na concentração de ácidos graxos livres e, como resultado, existe a possiblidade de maior formação de sabão dentro da cela eletroquímica causando uma depleção do sinal analítico, portanto, há uma tendência de separação entre as amostras expiradas das demais amostras não expiradas.

Analisando o perfil voltametrico de todas as amostras, uma leve separação pode ser observada entre as amostras de canola e milho, das amostras soja e girassol em torno dos potenciais -0.9 a -0.81V e de -0.15 a -0.09V respectivamente. Também pode ser observado uma pequena diferença entre o perfil do voltametrico da classe canola em relação às demais entre os potenciais -0.580 V a -0.650 V. Esta informação pode ser melhor visualizada nos voltamogramas médios de cada classe apresentada na **Figura 3b**.

Contudo é praticamente impossível por inspeção visual a distinção entre todos os tipos de óleos vegetais. Dessa forma, o uso de ferramentas quimiométricas tornou-se indispensável para essa tarefa.

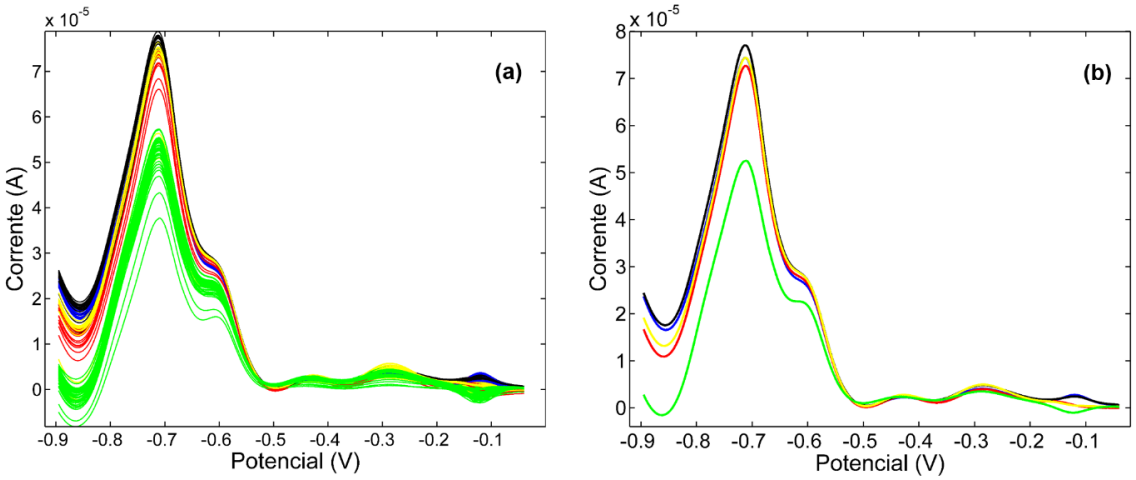


Figura 3 – (a) Voltamogramas das 114 amostras de óleos vegetais. (b) Voltamogramas médio amostras de óleos vegetais estudadas (canola, girassol, milho, soja e expirada).

4.3.1 – Análise por componentes principais

A análise exploratória dos dados foi realizada usando PCA afim de encontrar formação de agrupados entre as diferentes classes estudadas. Na **Figura 4a** e **Figura 4b** é apresentado os escores de PCA para as componentes: PC1 versus PC2 e PC3 versus PC4, respectivamente.

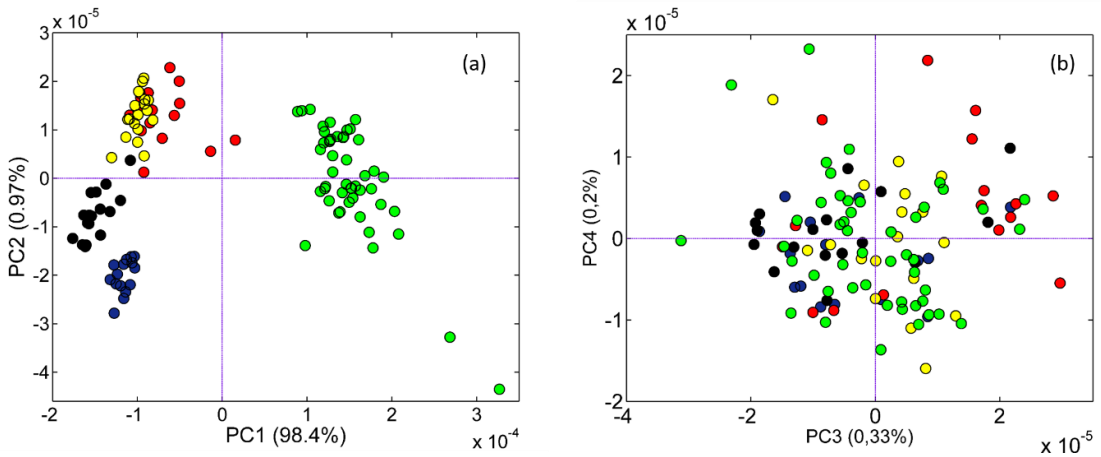


Figura 4 – (a) Gráficos de escores de PC1 x PC2. (b) Gráficos de escores de PC3 x PC4 para as 114 amostras de óleos vegetais estudadas (canola, girassol, milho, soja e expirada).

Aplicando PCA no conjunto de dados, uma distinção clara entre as amostras óleos expirados e não expirados podem ser vistos ao longo de PC1 (**Figura 10a**). Com relação as amostras não expiradas, três agrupamentos podem ser observados; dois destes formados por amostras de canola e milho e um último formado pela sobreposição parcial das classes girassol e soja (**Figura 10a**). Portanto, torna-se necessário o desenvolvimento de modelos baseados em métodos de reconhecimento de padrão supervisionado.

4.3.2 – Formação dos conjuntos de dados

O conjunto de dados foi então dividido em conjuntos de treinamento com 76 amostras (10 canola, 11 girassol, 12 milho, 13 soja e 30 expiradas) e, teste com 38 amostras (5 canola, 5 girassol, 5 milho, 5 soja e 18 expiradas) empregando algoritmo Kernnard-Stone [75].

4.4 - Classificação

4.4.1. – PLS-DA

O modelo PLS-DA foi construído para as cinco classes de óleos estudadas usando validação cruzada completa. O número de fatores ótimos a ser usado no modelo foi determinado com base na função da taxa de erro, e com auxilio dos gráficos dos coeficientes de regressão (ver apendece 1). Dezenove fatores corresponde a menor taxa de erro para o modelo PLS-DA, o perfil da taxa de erro em virtude da inclusão de fatores ao modelo é apresentado na **Figura 5**.

Para atribuição das amostras a uma dada classe, foi adotatado o limiar usando a função multi objetiva ROC do inglês, *Receiver Operating Characteristics*.

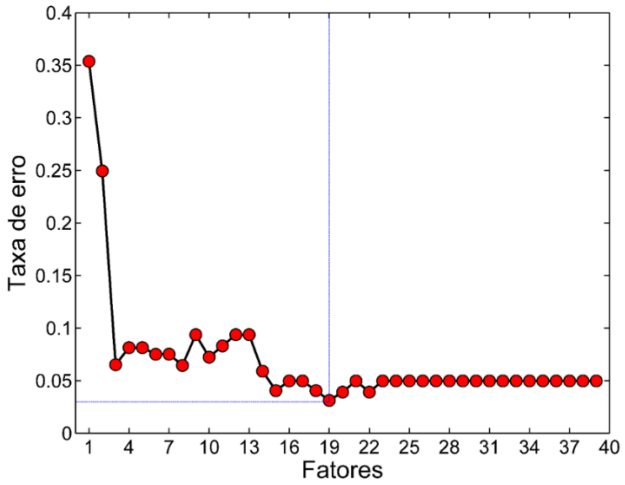


Figura 5: Número de fatores selecionados para o modelo *full* PLS-DA

O resultado da classificação PLS-DA é apresentado na **Tabela 2** na forma de matriz de confusão. O desempenho do modelo é mostrado em termos dos valores da especificidade, sensibilidade, % TCC para o modelo global e conjuntos de validação cruzada e teste.

Tabela 2: matriz de confusão para classificação PLS-DA

		Validação cruzada						Test					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	N_{teste}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	4	-	1	-	-
	GI	11	-	9	-	2	-	5	-	5	-	-	-
	MI	12	-	-	11	1	-	5	-	-	5	-	-
	SO	13	-	-	-	13	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	18
Sensibilidade			1	0.82	0.92	1	1		0.80	1	1	1	1
Especificidade			1	1	1	0.95	1		1	1	0.97	1	1
TCC (%) conjunto			96						97.4				
TCC (%) modelo			96.7										

CA: classe canola**, GI:** classe girassol**, MI:** classe milho**, SO:** classe soja**, EX:** classe expirado**, N_{cv}:** número de amostras por classe, para o conjunto de validação cruzada. N_{teste}: número de amostras para o conjunto de teste.

Como pode ser visto na **Tabela 2**, o conjunto de validação cruzada resultou em três amostras mal classificadas obtendo uma TCC correspondente a 96%. Uma única amostra foi erroneamente classificada no conjunto de teste, a TCC alcançada nesse conjunto correspondeu a 97,4%, além disto, apenas as amostras pertencentes a classe dos óleos expirados atingiu valor máximo em termos da especificidade e sensibilidade

4.4.2 – Classificação SIMCA

Modelos SIMCA foram construídos individualmente para classe de óleo estudada, o número de *PC's* exigidos em cada classe foi determinda tomando como referência, o ponto correspondente a menor taxa de erro em virtude da inclusão de novas componentes principais em cada classe.

O desempenho do SIMCA é apresentado na **Tabela 3** em termos dos valores da especificidade, sensibilidade, % TCC correspondente ao modelo global, conjunto de validação cruzada e teste, bem como matriz de confusão para as classes em estudo.

Tabela 3: matriz de confusão para classificação SIMCA

		Validação cruzada						**Test**					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	N_{teste}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	5	-	-	-	-
	GI	11	-	9	-	2	-	5	-	3	-	2	-
	MI	12	-	-	11	1	-	5	-	-	5	-	-
	SO	13	-	1	-	12	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	18
Sensibilidade			1	0,82	0,92	0,92	1		1	0,60	1	1	1
Especificidade			1	0,98	1	0,95	1		1	1	1	0,94	1
TCC (%) conjunto			94,7						94,7				
TCC (%) modelo			94,7										

CA: classe canola**, GI:** classe girassol**, MI:** classe milho**, SO:** classe soja**, EX:** classe expirado**, N_{cv}:** número de amostras por classe, para o conjunto de validação cruzada. N_{teste}: número de amostras para o conjunto de teste.

O desempenho em termos da taxa de classificação correta para o modelo global SIMCA correspondeu a 94,7%. No total quatro amostras foram mal classificadas no conjunto de validação cruzada, e apenas nas classes canola e expirados os valores correspondente a sensibilidade eespecificidade foram máximos. No conjunto teste, o máximo valor para sensibilidade, especificidade e taxa de classificação correta ocorreram apenas nas classes canola, milho e expirados.

4.4.3 – SPA-LDA

A **Figura 6a** apresenta o voltamograma médio das 114 amostras com a indicação dos doze potenciais selecionados pelo SPA-LDA. Observa-se que os potenciais selecionados pelo SPA-LDA se encontram distribuídos em regiões correspondentes a picos dispostos ao longo de todo voltamograma. A **Figura 6b** mostra os gráficos das funções discriminantes obtidas pelo SPA-LDA. Uma excelente discrimação entre amostras das classes expirados, milho e canola pode ser visualizada na **Figura 6b**, contundo, as amostras das classes girassol e soja são mais próxima entre si, resultado em um erro apresentado na classificação como apresentado na **Tabela 4**.

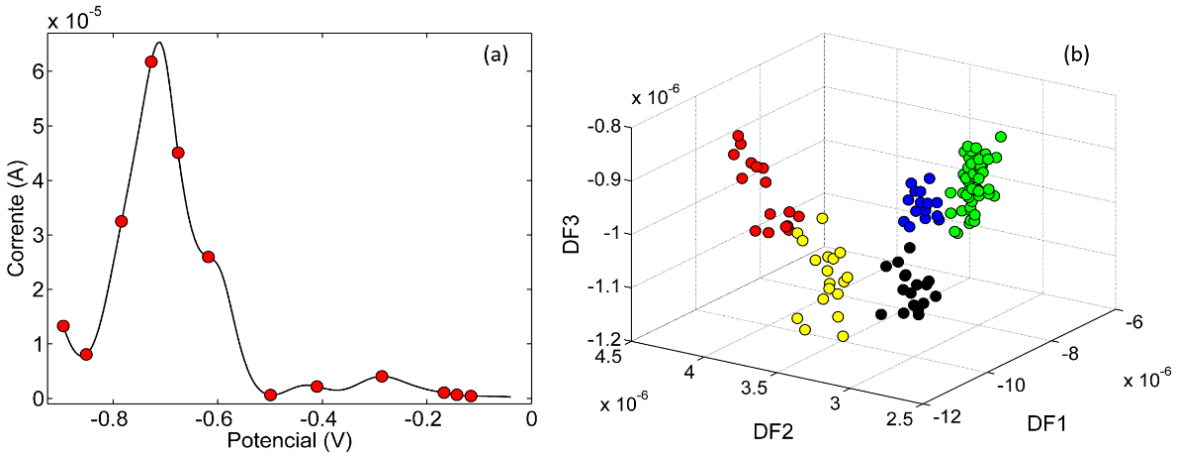


Figura 6 - (a) variáveis selecionadas por SPA-LDA e (b) gráfico de escores para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (● canola, ● girassol, ● milho, ● soja e ● expirada).

A **Tabela 4** apresenta o resultado da classicação obtida pelo modelo SPA-LDA para os conjuntos de validação cruzada e teste.

Tabela 4 - Matriz de confusão para classificação SPA-LDA para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

		Validação cruzada						Teste					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	N_{teste}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	5	-	-	-	-
	GI	11	-	11	-	-	-	5	-	4	-	1	-
	MI	12	-	-	12	-	-	5	-	-	5	-	-
	SO	13	-	-	-	13	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	18
Sensibilidade			1	1	1	1	1		1	0,80	1	1	1
Especificidade			1	1	1	1	1		1	1	1	0,97	1
TCC (%) conjunto			100						97,4				
TCC (%) modelo			98,7										

CA: classe canola**, GI:** classe girassol**, MI:** classe milho**, SO:** classe soja**, EX:** classe expirado**, N_{cv}:** número de amostras por classe, para o conjunto de validação cruzada. N_{teste}: número de amostras para o conjunto de teste.

O desempenho do modelo SPA-LDA para o conjunto de validação cruzada atingiu uma TCC de 100%, consequentemente máximo valores de sensibilidade e especificidade foram obtidos para as cinco classes nesse conjunto. Contudo no conjunto teste, uma amostra da classe girassol foi mal atribuida, e classificada como pertencente a classe soja, resultando em uma taxa de classificação correta igual a 97,4%.

4.4.4 – GA-LDA

O modelo GA-LDA resultou na seleção de seis potenciais que pode ser visualizado no voltamograma médio das 114 amostras apresentado na **Figura 7a**. Comparando o número de variáveis selecionadas nos modelos SPA-LDA e GA-LDA, o modelo GA-LDA apresenta-se mais parcimonioso, contudo, as variáveis selecionadas nesse modelo resultam em potenciais localizados em regiões com pouca intensidade de corrente,

resultando em uma tendência de separação menos efetiva como pode ser visto nas funções discriminantes apresentadas na **Figura 7b**.

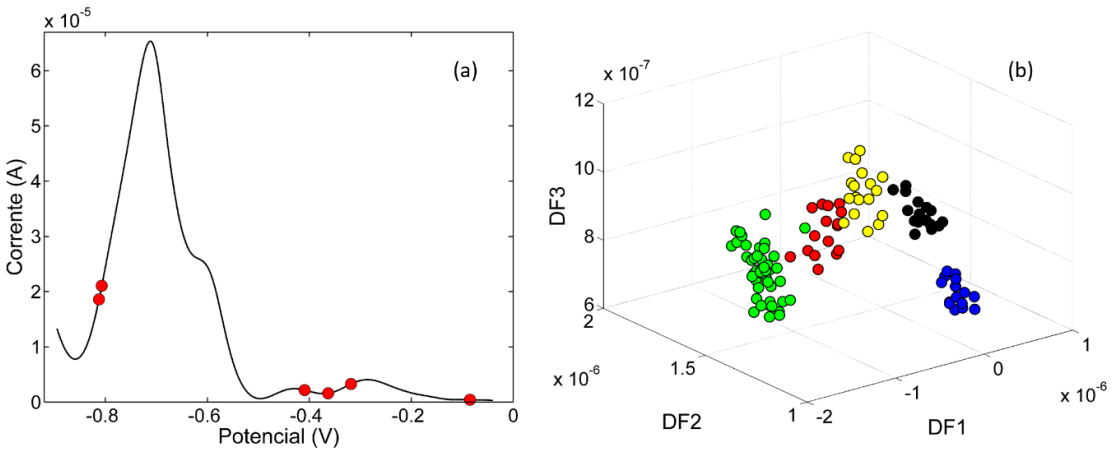


Figura 7 - (a) Variáveis selecionadas por GA-LDA e (b) gráfico de escores para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (● canola, ● girassol, ● milho, ● soja e ● expirada) Fonte própria.

O resultado obtido no modelo GA-LDA em detalhes é apresentado na **Tabela 5** na forma da matriz de confusão, taxa de classificação correta, sensibilidade e especificidade. O modelo LDA obtido apartir dos potenciais selecionadas pelo GA, resultou na classificação correta de todas as amostras do conjunto de teste. Contudo, três amostras foram mal classificadas no conjunto de validação cruzada.

Tabela 5- Matriz de confusão para classificação GA-LDA para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas

		Validação cruzada						Test					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	Nt_{este}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	5	-	-	-	-
	GI	11	-	10	-	1	-	5	-	5	-	-	-
	MI	12	-	-	11	1	-	5	-	-	5	-	-
	SO	13	-	1	-	12	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	5
Sensibilidade			1	0.91	0.92	0.92	1		1	1	1	1	1
Especificidade			1	0.98	1	0.97	1		1	1	1	1	1
TCC (%) conjunto			96						100				
TCC (%) modelo			98										

4.4.5 – *i*SPA-PLS-DA

A seleção de intervalos não sobrepostos foi realizada empregando os voltamogramas de todas as amostras sobre as seguintes quantidades de intervalos *w* 1, 5, 10, 15 e 20. Para utilizar *w* = 1, o algoritmo *i*SPA-PLS-DA foi forçado a selecionar um único intervalo, correspondente a menor taxa de erro. Essa estratégia aqui desenvolvida, foi chamada de seleção de intervalo único em Análise Discriminante por Mínimos Quadrados Parciais *i*PLS-DA. Para *w* intervalos a serem selecionados, o algoritmo faz uso da etapa de projeção. Validação cruzada foi utilizada para cálculo de todos os modelos *i*SPA-PLS-DA, onde, o número de fatores utilizados variou de 1 até o número de fatores encontrado no modelo PLS-DA.

O resultado obtido para cada modelo *i*SPA-PLS-DA calculado para *w* = 1, 5, 10, 15 e 20 é mostrado em detalhes na **Tabela 6** em termo da taxa de classificação correta.

Tabela 6 - Resultados das classificações obtidas, *i*PLS-DA, *i*SPA-PLS-DA (*w*= 5, 10, 15 e 20) e para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

Modelo	Taxa de Classificação Correta (%)			Fatores	Intervalos selecionados
	Validação cruzada	Teste	Global		
*i*PLS-DA (*w* = 20)	81,6	76,3	79,0	8	1
*i*SPA-PLS-DA (*w* = 5)	94,7	89,5	92,1	5	2
*i*SPA-PLS-DA (*w* = 10)	96,0	97,4	96,7	5	4
*i*SPA-PLS-DA (*w* = 15)	97,4	97,4	97,4	5	4
*i*SPA-PLS-DA (*w* = 20)	97,4	100	98,7	13	13

Avaliando o desempenho obtido pelo algoritmo *i*SPA-PLS-DA, os modelos que utilizaram *w* = 15 e 20 intervalos resultaram na mesma TCC (97,4%) para conjunto de validação cruzada. Para o conjunto de teste, o modelo *i*SPA-PLS-DA para *w* = 20 intevalos, alcançou uma taxa de classificação correta correspondente a 98,7%. Além disso, a TCC obtida para *i*SPA-PLS-DA foi superior à PLS-DA e *i*PLS-DA, as quais

obtiveram 98,7; 96,5 e 79% de classificações corretas, respectivamente. Por outro lado, comparando-se com os metodos tradicionais com seleção de variáveis individuais, o desempenho de *i*SPA-PLS-DA é igual ao obtido por SPA-LDA. Neste caso, a diferença entre elas está no número de amostras classificadas incorretamente nos conjuntos de validação cruzada e teste, conforme é mostrado nas matrizes de confusão da **Tabelas 4** e **Tabela 7** abaixo.

Tabela 7 - Matriz de confusão para classificação *iSPA*-PLS-DA (w = 20) para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

		Validação cruzada						Teste					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	N_{teste}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	5	-	-	-	-
	GI	11	-	11	-	-	-	5	-	5	-	-	-
	MI	12	-	-	11	1	-	5	-	-	5	-	-
	SO	13	-	1	-	12	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	18
Sensibilidade			1	1	0,92	0,92	1		1	1	1	1	1
Especificidade			1	0,98	1	0,98	1		1	1	1	1	1
TCC (%) conjunto			97,4					100					
TCC (%) modelo		98,7											

CA: classe canola**, GI:** classe girassol**, MI:** classe milho**, SO:** classe soja**, EX:** expirado**, N_{cv}:** número de amostras por classe, para o conjunto de validação cruzada. N_{teste}: número de amostras para o conjunto de teste.

A **Figura 9a** apresenta o intervalo correspondente ao melhor modelo *i*PLS-DA obtido apartir da divisão dos voltamogramas em 20 intervalos. Na **Figura 9b-e** é apresentado os intervalos selecionados para cada modelo *i*SPA-PLS-DA correspondente a divisão w = 5, 10, 15 e 20.

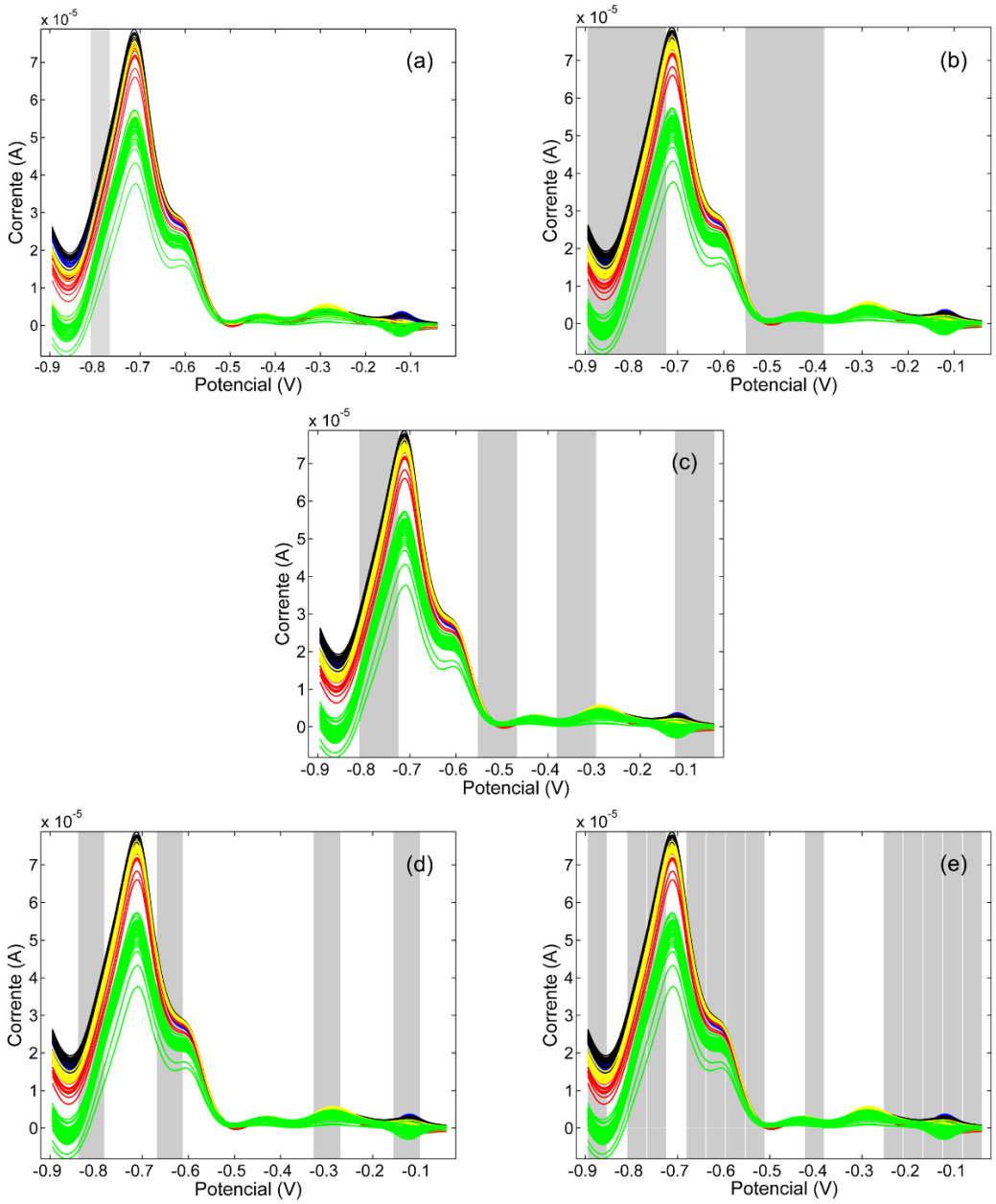


Figura 9 - (a) Intervalo selecionado para o modelo *i*PLS-DA (*w* = 20) e intervalos selecionados para os modelos *i*SPA-PLS-DA (b) *i*SPA-PLS-DA (*w* = 5), (c) *i*SPA-PLS-DA (*w* = 10), (d) *i*SPA-PLS-DA (*w* = 15), *i*SPA-PLS-DA (*w* = 20) (— canola, — girassol, — milho, — soja e — expirada) (Fonte própria).

4.4.6 – *i*SPA-SIMCA

Para avaliar do desempenho da classificação utilizando o algoritmo *i*SPA-SIMCA proposto, os voltamogramas foram divididos nas seguintes quantidades de intervalos *w* não sobrepostos 1, 5, 10, 15 e 20. Todos os modelos *i*SPA-SIMCA foram validados utilizando validação cruzada completa, modificando a quantidade de componentes

principais em cada classe partindo de 1 até ao máximo de componentes principais encontrado no modelo SIMCA.

Na **Figura 10a** é apresentado o intervalo selecionado correspondente ao modelo iSIMCA para *w* = 20 responsável pela maior taxa de classificação correta. Os intervalos selecionados nos modelos *i*SPA-SIMCA correspondente a divisão em *w* intervalos igual 5, 10, 15 e 20 são mostrados na **Figura 10b-e**.

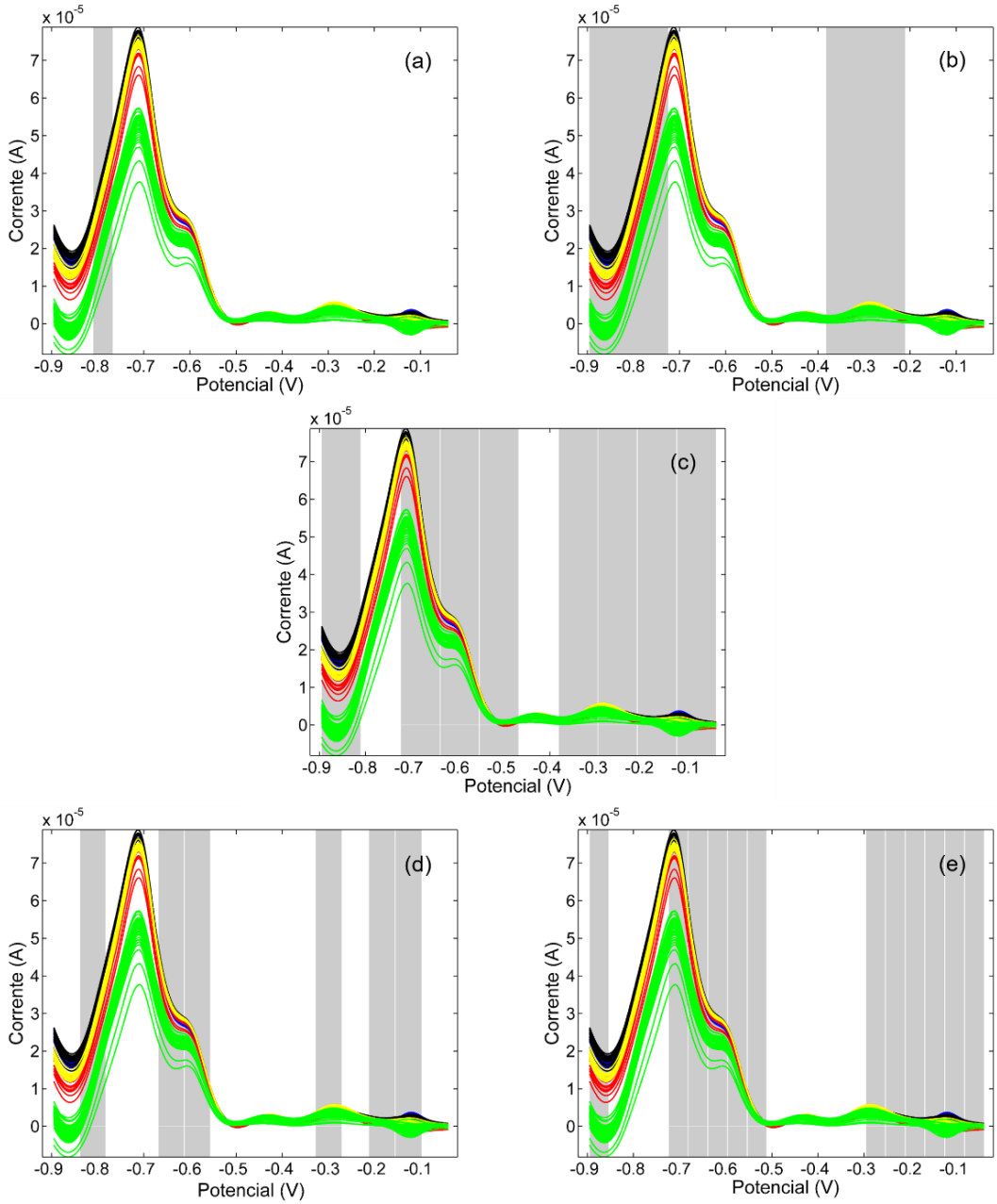


Figura 10 - (a) Intervalo selecionado para o modelo *i*PLS-DA (*w* = 20) e intervalos selecionados para os modelos *i*SPA-PLS-DA (b) *i*SPA-PLS-DA (*w* = 5), (c) *i*SPA-PLS-DA (*w* = 10), (d) *i*SPA-PLS-DA (*w* = 15), *i*SPA-PLS-DA (*w* = 20) (— canola, — girassol, — milho, — soja e — expirada).

Na **Tabela 8** são apresentados os resultados obtidos nos conjutos de validação cruzada e teste para cada modelo *i*SPA-SIMCA correspondente as seguintes quantidades de intervalos (*w* igual a 1, 5, 10, 15 e 20).

Tabela 8 - Resultados das classificações obtidas, *i*SIMCA, *i*SPA-SIMCA (*w*= 5, 10, 15 e 20) e para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

Modelo	Taxa de Classificação Correta (%)			Fatores	Intervalos selecionados
	Validação cruzada	Teste	Global		
*i*SIMCA (*w* = 20)	92	100	96,0	1-3-1-1-1	1
*i*SPA-SIMCA (*w* = 5)	97,4	100	98,7	1-3-1-1-1	2
*i*SPA-SIMCA (*w* = 10)	94,7	94,7	94,7	1-3-1-1-1	8
*i*SPA-SIMCA (*w* = 15)	96,0	94,7	95,3	1-3-1-1-1	6
*i*SPA-SIMCA (*w* = 20)	97,4	97,4	97,4	1-3-1-1-1	12

Observando-se os resultados de SIMCA, *i*SIMCA e *i*SPA-SIMCA na **Tabela 8**, podemos observar que os melhores resultados para os modelos *i*SIMCA e *i*SPA-SIMCA foram alcançados utilizando a divisão de 20 e 5 intervalos, respectivamente. A taxa de classificação correta para *i*SPA-SIMCA é superior à *i*SIMCA e SIMCA, as quais obtiveram 98,7; 96 e 95% de classificações corretas, respectivamente. Neste caso, observados que as técnicas com seleção de intervalos melhoram a capacidade classificatória quando comparada com o modelo que emprega o voltamograma completo. Por outro lado, comparando-se com as técnicas tradicionais com seleção de variáveis individuais, o desempenho de *i*SPA-SIMCA também é igual ao obtido por SPA-LDA. Para *i*SPA-SIMCA, duas amostras foram classificadas incorretamente no conjunto de validação cruzada (uma de milho como soja e outra de soja como girassol), enquanto para SPA-LDA apenas uma amostra de girassol foi classificada incorretamente como soja no conjunto de teste, conforme é mostrado na matriz de confusão da **Tabela 9** abaixo.

Tabela 9: Matriz de confusão para classificação *i*SPA-SIMCA para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

		Validação cruzada						Teste					
Classe atribuída		N_{cv}	**CA**	**GI**	**MI**	**SO**	**EX**	N_{teste}	**CA**	**GI**	**MI**	**SO**	**EX**
Classe verdadeira	**CA**	10	10	-	-	-	-	5	5	-	-	-	-
	GI	11	-	11	-	-	-	5	-	5	-	-	-
	MI	12	-	-	11	1	-	5	-	-	5	-	-
	SO	13	-	1	-	12	-	5	-	-	-	5	-
	EX	30	-	-	-	-	30	18	-	-	-	-	18
Sensibilidade			1	1	0,92	0,92	1		1	1	1	1	1
Especificidade			1	0,98	1	0,98	1		1	1	1	1	1
TCC (%) conjunto			97,4						100				
TCC (%) modelo			98,7										

CA: classe canola**, GI:** classe girassol**, MI:** classe milho**, SO:** classe soja**, EX:** classe expirado**, N_{cv}:** número de amostras por classe, para o conjunto de validação cruzada. N_{teste}: número de amostras para o conjunto de teste.

4.5. Considerações Finais

Neste capítulo, foi avaliada a classificação de óleos vegetais de acordo com o tipo (canola, milho, girassol e soja) e estado de conservação (expirado/não expirado) utilizando dados de voltametria de onda quadrada. O desempenho dos algoritmos propostos *i*SPA-PLS-DA e *i*SPA-SIMCA, forneceram bons resultados com taxas de classificação corretas sempre maiores aos obtidos pelos modelos PLS-DA e SIMCA.

As abordagens que empregaram seleção de um único intervalo (*i*PLS-DA e *i*SIMCA), sempre obtiveram taxa de classificação correta muito inferior aos modelos PLS-DA, SIMCA, SPA-LDA, GA-LDA e consequentemente, aos modelos de seleção de intervalos *i*SPA-PLS-DA e *i*SPA-SIMCA.

Quando comparado, resultados obtidos pelos modelos de seleção de variáveis individuais SPA-LDA e GA-LDA com algoritmos propostos, as taxas de classificação correta resultante dos modelos otimizados *i*SPA-PLS-DA e *i*SPA-SIMCA, foram sempre iguais ou levemente superiores aos obitidos pelos modelos SPA-LDA e GA-LDA.

Identificação de adulteração de misturas de biodiesel/diesel (B5)

Capítulo 5

5. IDENTIFICAÇÃO DE MISTURAS BIODIESEL/DIESEL ADULTERADAS COM ÓLEO VEGETAL EMPREGANDO ESPECTROSCOPIA DE ABSORÇÃO MOLECULAR UV-VIS

5.1 Apresentação

Neste estudo de caso, a classificação de misturas biodiesel/diesel é avaliada com respeito à presença ou não, de óleo de soja constituindo as blendas B5, utilizando espectroscopia de absorção molecular do Uv-visível. O desempenho dos algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos são comparados com PLS-DA, SIMCA, GA-LDA, SPA-LDA, *i*PLS-DA e *i*SIMCA em termos de sensibilidade, especificidade e taxa de classificação correta.

5.2 Experimental

O conjunto de dados é formado por 90 amostras de mistura biodiesel/diesel. 31 amostras formam a classe não adulteradas (mistura biodiesel/diesel B5), e 59 amostras conferem a classe de amostras B5 adulteradas com óleo de soja cru (OB5).

As amostras de biodiesel puro foram sintetizadas a partir de óleos de soja refinados comercialmente disponíveis. Etanol absoluto foi utilizado para promover a reação de transesterificação, e como catalisador para reação foi utilizado Hidróxido de Potássio PA. A mistura reacional foi submetida à agitação magnética e aquecimento a 45°C por 1 hora. Após o tempo reacional, as amostras foram deixadas em repouso para separação da glicerina e purificação.

As amostras de diesel puro foram cedidas pela Petrobras Distribuidora, localizada em Cabedelo no estado da Paraíba. As amostras de óleo vegetal de soja foram compradas no comércio local de Campina Grande no estado da Paraíba, com diferentes marcas e lotes.

As misturas biodiesel/diesel (B5) foram produzidos em balões volumétrico de 50 ml. Na classe B5, a relação diesel puro/biodiesel puro correspondeu a 95% (v/v) de diesel e 5% (v/v) de biodiesel. As amostras da classe adulteradas com óleo de soja cru foram produzidas também em balões de 50 ml, nessa classe, a porcentagem de diesel puro na mistura final equilvaleu a 95% (v/v), e os 5% restantes do volume da mistura foi composto de biodiesel puro e óleo cru. A **Figura 10** ilustra o procedimento para preparação das amostras da classe B5 e OB5.

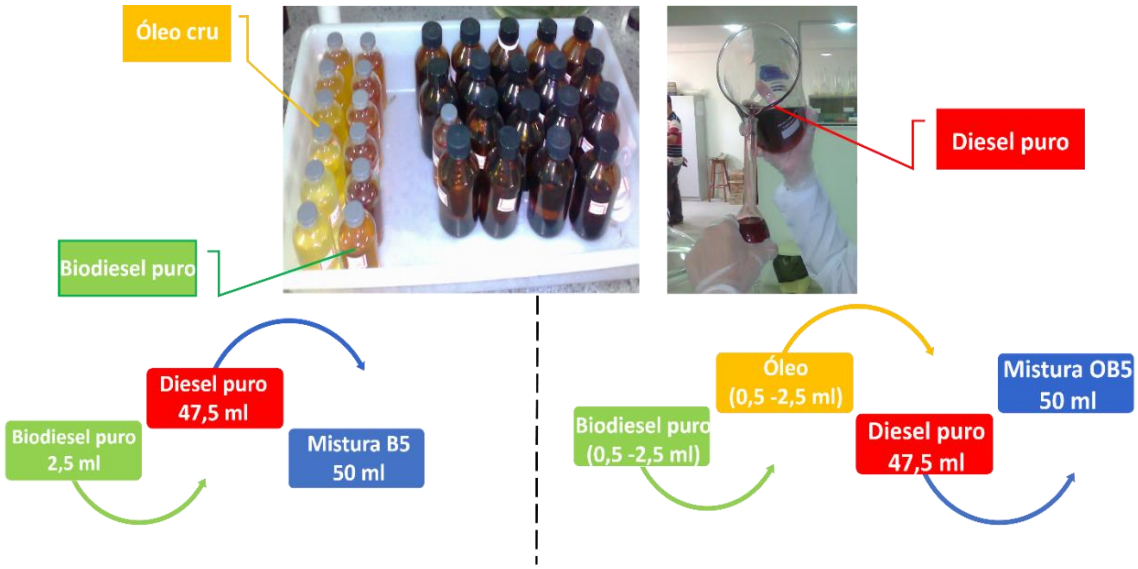


Figura 10 – Metodologia empreganda na confecção das 90 amostras de biodiesel/diesel (Fonte própria).

Os espectros das 90 amostras foram obtidos na região compreendida entre 410 – 850 nm empregando um espectrofotômetro da Perkin Elmer, modelo Lambda 750, cubeta de quartzo com caminho óptico de 1 cm foi utilizado.

5.3. Resultados e discussão

O perfil espectral das 90 amostras de misturas biodiesel/diesel registrados entre 410 a 850 nm são mostrados na **Figura 11**. Uma parte das amostras da classe B5 puras possui uma leve tendência de separação das amostras da classe OB5 puro em torno de 450 nm.

Para melhor observação da tendência natural de separação das amostras OB5 das B5, o emprego de ferrementas de analise exploratória torna-se necessária.

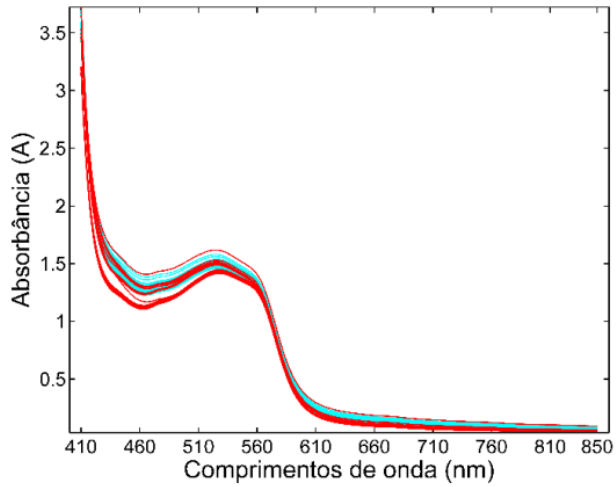


Figura 11 - Espectros UV-visível das 90 amostras das misturas biodiesel/diesel (girassol, milho) (Fonte própria).

5.3.1. Analise exploratória

A **Figura. 12** apresenta o gráfico de escores para as quatros primeiras componentes principais. Analisando-se os gráficos dos escores de PC1 versus PC2 podemos verificar um pequeno agrupamento das amostras B5 e uma sobreposição parcial entre as amostras B5 e OB5 em torno de PC2. Em virtude da sobreposição apresentadas nos escores de PCA, torna-se necessário o desenvolvimento de modelos baseados em métodos de reconhecimento de padrão supervisionado.

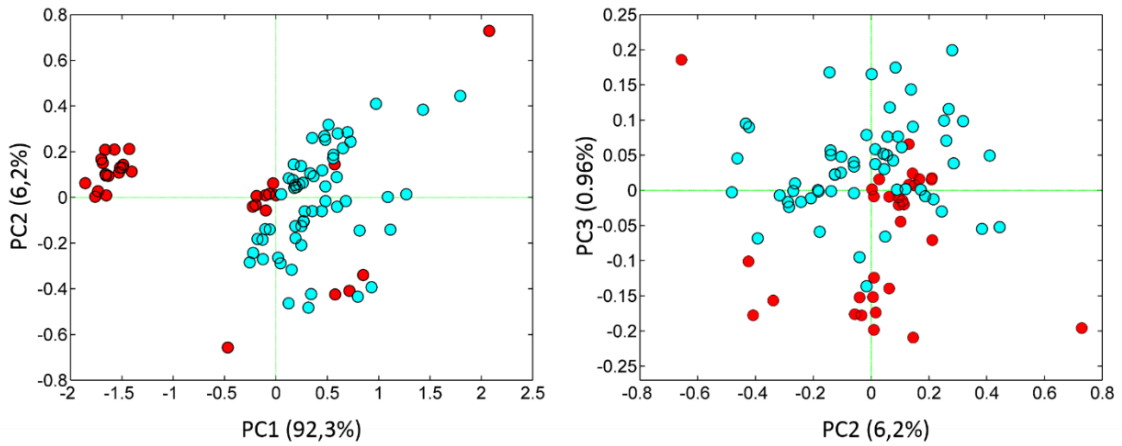


Figura 12 - Gráficos de (a e b) escores de PCA para as 90 amostras misturas biodiesel/diesel estudadas (OB5, B5) (Fonte própria).

5.3.2 – Formação dos conjuntos de dados

O conjunto de dados foi então dividido em conjuntos de treinamento com 61 amostras (21 misturas B5 e 40 misturas /diesel OB5) e 29 amostras no conjunto de teste (10 misturas B5 e 19 misturas OB5) empregando algoritmo Kernnard-Stone [75].

5.4 - Classificação

5.4.1. – PLS-DA

Um único modelo PLS-DA foi construído para as duas classes de misturas biodiesel/diesel estudadas. A quantidade de fatores ótimos usado no modelo PLS-DA final foi determinado empregando validação cruzada completa. A função da taxa de erro versus o número de fatores e os gráficos dos coeficientes de regressão (ver apendece 2) corroboraram para escolha de 13 o número de fatores ideal.

Para atribuição das amostras a uma dada classe, o limiar utilizado foi determinado empregando a função multi objetiva ROC do inglês, *Receiver Operating Characteristics*.

Na **Tabela 10** são mostrados os valores da especificidade, sensibilidade e da taxa de classificação correta referente aos conjuntos de validação cruzada e teste para o modelo PLS-DA. Como pode ser visto na **Tabela 10**, ambos os conjutos de validação cruzada e teste obtiveram máximos valores de especificidade, sensibilidade e TCC.

Tabela 10: Matriz de confusão para o modelo PLS-DA

		Validação cruzada				Test	
Classe atribuída		N_{cv}	**B5**	**OB5**	Nt_{este}	**B5**	**OB5**
Classe verdadeira	**B5**	21	21	-	10	10	-
	OB5	40	-	40	19	-	19
Sensibilidade			1	1		1	1
Especificidade			1	1		1	1
TCC (%) conjunto			100		100		
TCC (%) modelo			100				

B5: classe**:** mistura biodiesel/diesel, **OB5:** classe**:** mistura biodiesel/diesel/óleo**,** $\mathbf{N_{cv}}$**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

5.4.2 – Classificação SIMCA

Um modelo SIMCA foi construído individualmente para classe de mistura biodiesel/diesel estudada. O número de *PC's* exigidos em cada classe foi determinda empregando o ponto correspondente a menor taxa de erro em virtude da inclusão de novas componentes principais em cada classe. Para classe das amostras B5 a quantidade de *PC's* exigidos foi igual a quatro, e na classe OB5 uma única componente principal foi requerida.

A **Tabela 11** apresenta o resultado referente ao modelo SIMCA em termo dos valores da especificidade, sensibilidade e taxa de classificação correta para o modelo global, conjunto de validação cruzada e teste.

Tabela 11: matriz de confusão para o modelo SIMCA

		Validação cruzada				Test		
Classe atribuída		N_{cv}	**B5**	**OB5**	**N/atrib**	Nt_{este}	**B5**	**OB5**	**N/atrib**
Classe verdadeira	**B5**	21	15	1	5	10	10	-	-
	OB5	40	-	34	6	19	-	19	-
Sensibilidade			0,94	1			1	1	
Especificidade			1	0.94			1	1	
TCC (%) conjunto			80,0				100		
TCC (%) modelo			90						

B5: classe**:** mistura biodiesel/diesel, **OB5:** classe**:** mistura biodiesel/diesel/óleo**,** $\mathbf{N_{cv}}$***:*** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste, **N/atrib:** amostras não atribuídas em nenhuma classe**.**

O desempenho do modelo SIMCA atingiu uma taxa de classificação correta de 90%. Para o conjunto de teste 100% das amostras foram bem classificadas e máximos valores de sensibilidade e especificidade foram obtidos. No conjunto de validação cruzada 12 amostras foram mal classificadas resultando em uma TCC de 80%.

5.4.3 – SPA-LDA

O modelo SPA-LDA foi calculado empregando validação cruzada completa. No total 7 variáves distribuídas ao longo de toda região espectral foram selecionadas como apresentada na **Figura 12a**. O gráfico das funções discriminantes obtidas pelo SPA-LDA é exibido na **Figura 12b**.

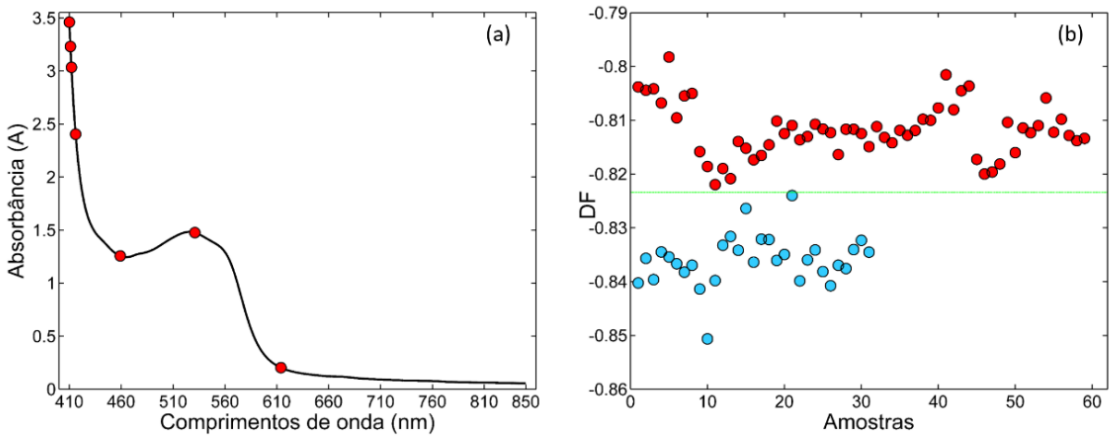


Figura 12 - (a) variáveis selecionadas por SPA-LDA e (b) gráfico de escores de LDA para os dados UV-vis das 90 amostras de misturas biodiesel/diesel estudadas (○OB5, ●B5) (Fonte própria.

O modelo SPA-LDA resultou na classificação correta de todas as amostras de biodiesel estudadas como pode ser visto na **Tabela 12**

Tabela 12: matriz de confusão para o modelo SPA-LDA

		Validação cruzada				Test	
Classe atribuída		N_{cv}	**B5**	**OB5**	Nt_{este}	**B5**	**OB5**
Classe verdadeira	**B5**	21	21	-	10	10	-
	OB5	40	-	40	19	-	19
Sensibilidade			1	1		1	1
Especificidade			1	1		1	1
TCC (%) conjunto			100		100		
TCC (%) modelo			100				

B5: classe**:** mistura biodiesel/diesel, **OB5:** classe**:** mistura biodiesel/diesel/óleo**,** N_{cv}**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

5.4.4 – GA-LDA

A seleção de variáveis empregando GA-LDA resultou na escolha de 9 comprimentos de onda apresentados no espectro médio das 90 amostras na **Figura 13a**.

Apartir das variáveis obtidas no modelo GA-LDA, uma boa discriminação entre as duas classes de misturas estudada foi obtida, como pode ser vista na função discriminante de Fischer apresentada na **Figura 13b**.

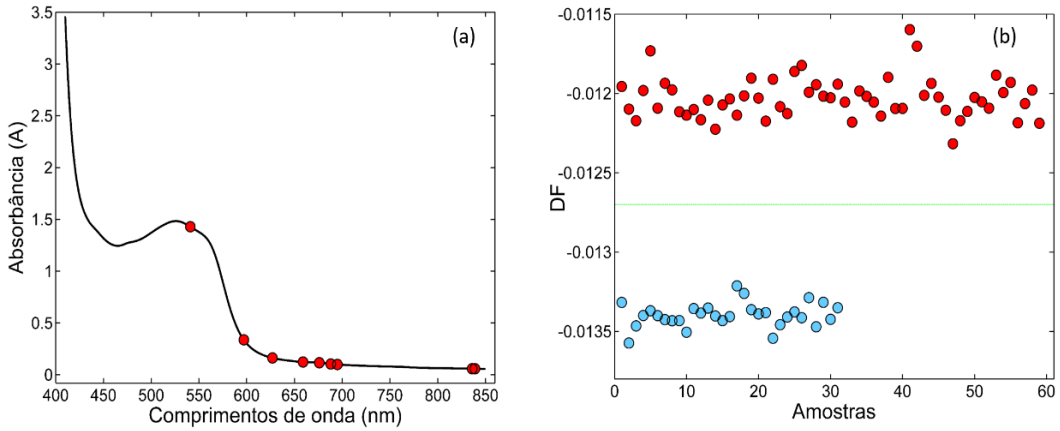


Figura 13 - (a) variáveis selecionadas pelo modelo GA-LDA e (b) gráfico de escores de LDA para os dados UV-vis das 90 amostras de misturas biodiesel/diesel estudadas (OB5, B5) (Fonte própria).

Embora o modelo GA-LDA tenha conseguido classificar corretamente todas as amostras em suas respectivas classes, ver **Tabela 13**, as variáveis selecionadas por esse modelo resultaram em comprimentos de onda localizados em regiões com baixa intensidade.

Tabela 13: matriz de confusão para o modelo GA-LDA

		Validação cruzada				Test	
Classe atribuída		N_{cv}	**B5**	**OB5**	Nt_{este}	**B5**	**OB5**
Classe verdadeira	**B5**	21	21	-	10	10	-
	OB5	40	-	40	19	-	19
Sensibilidade			1	1		1	1
Especificidade			1	1		1	1
TCC (%) conjunto			100		100		
TCC (%) modelo			100				

B5: classe**:** mistura biodiesel/diesel, **OB5:** classe**:** mistura biodiesel/diesel/óleo**,** $\mathbf{N_{cv}}$**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por clase para o conjunto de teste.

5.4.5 – *i*SPA-PLS-DA

Modelos *i*SPA-PLS-DA utilizando 1, 5, 10, 15 e 20 intervalos foram construído empregando validação cruzada completa. A quantidade de fatores empregados na contruição de cada modelo variou de 1 até o máximo de fatores encontrado no modelo PLS-DA.

Tabela 14: Resultados das classificações obtidas por *i*PLS-DA e *i*SPA-PLS-DA para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

Modelo	Taxa de Classificação Correta (%)			Fatores	Intervalos selecionados
	Validação cruzada	Teste	Global		
*i*PLS-DA (*w* = 5)	100	96,5	98,2	13	1
ALGORITMO PROPOSTO					
*i*SPA-PLS-DA (*w* = 5)	100	96,5	98,2	13	1
*i*SPA-PLS-DA (*w* = 10)	100	100	100	13	8
*i*SPA-PLS-DA (*w* = 15)	100	100	100	13	8
*i*SPA-PLS-DA (*w* = 20)	100	100	100	13	13

Na **Tabela 14** é apresentado o resultado obtido em cada modelo *i*SPA-PLS-DA correspondente a divisão de intervalos *w* igual a 1, 5, 10, 15 e 20.

A taxa de classificação correta obtida para o modelo *i*SPA-PLS-DA dividindo os espectros em 15 intervalos alcançou 100% das amostras corretamente classificadas. Esse desempenho é ligeiramente mais parcimonioso quando comparado com o modelo PLS-DA, e alcançou uma TCC igual aos obtidos com as técnicas tradicionais de seleção de variáveis como SPA-LDA e GA-LDA. Além disso, o desempenho alcançado no modelo *i*PLS-DA dividindo os espectros em 5 intervalos resultou em uma TCC igual 98,2%.

A **Figura 14a** apresenta o intervalo correspondente ao modelo iPLS-DA com maior TCC obtido proveniente divisão dos espectros em 5 intervalos. Na Figura 14b-e é

apresentado os intervalos selecionados para cada modelo iSPA-PLS-DA correspondente a divisão w igual 5, 10, 15 e 20.

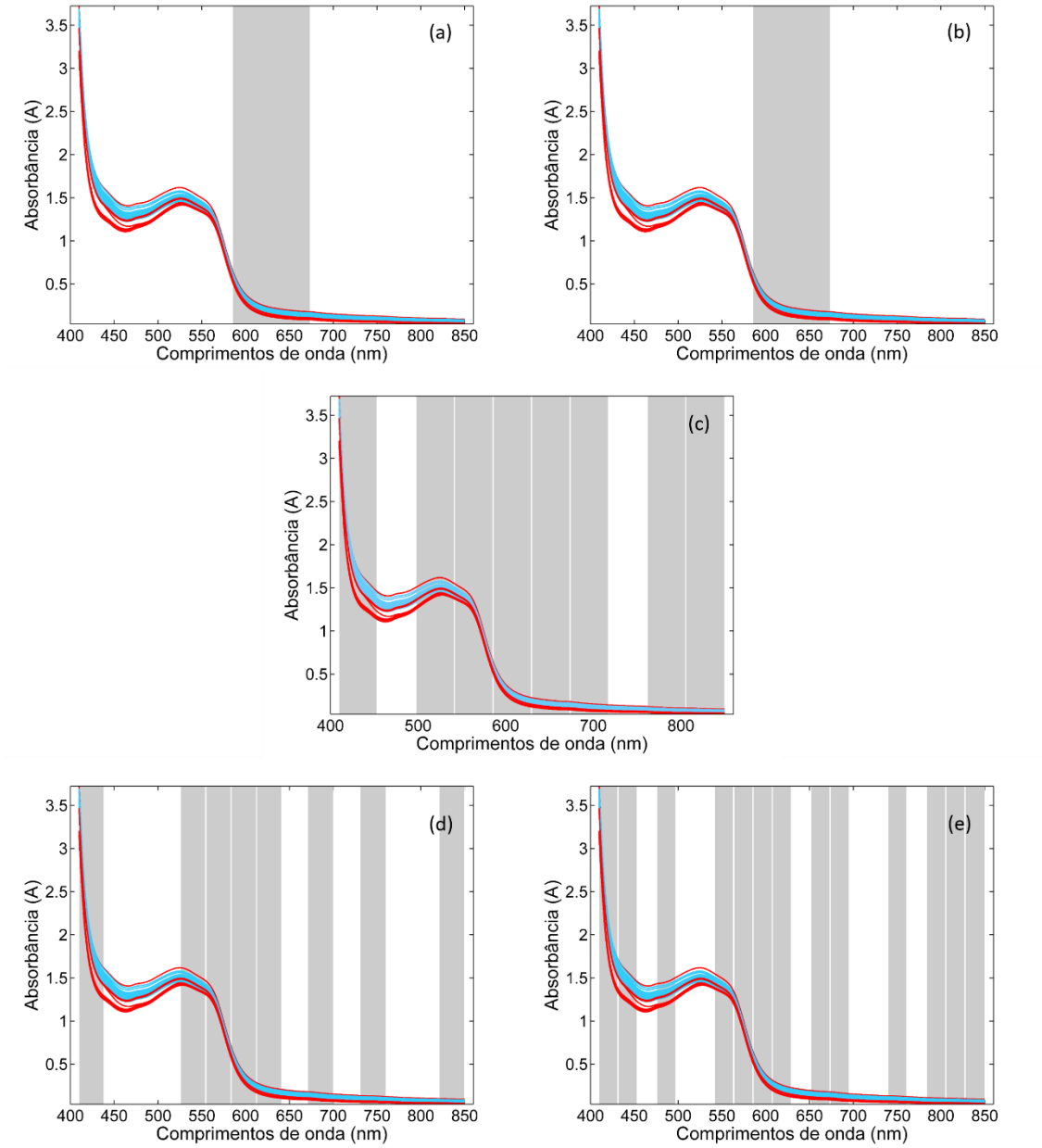


Figura 14 - (a) Intervalo selecionado para o modelo *i*PLS-DA (*w* = 5) e intervalos selecionados para os modelos *i*SPA-PLS-DA (b) *i*SPA-PLS-DA (*w* = 5), (c) *i*SPA-PLS-DA (*w* = 10), (d) *i*SPA-PLS-DA (*w* = 15), *i*SPA-PLS-DA (*w* = 20) (girassol, milho) (Fonte própria).

5.4.6 – *i*SPA-SIMCA

O desempenho da classificação utilizando o algoritmo *i*SPA-SIMCA foi realizado empregando seguintes quantidades de intervalos *w* não sobrepostos 1, 5, 10, 15 e 20.

Validação cruzada foi utilizada para encontrar a quantidade ideal de *PC's* em cada modelo *i*SPA-SIMCA. Para evitar sobreajuste, a quantidade de componentes principais usada em cada modelo *i*SPA-SIMCA não ultrapassou o número de componentes principais usado no modelo SIMCA.

Os intervalos selecionados nos modelos *i*SIMCA para *w* igual 20 e *i*SPA-SIMCA correspondente a divisão em *w* intervalos igual a 5, 10, 15 e 20 são mostrados na **Figura 15a-e**.

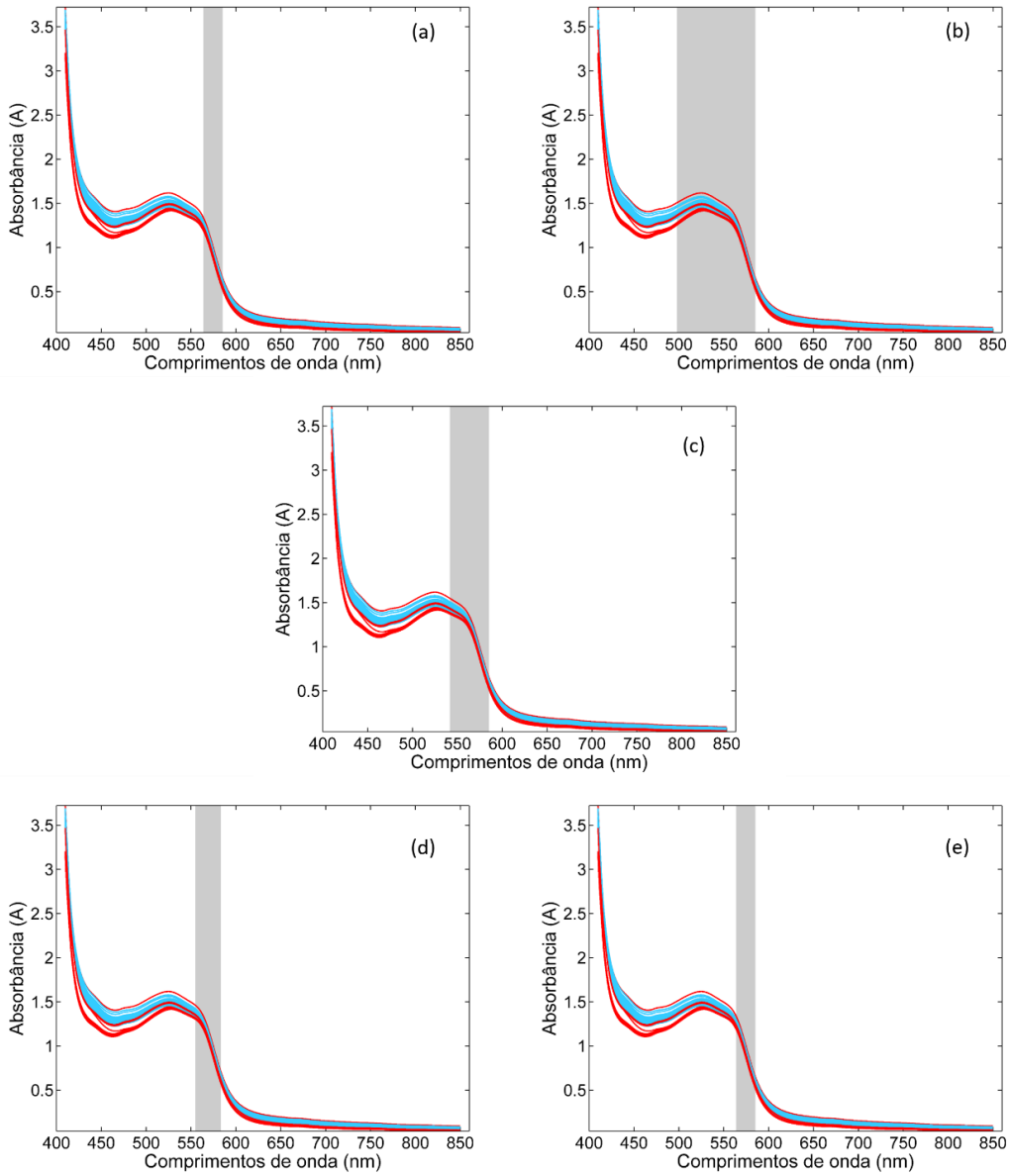


Figura 15 - (a) Intervalo selecionado para o modelo *i*PLS-DA (*w* = 20) e intervalos selecionados para os modelos *i*SPA-PLS-DA (b) *i*SPA-PLS-DA (*w* = 5), (c) *i*SPA-PLS-DA (*w* = 10), (d) *i*SPA-PLS-DA (*w* = 15), *i*SPA-PLS-DA (*w* = 20) (girassol, milho) (Fonte própria).

Na **Tabela 15** são apresentados os resultados obtidos nos conjutos de validação cruzada e teste para cada modelo *i*SPA-SIMCA correspondente aos intervalos *w* igual a 1, 5, 10, 15 e 20.

Tabela 15: Resultados das classificações obtidas *i*SIMCA, e *i*SPA-SIMCA para os dados obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas.

Modelo	Taxa de Classificação Correta (%)			Fatores	Intervalos selecionados
	Validação cruzada	Teste	Global		
*i*SIMCA (*w* = 15)	96,7	100	98,3	2-1	1
ALGORITMO PROPOSTO					
*i*SPA-SIMCA (*w* = 5)	82,0	100	91	3-2	1
*i*SPA-SIMCA (*w* = 10)	85,2	100	92,6	3-2	1
*i*SPA-SIMCA (*w* = 15)	96,7	100	98,3	2-1	1
*i*SPA-SIMCA (*w* = 20)	96,3	100	98,2	2-1	1

Observando-se os resultados de SIMCA, *i*SIMCA e *i*SPA-SIMCA na **Tabela 15**, podemos observar que os resultados para os modelos *i*SIMCA e *i*SPA-SIMCA são equivalentes, uma vez que em todos modelos *i*SPA-SIMCA avaliados (espectros divididos em *w* = 5, 10, 15 e 20) um único intervalo foi selecionado. A taxa de classificação correta para ambos os modelos de seleção de intervalos anteriormente citados, foi superior a taxa de classificação correta obtida pelo SIMCA, as quais obtiveram 98,3 e 90% de classificações corretas, respectivamente. Neste caso, as técnicas com seleção de intervalos melhoram a capacidade classificatória quando comparada com o modelo que emprega a região espectral completa. As técnicas tradicionais com seleção de variáveis individuais (SPA-LDA e GA-LDA) obtiveram melhores desempenho quando comparados aos classificadores *i*SPA-SIMCA e *i*SIMCA. Para o *i*SPA-SIMCA e *i*SIMCA (divididos em 15 intervalos) duas amostras foram classificadas incorretamente no conjunto de validação cruzada (duas amostras classe B5 foram classificadas amostras da classe OB5), conforme é mostrado na matriz de confusão da **Tabela 16**.

Tabela 16 - Matriz de confusão para classificação iSPA-SIMCA (w = 15) para os dados espectroscópicos UV-vis das 90 amostras de misturas biodiesel/diesel estudadas

		Validação cruzada				Test	
Classe atribuída		N_{cv}	**B5**	**OB5**	Nt_{este}	**B5**	**OB5**
Classe verdadeira	**B5**	21	19	2	10	10	-
	OB5	40	-	40	19	-	19
Sensibilidade			0,97	1		1	1
Especificidade			1	0,97		1	1
TCC (%) conjunto			96,7		100		
TCC (%) modelo			98,3				

B5: classe**:** mistura biodiesel/diesel, **OB5:** classe**:** mistura biodiesel/diesel/óleo**,** $\mathbf{N_{cv}}$***:*** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

5.5. Considerações Finais

Neste capítulo, a classificação de misturas biodiesel/diesel é avaliada com respeito a presença ou não, de óleo de soja constituindo as blendas B5, utilizando espectrometria do UV-visível O desempenho do algoritmo *i*SPA-PLS-DA forneceu um resultado sastifatório com taxa de classificação correta iguais aos obtidos pelos modelos PLS-DA, SPA-LDA e GA-LDA. Já o desempenho do algoritmo *i*SPA-SIMCA, atingiu taxa de classificação correta superior ao obtido no modelo SIMCA e levemente inferior aos obtidos nos modelos PLS-DA, *i*SPA-PLS-DA SPA-LDA e GA-LDA.

Os modelos empregando seleção de um único intervalo (*i*PLS-DA e *i*SIMCA), obtiveram taxa de classificação correta muito próxima ao obtidos nos modelos PLS-DA, *i*SPA-PLS-DA SPA-LDA e GA-LDA. Os modelos *i*SPA-PLS-DA (w = 15), *i*PLS-DA (w = 5) *i*SPA-SIMCA (w = 15) e iSIMCA (w = 15) obitveram seus intervalos localizados sobre a região correspondente a composição química das misturas B5 e OB5.

TRIAGEM DE ÓLEOS DE SOJA

Capítulo 6

6. TRIAGEM DE ÓLEOS DE SOJA COM RESPEITO AO ESTADO DE CONSERVAÇÃO UTILIZANDO ESPECTROSCOPIA DO INFRAVERMELHO PRÓXIMO.

6.1 Apresentação

Neste estudo de caso é avaliada a classificação de óleos vegetais de acordo com o estado de conservação (expirado e não expirado) utilizando dados de espectroscopia do infravermelho próximo. O desempenho dos algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos são comparados com PLS-DA, SIMCA, GA-LDA, SPA-LDA, *i*PLS-DA e *i*SIMCA em termos de sensibilidade, especificidade e taxa de classificação correta.

6.2 Experimental

O banco de dado é composto de 50 amostras de óleos de soja de diferentes marcas e lotes. Trinta destas amostras foram armazenadas, em seus frascos comerciais originais sem controle das condições ambientais até passagem da data de validade indicado no rótulo. Para comprovar realmente, que as amostras estavam inapropriadas para o consumo, o índice de peróxidos foi determinado para todas as amostras.

O ensaio de índice de peróxido foi realizado misturando-se, 5g de óleo a 30 mL de uma solução 3:2 de ácido acético/clorofórmio até a dissolução da amostra. Em seguida, 0,5 mL de iodeto de potássio saturado foi adicionado a mistura e, deixado em repouso ao abrigo da luz por exatamente um minuto. Na sequência, adicionou-se 30 mL de água deionizada e 1 mL de amido como indicador. A solução foi titulada com tiossulfato de sódio 0,01 mol / L até a cor escura desaparecesse. As amostras que possuíam valores do índice de peróxidos característicos de oléos inapropriado para o consumo formou a classe nomeada de expirada (EX). Mais tarde, 20 amostras foram compradas para compor a

classe de amostras não expiradas (NEX). Todas as amostras foram adquiridas na cidade de Campina Grande, localizada no Estado da Paraíba, Brasil.

Os espectros dos óleos vegetais foram registrados em triplicata no intervalo de 380 a 2300 nanometro com resolução de 1 nm, usando um espectrofotômetro da Perkim Elmer, modelo Lambda 750 equipado com células de quartzo de 1 cm de trajeto óptico. Mais detalhes experimentais são fornecidos na referência [78].

6.3 Resultados e discussão

Os espectros das 50 amostras de óleos de soja registrados no intervalo de 380 a 2300 nm são mostrados na **Figura 16a**. A região espectral entre 1100 à 1600 nm (sombreada em cinza) foi selecionada para esse estudo. A região espectral selecionada, foi pré-processada usando derivada por filtro Savitzky – Golay para remoção de linha de base e ruído aleatório sob diversas condições (ordem do polinômio, tamanho de janelas e ordem da derivada). Na **Figura 16b** é mostrado os espectros derivados aplicado derivada Savitzky – Golay, polinômio de segunda ordem e janela de 17 pontos. Como pode ser observado no gráfico dos espectros derivados, nenhuma tendência de separação entre as duas classes pode ser observada.

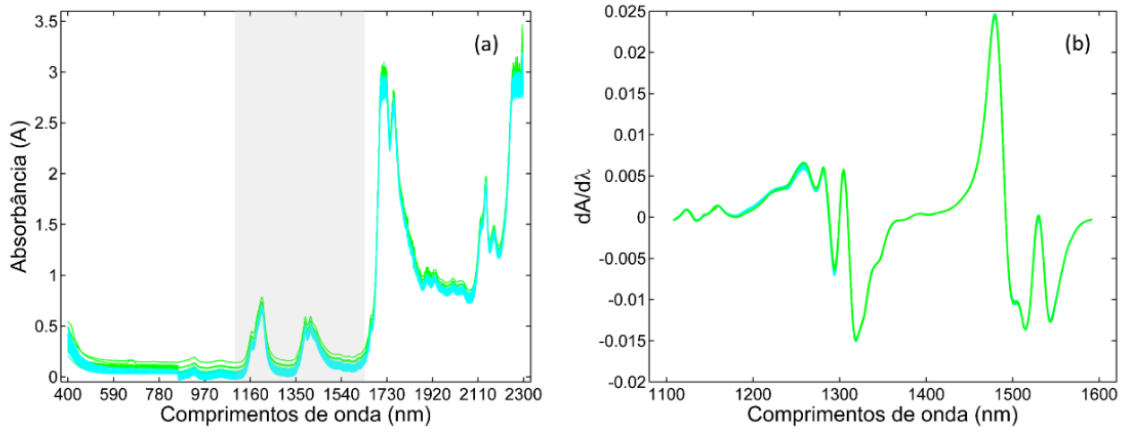


Figura 16 - (a) Espectros das 50 amostras de óleos de soja registrados entre os comprimentos de onda 380 a 2300 nm, (— NEX, — EX), região sombreada em cinza corresponde a parte dos espectros empregados nesse estudo. (b) Espectro pré processados das 50 amostras usada nesse estudo (Fonte própria).

6.3.1. Analise exploratória

O gráfico de escores correspondente as três primeiras componentes principais são apresentadas na **Figura 17a-b**. Analisando-se os escores de PC1 versus PC2 e PC2 versus PC3, podemos verificar as amostras de ambas classes estão parcialmente sobrepostas em torno de PC2 e PC3. Em virtude do resultado apresentado nos escores de PCA, torna-se necessário o desenvolvimento de modelos baseados em métodos de reconhecimento de padrão supervisionado.

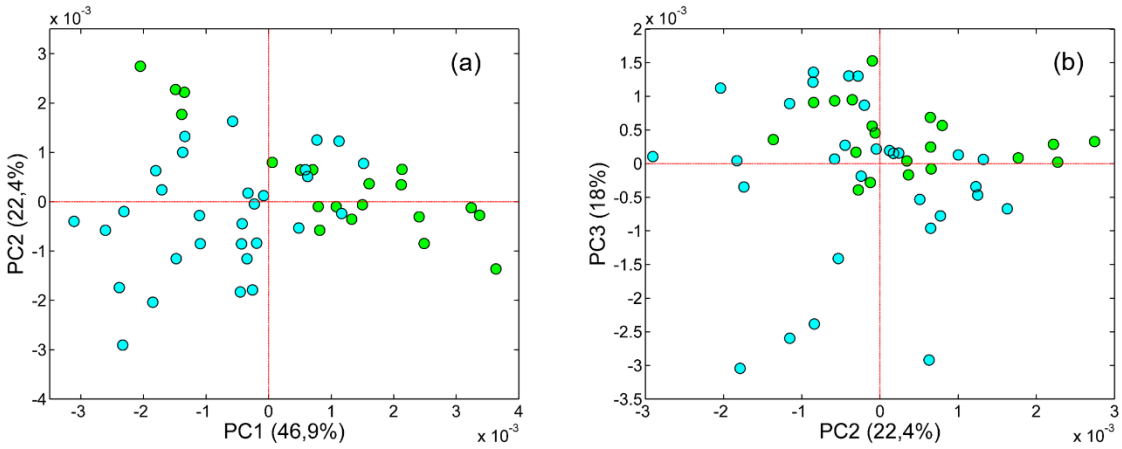


Figura 17 - Gráficos de escores de PCA (a) PC1 versus PC2 (b) PC2 versus PC3 para as 90 amostras misturas biodiesel/diesel estudadas (EX, NEX) (Fonte própria).

6.3.2 – Formação dos conjuntos de dados

O conjunto de dados foi então dividido em conjuntos de treinamento com 35 amostras (15 amostras não expiradas - NEX e 20 amostras expiradas - EX) e 29 amostras no conjunto de teste (5 amostras não expiradas - NEX e 10 amostras expiradas - EX) empregando algoritmo Kernnard-Stone [75].

6.4 - Classificação

6.4.1. – PLS-DA

Um único modelo PLS-DA foi construído para as duas classes de óleos vegetais estudadas. O número de fatores ótimos usado no modelo PLS-DA final foi determinado

utilizando validação cruzada completa. A função da taxa de erro versus o número de fatores, e os gráficos dos coeficientes de regressão (ver apendece 3) confirmam que a quantidade ideal de fatores a ser usado no modelo é igual a 7. Para atribuição das amostras a uma dada classe, o limiar utilizado foi determinado empregando a função multi objetiva ROC do inglês, *Receiver Operating Characteristics*.

Os valores da especificidade, sensibilidade e da taxa de classificação correta referente aos conjuntos de validação cruzada e teste para o modelo PLS-DA são mostrados na **Tabela 17**. Como pode ser visto na **Tabela 17**, o conjunto de validação cruzada obtive uma TCC de 85.7%. Já o conjunto de teste alcançou 100% de classificação correta, além de máximos valores de especificidade, sensibilidade.

Tabela 9: Matriz de confusão para classificação PLS-DA para os dados obtidos por infravermelho próximo para as 50 amostras de óleos vegetais estudadas

		Validação cruzada				Teste	
Classe atribuída		N_{cv}	**NEX**	**EX**	Nt_{este}	**NEX**	**EX**
Classe verdadeira	**NEX**	15	12	3	5	5	-
	EX	20	2	18	10	-	10
Sensibilidade			0.80	0.90		1	1
Especificidade			0.90	0.80		1	1
TCC (%) conjunto			85.7		100		
TCC (%) modelo			92.8				

NEX: Não expiradas, **EX:** Expiradas**,** $\mathbf{N_{cv}}$**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

6.4.2 – Classificação SIMCA

Modelos SIMCA foram construídos individualmente para classe classes de óleos vegetais estudadas. A quantidade de componentes principais exigidos para cada classe foi determinda empregando o ponto correspondente a menor taxa de erro em virtude da inclusão de novas componentes principais em cada classe. Para classe das amostras NEX

a quantidade de *PC's* exigidos foi igual a um, e na classe EX, três componentes principal foram requeridas.

Na **Tabela 18** é apresentado o resultado obtido no modelo SIMCA em termo dos valores da especificidade, sensibilidade e taxa de classificação correta para o modelo global, conjunto de validação cruzada e teste.

Tabela 18: Matriz de confusão para classificação SIMCA para os dados obtidos por infravermelho próximo para as 50 amostras de óleos vegetais estudadas

		Validação cruzada				Teste	
Classe atribuída		N_{cv}	**NEX**	**EX**	Nt_{este}	**NEX**	**EX**
Classe verdadeira	**NEX**	15	14	1	5	5	-
	EX	20	2	18	10	4	6
Sensibilidade			0.93	0.90		1	0.60
Especificidade			0.90	0.93		0.60	1
TCC (%) conjunto			91.4		73.3		
TCC (%) modelo			82.3				

NEX: Não expiradas, **EX:** Expiradas**,** N_{cv}**:** número de amostras por classe para o conjunto de validação cruzada. Nt_{este}: número de amostras por classe para o conjunto de teste.

O desempenho do modelo SIMCA atingiu uma taxa de classificação correta de 82.3%. No conjunto de teste apenas 73,3% das amostras foram bem classificadas. No conjunto de validação cruzada apenas 3 amostras foram mal classificadas resultando em uma TCC de 91,4%.

6.4.3 – SPA-LDA

O modelo SPA-LDA foi calculado empregando validação cruzada completa. No total 7 variáves distribuídas ao longo de toda região espectral foram selecionadas como apresentada na **Figura 18a**. O gráfico das funções discriminantes obtidas pelo SPA-LDA é exibido na **Figura 18b**.

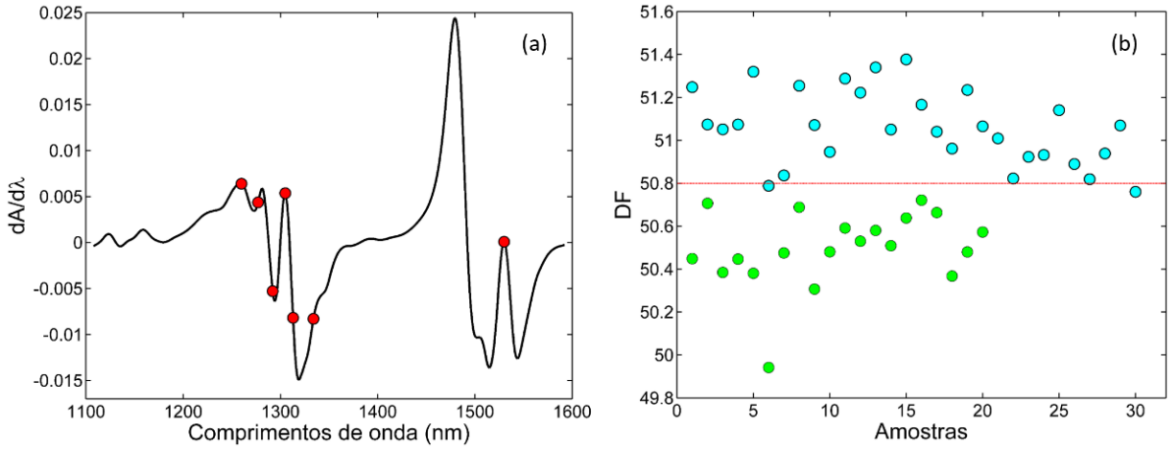


Figura 18 - (a) variáveis selecionadas pelo modelo SPA-LDA e (b) gráfico de escores de LDA para os dados de infravermelho próximo das 50 amostras de óleos vegetais estudadas (●EX, ●NEX) (Fonte própria).

O modelo SPA-LDA global resultou em uma taxa de classificação correta igual a 95,2 % como pode ser visto na **Tabela 19**.

Tabela 19 - Matriz de confusão para classificação SPA-LDA para os dados obtidos por infravermelho próximo para as 50 amostras de óleos vegetais estudadas

		Validação cruzada				Teste	
Classe atribuída		N_{cv}	**NEX**	**EX**	Nt_{este}	**NEX**	**EX**
Classe verdadeira	**NEX**	15	15	-	5	5	-
	EX	20	1	19	10	1	9
Sensibilidade			1	0,95		1	1
Especificidade			0,95	1		1	1
CCR (%) conjunto			97,1		93,3		
CCR (%) modelo			95,2				

NEX: Não expiradas, **EX:** Expiradas**,** $\mathbf{N_{cv}}$**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste

6.4.4 – GA-LDA

A seleção de variáveis empregando GA-LDA resultou na escolha de 17 comprimentos de onda apresentados no espectro médio das 50 amostras na **Figura 19a**.

No modelo GA-LDA, uma boa discriminação entre as duas classes de óleos vegetais estudada foi obtida, como pode ser vista na função discriminante de Fischer apresentada na **Figura 19b**.

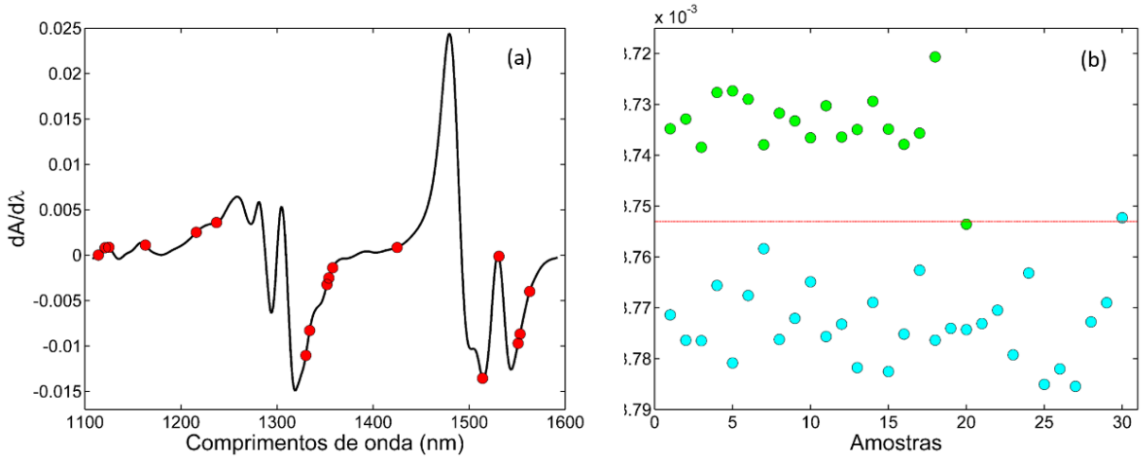


Figura 19 - (a) variáveis selecionadas pelo modelo GA-LDA e (b) gráfico de escores de LDA para os dados de infravermelho próximo das 50 amostras de óleos vegetais estudadas (EX, NEX) (Fonte própria).

Embora o modelo GA-LDA tenha conseguido classificar corretamente todas as amostras em suas respectivas classes, ver **Tabela 20**, as variáveis selecionadas por esse modelo resultaram em comprimentos de onda localizados em regiões com baixa intensidade.

Tabela 20 - Matriz de confusão para classificação GA-LDA para os dados obtidos por infravermelho próximo para as 50 amostras de óleos vegetais estudadas

		Validação cruzada				Teste	
Classe atribuída		N_{cv}	**NEX**	**EX**	Nt_{este}	**NEX**	**EX**
Classe verdadeira	**NEX**	15	15	-	5	4	1
	EX	20	-	20	10	1	9
Sensibilidade			1	1		0,8	0,9
Especificidade			1	1		0,9	0,8
CCR (%) conjunto			100		86,7		
CCR (%) modelo			93,3				

NEX: Não expiradas, **EX:** Expiradas**,** N_{cv}**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

6.4.5 – *i*SPA-PLS-DA

Modelos *i*SPA-PLS-DA utilizando 1, 5, 10, 15 e 20 intervalos foram construído empregando validação cruzada completa. A quantidade de fatores empregados na contruição de cada modelo variou de 1 até o máximo de fatores encontrado no modelo PLS-DA.

Na **Tabela 21** é apresentado o resultado obtido em cada modelo *i*SPA-PLS-DA correspondente a divisão de intervalos *w* igual a 1, 10, 20 e 30.

Tabela 21 - Resultados das classificações obtidas por *i*PLS-DA e *i*SPA-PLS-DA para os dados obtidos por espectroscopia do infravermelho próximo para as 50 amostras de óleos vegetais estudadas.

Modelo	**Taxa de Classificação Correta (%)**			**Fatores**	**Intervalos selecionados**
	Validação cruzada	**Teste**	**Global**		
*i*PLS-DA (*w* = 10)	94,3	93,3	93,8	7	1
ALGORITMOS PROPOSTOS					
*i*SPA-PLS-DA (*w* = 10)	94,3	93,3	93,8	7	1
*i*SPA-PLS-DA (*w* = 20)	97,1	86,7	91,9	7	2
*i*SPA-PLS-DA (*w* = 30)	97,1	93,3	95,2	7	2

O modelo *i*SPA-PLS-DA correspondente a maior taxa de classificação correta foi obtida dividindo os espectros em 30 intervalos, 95,2% das amostras foram corretamente classificadas. Esse desempenho é mais parcimonioso quando comparado com o modelo PLS-DA, e alcançou uma TCC igual ao obtido com seleção de variáveis pelo SPA-LDA. O desempenho alcançado no modelo *i*PLS-DA dividindo os espectros em 10 intervalos resultou em uma TCC igual 93,8%.

A **Figura 20a** apresenta o intervalo correspondente ao modelo *i*PLS-DA com maior TCC obtido proveniente divisão dos espectros em 10 intervalos. Na **Figura 20b-d** são apresentados os intervalos selecionados para cada modelo *i*SPA-PLS-DA correspondente a divisão *w* igual 10, 20 e 30.

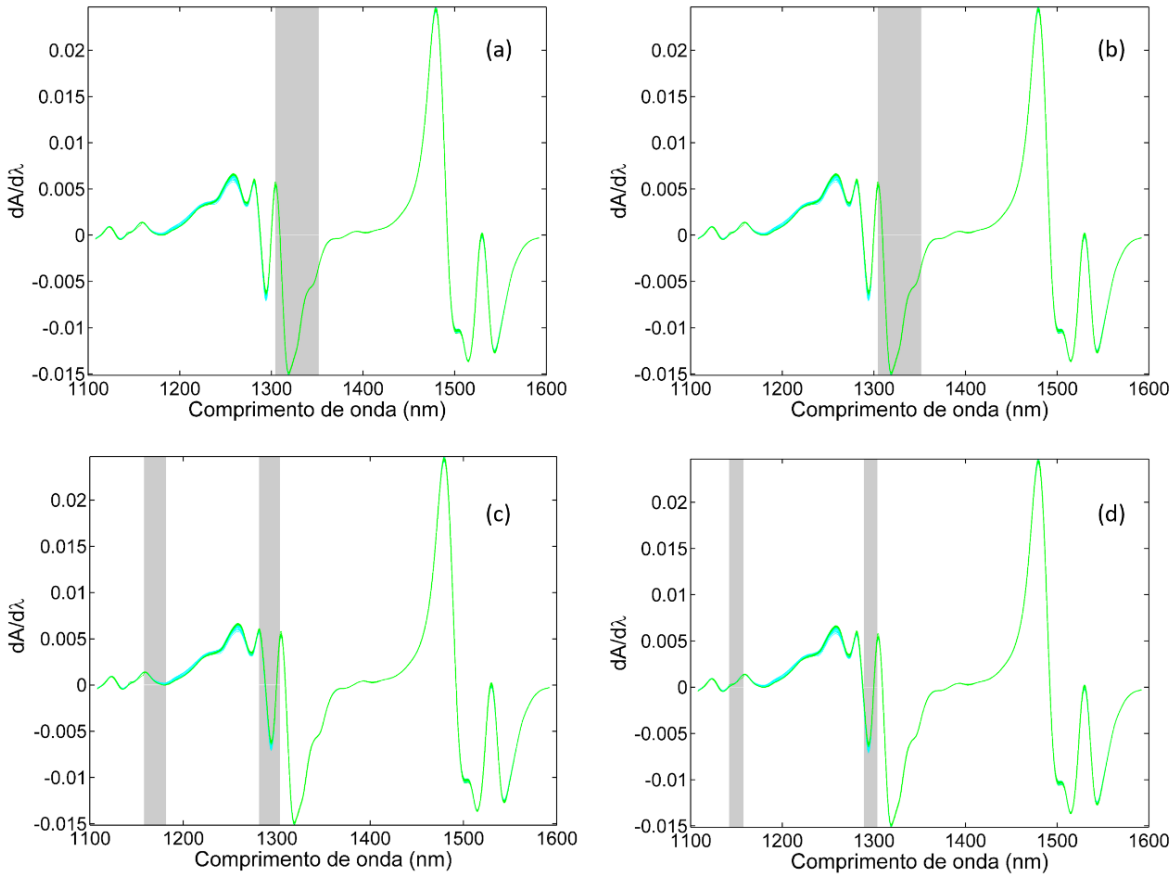


Figura 20 - (a) Intervalo selecionado para o modelo *i*PLS-DA (*w* = 10) e intervalos selecionados para os modelos *i*SPA-PLS-DA (b) *i*SPA-PLS-DA (*w* = 10), (c) *i*SPA-PLS-DA (*w* = 20), (d) *i*SPA-PLS-DA (*w* = 30) para os dados de infravermelho próximo das 50 amostras de óleos vegetais estudadas (— NEX, — EX) (Fonte própria).

6.4.6 – *i*SPA-SIMCA

O desempenho da classificação utilizando o algoritmo *i*SPA-SIMCA foi realizado empregando seguintes quantidades de intervalos *w* não sobrepostos 1, 10, 20 e 30. Para encontrar a quantidade ideal de *PC's* em cada modelo *i*SPA-SIMCA foi empregado validação cruzada. Para evitar sobreajuste, a quantidade de componentes principais usada em cada modelo *i*SPA-SIMCA não ultrapassou o número de componentes principais usado no modelo SIMCA.

Os intervalos selecionados nos modelos *i*SIMCA para *w* igual 10 e *i*SPA-SIMCA correspondente a divisão em *w* intervalos igual a 10, 20 e 30 são mostrados na **Figura 21a-d**.

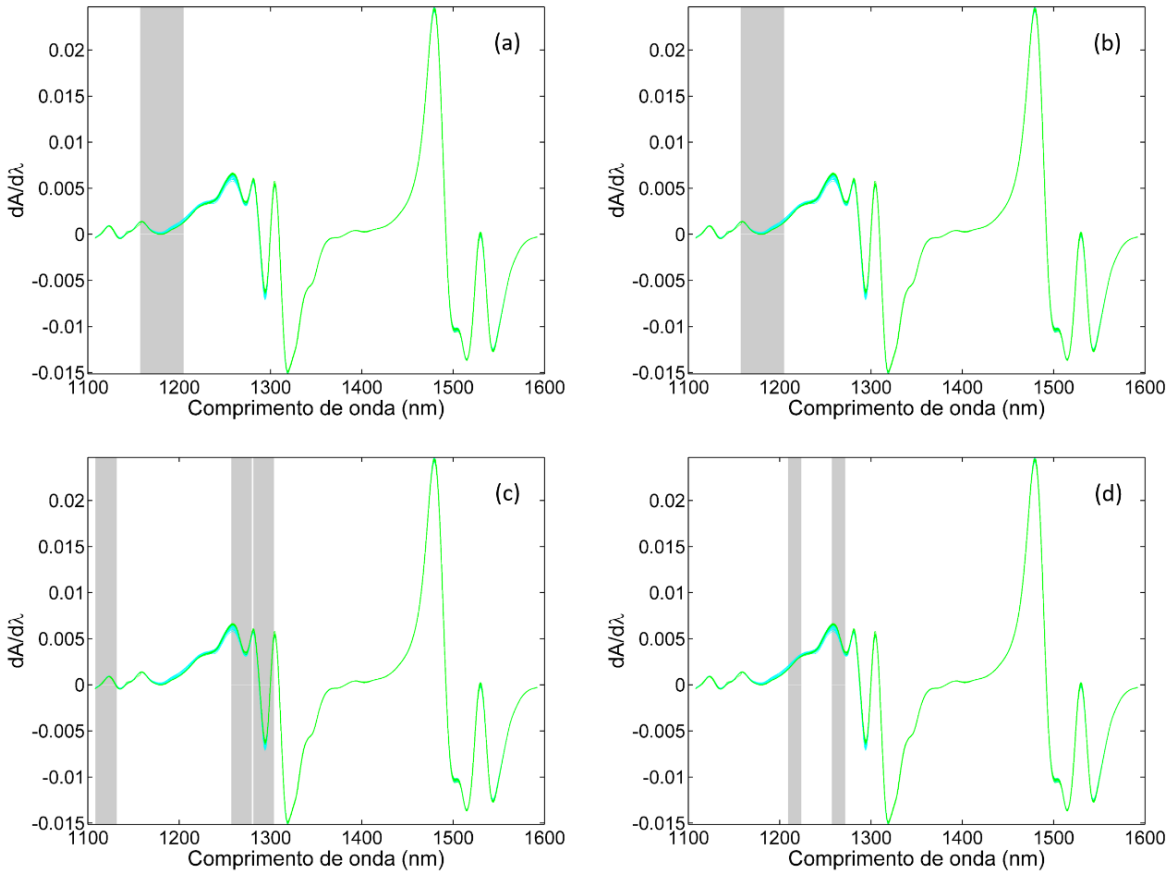


Figura 21 - (a) Intervalo selecionado para o modelo *i*SIMCA (*w* = 10) e intervalos selecionados para os modelos *i*SPA-SIMCA (b) *i*SPA-SIMCA (*w* = 10), (c) *i*SPA-SIMCA (*w* = 20), (d) *i*SPA-SIMCA (*w* = 30) para os dados de infravermelho próximo das 50 amostras de óleos vegetais estudadas (— NEX, — EX) (Fonte própria).

Na **Tabela 22** são apresentados os resultados obtidos nos conjutos de validação cruzada e teste para cada modelo *i*SPA-SIMCA correspondente aos intervalos *w* igual a 1, 10, 20 e 30.

Tabela 22 - Resultados das classificações obtidas por *i*SIMCA e *i*SPA-SIMCA para os dados obtidos por espectroscopia do infravermelho próximo para as 50 amostras de óleos vegetais estudadas.

Modelo	**Taxa de Classificação Correta (%)**			**Fatores**	**Intervalos selecionados**
	Validação cruzada	**Teste**	**Global**		
*i*SIMCA (*w* = 10)	94,3	93,3	93,8	1-3	1
ALGORITMOS PROPOSTOS					
*i*SPA-SIMCA (*w* = 10)	94,3	93,3	93,8	1-3	1
*i*SPA-SIMCA (*w* = 20)	94,3	86,7	90,5	1-3	3
*i*SPA-SIMCA (*w* = 30)	94,3	93,3	93,8	1-3	2

Os resultados dos modelos SIMCA, *i*SIMCA e *i*SPA-SIMCA apresentados na **Tabela 22** demonstram que os modelos *i*SIMCA e *i*SPA-SIMCA são equivalentes, uma vez que em todos modelos *i*SPA-SIMCA obtiveram a mesma taxa de classificação correta. A taxa de classificação correta para ambos os modelos de seleção de intervalos anteriormente citados, foi superior a taxa de classificação correta obtida pelo SIMCA, as quais obtiveram 93,8 e 82,3% de classificações corretas, respectivamente. Neste caso, as técnicas com seleção de intervalos melhoram a capacidade classificatória quando comparada com o modelo que emprega a região espectral completa. As técnicas tradicionais com seleção de variáveis individuais (SPA-LDA e GA-LDA) resultaram em maiores taxas de classicação correta quando comparados aos classificadores *i*SPA-SIMCA e *i*SIMCA. Para o *i*SPA-SIMCA para *w* igual a 30 intervalos, duas amostras foram classificadas incorretamente no conjunto de validação cruzada (sendo essas, duas amostras classe EX foram classificadas amostras da classe NEX), e uma amostras no conjunto de teste pertencente a classe EX foi mal classificada na classe NEX conforme é mostrado na matriz de confusão da **Tabela 23**.

Tabela 23 - Matriz de confusão para classificação *i*SPA-SIMCA (*w* = 30) para os dados obtidos por infravermelho próximo para as 50 amostras de óleos vegetais estudadas

		Validação cruzada				**Teste**	
Classe atribuída		N_{cv}	**NEX**	**EX**	Nt_{este}	**NEX**	**EX**
Classe verdadeira	**NEX**	15	15	-	5	5	-
	EX	20	2	18	10	1	9
Sensibilidade			1	0,9		1	0.9
Especificidade			0,9	1		0.9	1
CCR (%) conjunto			94,3		93,3		
CCR (%) modelo			93,8				

NEX: Não expiradas, **EX:** Expiradas**,** $\mathbf{N_{cv}}$**:** número de amostras por classe para o conjunto de validação cruzada. N_{teste}: número de amostras por classe para o conjunto de teste.

6.5. Considerações Finais

Neste capítulo, a classificação de oléos vegetais de soja foi avaliada a classificação de óleos vegetais de acordo com o estado de conservação (expirado e não expirado) utilizando dados de espectroscopia do infravermelho próximo O desempenho dos algoritmos *i*SPA-PLS-DA e iSPA-SIMCA forneceram resultado com taxa de classificação correta superiores aos aobitidos nos modelos PLS-DA e SIMCA, e iguais taxas de classicação obtidos pelos modelos SPA-LDA e GA-LDA. Os modelos *i*PLS-DA e *i*SIMCA obtiveram taxa de classificação correta maiores aos obtidos nos modelos PLS-DA e SIMCA.

Conclusões

Capítulo 7

Neste trabalho, foram desenvolvidos e avaliados, dois algoritmos que combinam a seleção de intervalos usando algoritmo das projeções sucessivas, em três estudos de caso. Os algoritmos *i*SPA-PLS-DA e *i*SPA-SIMCA propostos forneceram bons resultados nos três estudos de caso, com taxas de classificação corretas e sempre maiores ou iguais às obtidos pelos modelos PLS-DA e SIMCA utilizando todas as variáveis, *i*PLS-DA e *i*SIMCA com um único intervalo selecionado e o SPA-LDA e GA-LDA com seleção de variáveis individuas. Os *i*SPA-PLS-DA e *i*SPA-SIMCA podem ser consideradas uma abordagem promissora para uso em problemas de seleção por intervalo.

De forma geral, a possibilidade de utilização de intervalos variáveis, sem perda da precisão da classificação pode ser útil para a concepção de instrumentos (por exemplo, dedicados fotómetros) para uso de rotina.

Com propostas de continuidade deste trabalho as seguintes etapas serão desenvolvidas:

- Refinamento do código do *i*SPA-PLS-DA;
- Refinamento do código do *i*SPA-SIMCA;
- Ampliação dos algoritmos para ser usados validação por serie de teste
- Construção da interface gráfica para o *i*SPA-PLS-DA
- Construção da interface gráfica para o *i*SPA-SIMCA
- Avaliação dos métodos propostos em aplicações envolvendo dados de outras técnicas analíticas.

Referências
Capítulo 8

[1] M. Isabel López, M. Pilar Callao, Itziar Ruisánchez, A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach, Analytica Chimica Acta 891 (2015) 62–72.

[2] R.G Brereton, Chemometrics and Statistics: Multivariate Classification Techniques. In: Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier, 2013.

[3] Luis Cuadros-Rodríguez, Estefanía Pérez-Castaño, Cristina Ruiz-Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, Trends in Analytical Chemistry 80 (2016) 612–624

[4] M. Valcárcel, S. Cárdenas, B.M. Simonet, C. Carrillo Carrión, Principles of qualitative analysis in the chromatographic context, J. Chromatogr. A 1158 (2007) 234–240.

[5] B.L. Milman, L.A. Konopel'ko, Uncertainty of Qualitative chemical analysis: General methodology and binary test methods, J. Anal. Chem. 23 (2004) 1128–1141.

[6] D. Ballabio, V. Consonni, A. Mauri, R. Todeschini, Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 3. Variable selection in classification, Analytica Chimica Acta 657 (2010) 116–122.

[7] Z. Xiaobo, Z. Jiewen, M. J.W. Povey, M. H., M. Hanpin, Variables selection methods in near-infrared spectroscopy, Analytica Chimica Acta 667 (2010) 14–32.

[8] M. C. U. Araújo, R. K. H. Galvão, Linear Regression Modeling: Variable Selection. In: Beata Walczak; Romá Tauler Ferré; Steven Brown. (Org.). Comprehensive Chemometrics: Chemical and Biochemical Data Analysis (Four-Volume Set). 1ªed.Amsterdã: ELSEVIER, v. 03, (2009,) 233-283.

[9] M. B. Seasholtz, B. Kowalski, The parsimony principle applied to multivariate calibration, Analytica Chimica Acta 277 (1993) 165-177.

[10] Leardi R, Application of genetic algorithm-PLS for feature selection in spectral data sets. Journal of Chemometrics. 14: (2000) 643-655.

[11] F. Allegrini, A. C. Olivieri, (A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least - squares analysis. Analytica Chimica Acta 699: (2011) 18-25.

[12] C.S. Munita, L.P. Barroso, P. M. S. Oliveira, Stopping rule for variable selection using stepwise discriminant analysis. Journal of Radioanalytical and Nuclear Chemistry 269: (2006) 335-338.

[13] M.C. U Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems. 57: (2001) 65-73.

[14] M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D.P. Neto, G.E. José, T.C.B. Saldanha, The successive projections algorithm for spectral variable selection in classification problems. Chemometrics and Intelligent Laboratory Systems 78: (2005) 11-18.

[15] W. L. Yip T. C. Soosainather, K. Dyrstad, S. A. Sande, Classification of structurally related commercial contrast media by near infrared spectroscopy. Journal of Pharmaceutical and Biomedical Analysis 90 (2014) 148 - 160

[16] G. Foca, D. Salvo, A. Cino, C. Ferrari, D. P. Lo Fiego, G. Minelli, A. Ulrici. Classification of pig fat samples from different subcutaneous layers by means of fast and non-destructive analytical techniques. Food Research International 52 (2013) 185–197.

[17] R. S. Fernandes, F. S. L. Costa, P. Valderrama, P. H. Março, K. M. G. Lima. Journal of Pharmaceutical and Biomedical Analysis 66 (2012) 85– 90

[18] E. Ferrari, G. Foca, M. Vignali, L. Tassi, A. Ulrici, Adulteration of the anthocyanin content of red wines: Perspectives for authentication by Fourier Transform-Near InfraRed and ^{1}H NMR spectroscopies. Analytica Chimica Acta 701 (2011) 139– 151.

[19] C. V. Di Anibal, M. P. Callao, I. Ruisánchez, ^{1}H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs. Talanta 86 (2011) 316– 323.

[20] M. Bevilacqua, R. Nescatelli, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, Chemometric classification techniques as a tool for solving problems in analytical chemistry, J. AOAC Int. 97 (2014) 19-28.

[21] D. Ballabio, R. Todeschini, Multivariate classification for qualitative analysis, in: D-W. Sun (Ed.), Infrared spectroscopy for food quality analysis and control, Academic Press / Elsevier, Burlington, MA, 2009, pp. 83-104.

[22] M. Forina, P. Oliveri, S. Lanteri, M. Casale, Class-modeling techniques, classic and new, for old and new problems, Chemom. Intell. Lab. Syst. 93 (2008) 132-148.

[23] K. S. Booksh, B. R. Kowalski, Theory of Analytical Chemistry, Analytical Chemistry, 66: (1994) 15, 1.

[24] A. C. Olivieri, Recent advances in analytical calibration with multi-way data. Analytical. Methods 4 (2012) 1876- 1886.

[25] D. A. E. Skoog, J. J. Leary, Principles of instrumental analysis. 6. ed. New York: Saunders College Publishing, 1992.

[26] K.R. Beebe; R. J. Pell, B. Seasholtz, Chemometrics – A Pratical Guide. New York: Wiley, 1998.

[27] H. C. Goicoechea, A.C. Olivieri, MULTIVAR. A program for multivariate calibration incorporating net analyte signal calculations. TrAC Trends in Analytical Chemistry 19 (2000) 599-605.

[28] L. Hu, M. Ye, X Jiang, S. Feng, H, Zou, Advances in hyphenated analytical techniques for shotgun proteome and peptidome analysis—A review. Analytica Chimica Acta 598 (2007) 193–204.

[29] G. Lespes, J. Gigault, Hyphenated analytical techniques for multidimensional characterisation of submicron particles: A review. Anal. Chim. Acta 692 (2011) 26–41.

[30] H. Parastar, J. R. Radovic, M. Jalali-Heravi, S. Diez, J.M. Bayona, R. Tauler, Resolution and quantification of complex mixtures of polycyclic aromatic hydrocarbons in heavy fuel oil sample by means of GC × GC-TOFMS combined to multivariate curve resolution. Analytical Chemistry 83 (2011) 9289–9297.

[31] C. Pasquini, Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. Journal of the Brazilian Chemical Society. 14 (2003) 198-219.

[32] B.K. Lavine, W.S. Rayens, Classification: basic concepts. In: S.D. Brown, R. Rauler, B. Walczak (Eds.), Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. Vol. 3: Linear Regression Modeling. Non-Linear Regression Modeling. Classification. Feature Selection. Multivariate Robust Techniques, Elsevier, New York (2009), pp. 507–515.

[33] L. A. Berrueta, R. M. Alonso-Salces, K. Héberger, Supervised pattern recognition in food analysis. Journal of Chromatography A, v.1158, p.196–214, 2007.

[34] A. G. González, Use and misuse of supervised pattern recognition methods for interpreting compositional data. Journal of Chromatography A, (2007) 1158, 215–225.

[35] K.R. Beebe; R.J. Pell, B. Seasholtz. **Chemometrics – A Pratical Guide**. New York: Wiley, 1998.

[35] P. R. M. Correia, M. M. C. Ferreira, Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos. Química Nova, (2007) 30, 481–487.

[36] A. Candolfi, R. D. Maesschalck, , D. L. Massart; P. A. Hailey, A. C. E. Harrington, Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. Journal of Pharmaceutical and Biomedical Analysis, (1999) 19 923–935.

[37] M. Barker, W. Rayens, Partial least squares for discrimination. Journal of Chemometrics 17: (2003)166-173.

[38] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: taking the magic away. Journal of Chemometrics 28: (2014) 213-225.

[39] M. R. Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, R.J. Poppi, Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchemical Journal 109: (2013) 170-177.

[40] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models PLS-DA. Analitycal Methods. 5: (2013) 3790-3798.

[41] M. M. C. Ferreira, **Quimiometria - conceitos, métodos e aplicações**, Campinas-SP, Editora da UNICAMP, 2015.

[42] M. J. C. Pontes, **Algoritmo das projeções sucessivas para a seleção de variáveis espectrais em problemas de classificação**. 2009.123f Tese (Doutorado em Química) Curso de Pós Graduação em Química, Universidade Federal da Paraiba.

[44] Wold, S. Pattern recognition by means of disjoint principal component models, **Pattern Recognition**, handbook. Capitulo 33. p. 203 – 241. 1976.

[45] R. G. Brereton, Chemometrics: data Analysis for the laboratory and chemical plant. New York: John Wiley & Sons, 2003.

[46] D. L. Massart, et al., Handbook of Chemometrics and Qualimetrics: Part A. Amsterdam: Elsevier Science B. V., 1997.

[47] C. B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context. Chemometrics and Intelligent Laboratory Systems. 19: (1993) 1-33.

[48] R.K.H. Galvão, M. C. U. Araújo, Variable Selection. in: Walczak B; Tauler R; Brown S, (Eds) Comprehensive Chemometrics. Elsevier Inc. Oxford (2009) 233-283.

[49] S.F.C. Soares, A. A. Gomes, A. R. G. Filho, M. C. U. Araújo, R. K. H. Galvão The successive projections algorithm. TrAC Trends in Analytical Chemistry. 42: (2013) 84-98.

[50] A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Véras, E. C. Silva, The successive projections algorithm for interval selection in PLS. Microchemical Journal 100: (2013) 202-208.

[51] M.C. U Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemometrics and Intelligent Laboratory Systems. 57: (2001) 65-73.

[52] P. H. G. D. Diniz, M. F. Pistonesi, M. C. U. Araújo, Using iSPA-PLS and NIR spectroscopy for the determination of total polyphenols and moisture in commercial tea samples. Analytical. Methods 7: (2015) 3379-3384.

[53] M. Ghasemi-Varnamkhasti.; S. S Mohtasebi, M. L. Rodriguez-Mendeza, A. A. Gomes, M. C. U. Araújo, R. K. H. Galvão, Screening analysis of beer ageing using near infrared spectroscopy and the Successive Projections Algorithm for variable selection. Talanta 89: (2012) 286–291.

[54] U. T. C. P. Souto, M. J. C. Pontes, E.C. Silva, R. K. H. Galvão, M. C. U. Araújo, F. A. C. Sanches, F. A. S. Cunha, M. S. R. Oliveira, UV–Vis spectrometric classification of coffees by SPA–LDA. Food Chemistry 119: (2010) 368–371.

[55] C. B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context. Chemometrics and Intelligent Laboratory Systems. 19: (1993) 1-33.

[56]A. Höskuldsson, Variable and subset selection in PLS regression. Chemometrics and Intelligent Laboratory Systems 55: (2001) 23-38.

[57] C. H, Spiegelmen, M. J. McShane, M. J. Goetz, M. Motamedi , Q. L. Yue, G. L. Coté, Theoretical Justification of Wavelength Selection in PLS Calibration: Development of a New Algorithm Analytical Chemistry 70: (1998) 35-44.

[58] M. M. C. Ferreira, A. M. Antunes, M. S. Melgo, L.O.Volpe. Quimiometria I: calibração multivariada, um tutorial. Química Nova (1999).22 724–731

[59] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S. B. Engelsen, Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. Applied Spectroscopy 54: (2000) 413-419.

[60] Diniz, P. H. G. D., Barbosa, M. F., Milanez, K. D. T. M., Pistonesi, M. F., Araújo, M. C. U. Using UV–Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. Food Chemistry, 192 (2016) 374–379.

[61] G. A. Helfer, L. F. F. Silva, F. C. Bock, L. Marder, Análise de componentes principais por intervalos (iPCA) como método de seleção de região espectral no infravermelho médio e próximo para discriminação de óleos vegetais, Tecno-Lógica. 17 (2013) 108-116.

[62] R. Leardi, Lars Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, Journal of Chemometrics 18: (2004) 486-497

[63] M.Shariati-Rad, M. Hasani, Selection of individual variables versus intervals of variables in PLSR, Journal of Chemometrics 24: (2010) 45–56.

[64] A. A. Gomes, M. R. Alcaraz, H. C. Goicoechea, M. C. U. Araújo, The Successive Projections Algorithm for interval selection in trilinear partial least-squares with residual bilinearization, Analytica Chimica Acta 811: (2014) 13-22.

[65] A.A. Gomes, A.V. Schenone, H.C. Goicoechea, M.C.U. Araújo., Unfolded partial least squares/residual bilinearization combined with the Successive Projections Algorithm for interval selection: enhanced excitation-emission fluorescence data modeling in the presence of the inner filter effect, Anal Bioanal Chem. 407 (2015) 5649-5659.

[66] D. Suhandy, M. Yulia, Y. Ogawa, N. Kondo, Prediction of L-Ascorbic Acid using FTIR-ATR Terahertz Spectroscopy Combined with Interval Partial Least Squares (iPLS) Regression, Engineering in Agriculture, Environment and Food 3: (2013) 111-117.

[67] A. Borin, R. J. Poppi, Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil, Vibrational Spectroscopy 37: (2005) 27-32.

[68] Z. Xiaobo, Z. Jiewen, L. Yanxiao, Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of 'Fuji' apple based on BiPLS and FiPLS models, Vibrational Spectroscopy 44:(2007) 220-227.

[69] X. Li, C. Sun, L. Luo, Y. He, Determination of tea polyphenols content by infrared spectroscopy coupled with iPLS and random frog techniques doi:10.1016/j.compag.2015.01.005.

[70] Q. Kang, Q. Ru, Y. Liu, L. Xu, J. Liu, Y. Wang, Y. Zhang, H. Li, Q. Zhang, Q. Wu, On-line monitoring the extract process of Fu-fang Shuanghua oral solution using near

infrared spectroscopy and different PLS algorithms, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 152: (2016) 431-437.

[71] M. F. Ferrão, M. S. Viera, R. E. P. Pazos, D. Fachini, A. E. Gerbase, L. Marder, Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions, Fuel 90: (2011) 701-706.

[72] N. C. T. Mariani, G. H. A. Teixeira, M. G. Lima, T. B. Morgenstern, V. Nardini, L. C. C. Júnior, NIRS and iSPA-PLS for predicting total anthocyanin content in jaboticaba fruit, Food Chemistry 174: (2015) 643–648.

[73] B. K. Lavine, **Validation of classifer**. in: Walczak B; Tauler R; Brown S, (Eds) Comprehensive Chemometrics. Elsevier Inc. Oxford (2009) 587–599.

[74] R. L. Carneiro, J. W.B. Braga, C. B.G. Bottoli, R. J. Poppi. *A*pplication of genetic algorithm for selection of variables for the BLLS method applied to determination of pesticides and metabolites in wine. **Anal. Chim. Acta** 595 (2007) 51–58.

[75] Kennard, R. W., & Stone, L. A. Computer aided design of experiments. Technometrics 11 (1969) 137–148.

[76] F. F. Gambarra-Neto, G. Marino, M. C. U. Araújo, R. K. H. Galvão, M. J. C. Pontes, E. P. Medeiros, R. S. Lima, Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis, Talanta 77 (2009) 1660–1666.

[77] D. D. S. Fernandes, A. A. Gomes, M. M. Fontes, G. B. Costa, V. E. Almeida, M. C. U. Araújo, R. K. H. Galvão, G. Véras, UV-Vis Spectrometric Detection of Biodiesel/Diesel Blend Adulterations with Soybean Oil, Journal of the Brazilian Chemical Society. 25 (2014) 169-175

[78]G. B. Costa, D. D. S. Fernandes, A. A. Gomes, V. E. Almeida, G. Veras, Using near infrared spectroscopy to classify soybean oil according to expiration date, Food Chemistry 196 (2016) 539–543.

Anexo

http://dx.doi.org/10.5935/0103-5053.20130259
J. Braz. Chem. Soc., Vol. 25, No. 1, 169-175, 2014.
Printed in Brazil - ©2014 Sociedade Brasileira de Química
0103 - 5053 $6.00+0.00

Short Report

UV-Vis Spectrometric Detection of Biodiesel/Diesel Blend Adulterations with Soybean Oil

David D. S. Fernandes,[a] Adriano A. Gomes,[a] Marcelo M. de Fontes,[b] Gean B. da Costa,[b] Valber E. de Almeida,[c] Mario C. U. de Araújo,[a] Roberto K. H. Galvão[d] and Germano Véras[*,c]*

[a]*Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Departamento de Química, Universidade Federal da Paraíba, Caixa Postal 5093, 58051-970 João Pessoa-PB, Brazil*

[b]*Programa de Pós-Graduação em Ciências Agrárias and* [c]*Departamento de Química, CCT, Universidade Estadual da Paraíba, 58429-500 Campina Grande-PB, Brazil*

[d]*Divisão de Engenharia Eletrônica, Instituto Tecnológico de Aeronáutica, 12228-900 São José dos Campos-SP, Brazil*

Um método para detecção de adulterações de misturas de biodiesel/diesel (B5) com óleo de soja empregando espectrometria UV-Vis é proposto. O estudo envolve 90 amostras compreendendo misturas B5 com e sem a adição de óleo de soja (0,5 a 2,5% v/v). Uma discriminação apropriada foi obtida utilizando classificadores SIMCA (modelagem independente e flexível por analogia de classe), KNN (K-vizinhos mais próximos), PLS-DA (análise discriminante por mínimos quadrados parciais) e SPA-LDA (análise discriminante linear com algoritmo de projeções sucessivas).

A method for detecting adulterations of biodiesel/diesel blends (B5) with soybean oil using UV-Vis spectrometry is proposed. The study involves 90 samples comprising B5 blends with and without the addition of soybean oil (0.5 to 2.5% v/v). Suitable discrimination was achieved by using SIMCA (soft independent modeling of class analogy), KNN (K-nearest neighbors), PLS-DA (partial least squares discriminant analysis) and SPA-LDA (linear discriminant analysis with spectral variables selected by the successive projections algorithm) classifiers.

Keywords: biodiesel, adulteration, UV-Vis spectrometry, multivariate classification

Introduction

Since 2010, Brazilian regulations state that diesel fuel must be blended with 5% biodiesel prior to commercial distribution. This blend, termed B5, may have a variation of up to ± 0.5% (v/v) in biodiesel content, as established by the Brazilian national fuel authority (Agência Nacional de Petróleo, Gás natural e Biocombustível-ANP).[1] Within this scenario, concerns may be raised with regard to adulterations of B5 blends with raw vegetable oil,[2-11] which could be added by fuel retailers to increase profits. Such adulterations cause increase of engine wear[12] and constitute a crime against the popular economy.

The analytical method recommended by ANP for determination of biodiesel in diesel is based on the European standard EN 14078.[13] This method employs a single wavelength in the mid-infrared region, namely 5730 nm (1745 cm^{-1}), which corresponds to the peak of stretching band of carbonyl.[13] However, since this band is also found in vegetable oils, the reference method is unable to discriminate B5 blends from mixtures of diesel, biodiesel and vegetable oil. Such a discrimination cannot be carried out on the basis of refractive index, density or viscosity, either. In fact, diesel, biodiesel and vegetable oil all have values ranging from 0.82 to 0.92 kg m^{-3} for density[14-16] at 20 °C and from 1.4 to 1.5 for refractive index.[16] A better alternative might lie in the use of viscosity, which exhibits distinct values for vegetable oil, as compared to diesel and biodiesel. Viscosity values for soybean oil,[17] for example, range from 58.5 to 62.2 $mm^2 s^{-1}$, which is substantially larger compared to diesel (2.0-4.5 $mm^2 s^{-1}$)[14] and biodiesel (3.0-6.0 $mm^2 s^{-1}$).[15] However, as shown in the Supplementary Information, adulterations

*e-mail: germano.veras@pq.cnpq.br

Analytical Methods

PAPER

View Article Online
View Journal | View Issue

Cite this: *Anal. Methods*, 2016, 8, 7522

The successive projections algorithm for interval selection in partial least squares discriminant analysis

David Douglas de Sousa Fernandes,[a] Valber Elias Almeida,[a] Licarion Pinto,[a] Germano Véras,[b] Roberto Kawakami Harrop Galvão,[c] Adriano Araújo Gomes[d] and Mário Cesar Ugulino Araújo*[a]

This paper proposes a new interval selection approach for PLS-DA modelling, which is developed as an extension of the recently introduced iSPA-PLS method for multivariate calibration. The proposed iSPA-PLS-DA algorithm is tested in two case studies concerning the classification of five types of vegetable oils employing square-wave voltammetry and the classification of five species of bacteria (*Escherichia coli, Enterococcus faecalis, Streptococcus salivarius, Streptococcus oralis,* and *Staphylococcus aureus*) using digital images. For comparison, the iPLS-DA algorithm for interval selection is also employed, in standard and backward modes. In both case studies, iSPA-PLS-DA provided correct classification rates larger than or equal to those obtained by PLS-DA using all variables and iPLS-DA with single or multiple intervals.

Received 28th June 2016
Accepted 15th September 2016

DOI: 10.1039/c6ay01840h

www.rsc.org/methods

1 Introduction

The development of new analytical methods often benefits from the use of chemometrics modelling to handle data from multiple analytical channels.[1–3] Examples include near infrared (NIR),[4] fluorescence[5] and molecular absorption[6] spectrometries, voltammetric techniques[7] and methods based on digital images.[2,8–10] In this context, an important challenge consists in dealing with non-informative or redundant variables.[11] In fact, the use of a reduced set of representative variables often leads to multivariate calibration models with better prediction ability,[12–14] which is in line with the so-called Parsimony Principle.[15]

In many analytical problems, the variables are naturally disposed in a sequential order corresponding *e.g.* to wavelengths in a spectrum,[11] potentials in a voltammogram,[7] or variables or intervals of variables. In multiple linear regression, the selection of neighbour variables is usually avoided in order to prevent multi-collinearity problems.[17] However, in latent variable methods such as partial least squares (PLS) regression, the selection of intervals of variables is a more common approach.[18–22] A simple strategy employed in the well-known interval partial least squares (*i*PLS) method[18] consists of dividing the variables into intervals of equal width and selecting the best interval on the basis of a prediction performance criterion. For this purpose, each interval of variables is employed to build a PLS model, which is evaluated by using cross-validation or validation in an external set of samples.

Within this scope, the use of interval selection for classification purposes has received comparatively little attention. A typical approach consists of using the *i*PLS method with the

Apêndice 1

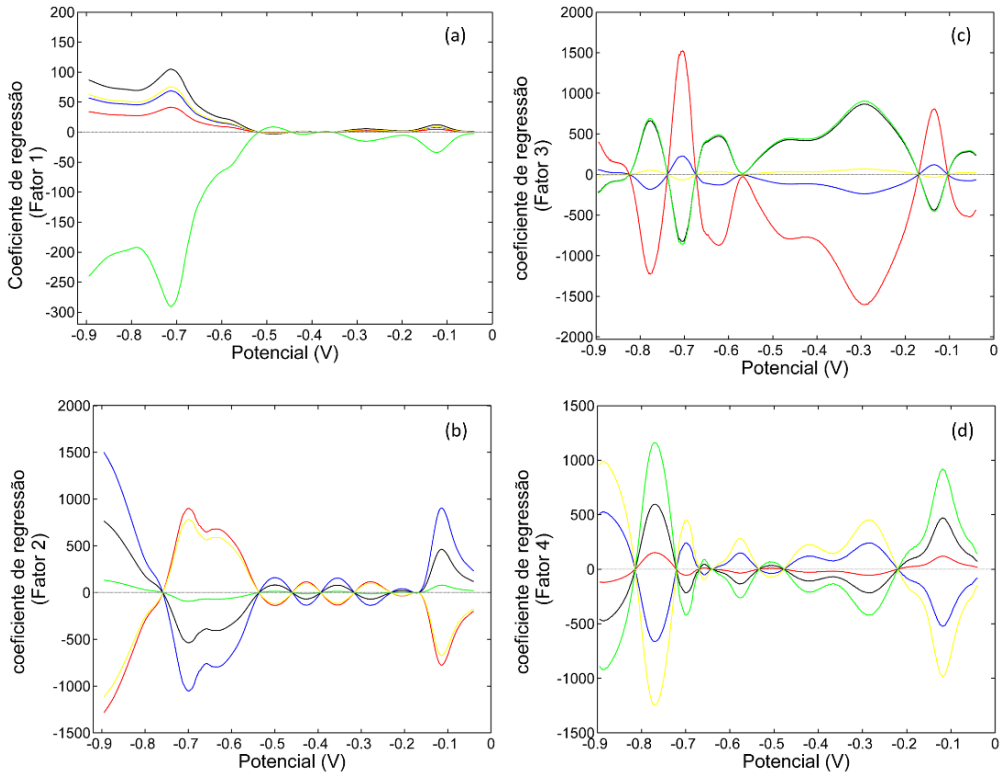


Figura 22 - (a-d) Coeficientes de regressão para o PLS-DA para o estudo: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (Fonte própria).

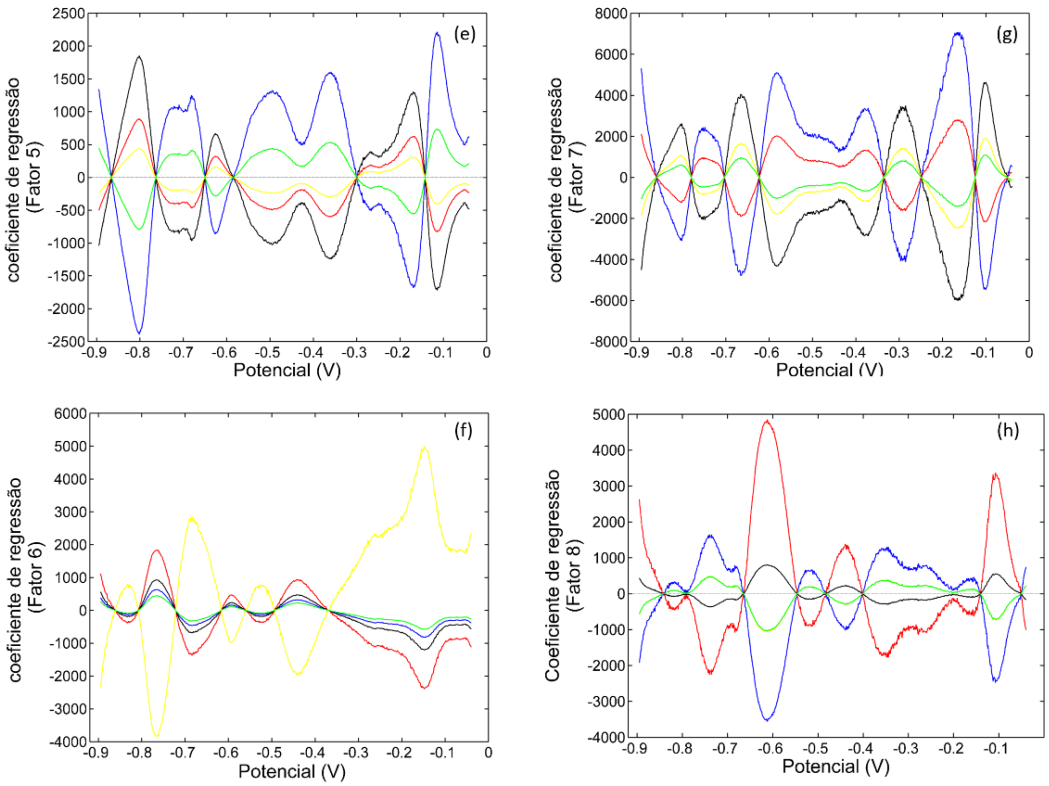


Figura 23 - (e-h) Coeficientes de regressão para o PLS-DA para o estudo: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (Fonte própria).

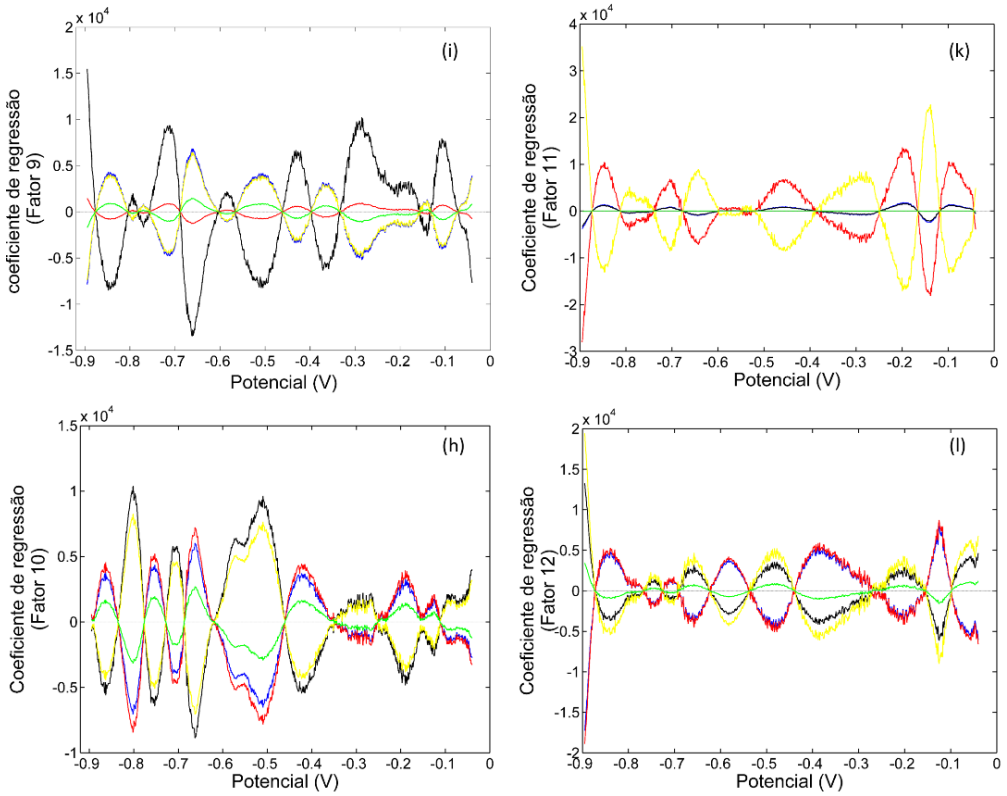


Figura 24 - (i-l) Coeficientes de regressão para o PLS-DA para o estudo: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (Fonte própria).

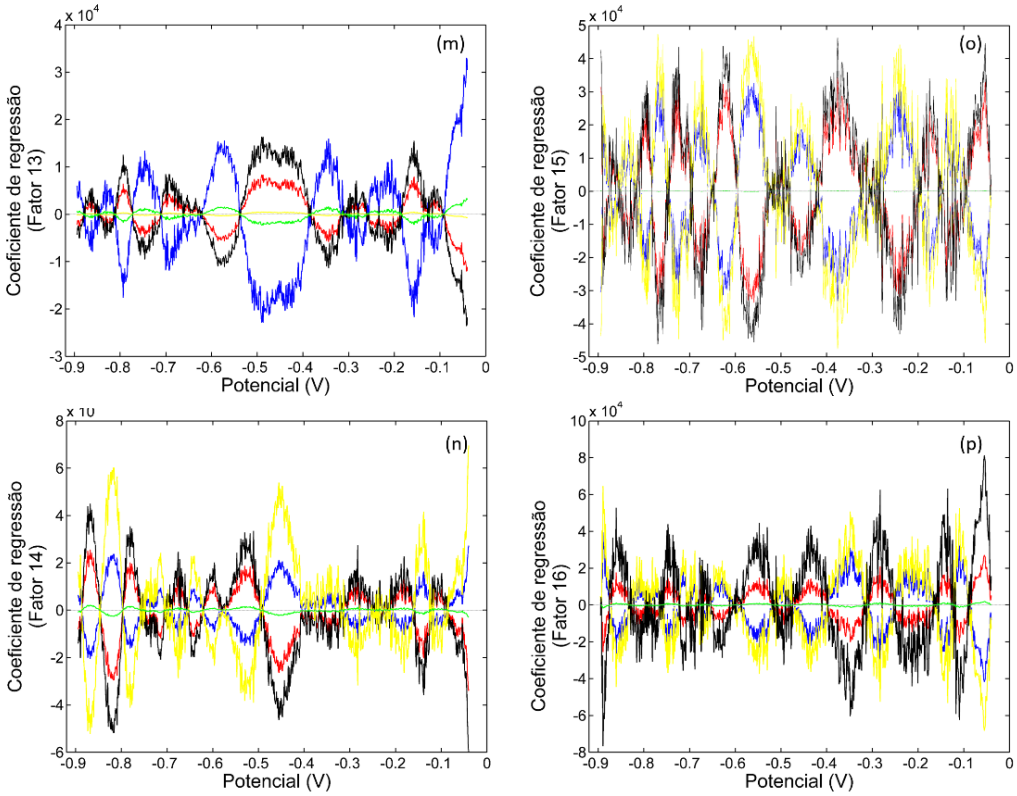


Figura 25 - (m-p) Coeficientes de regressão para o PLS-DA para o estudo: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (Fonte própria).

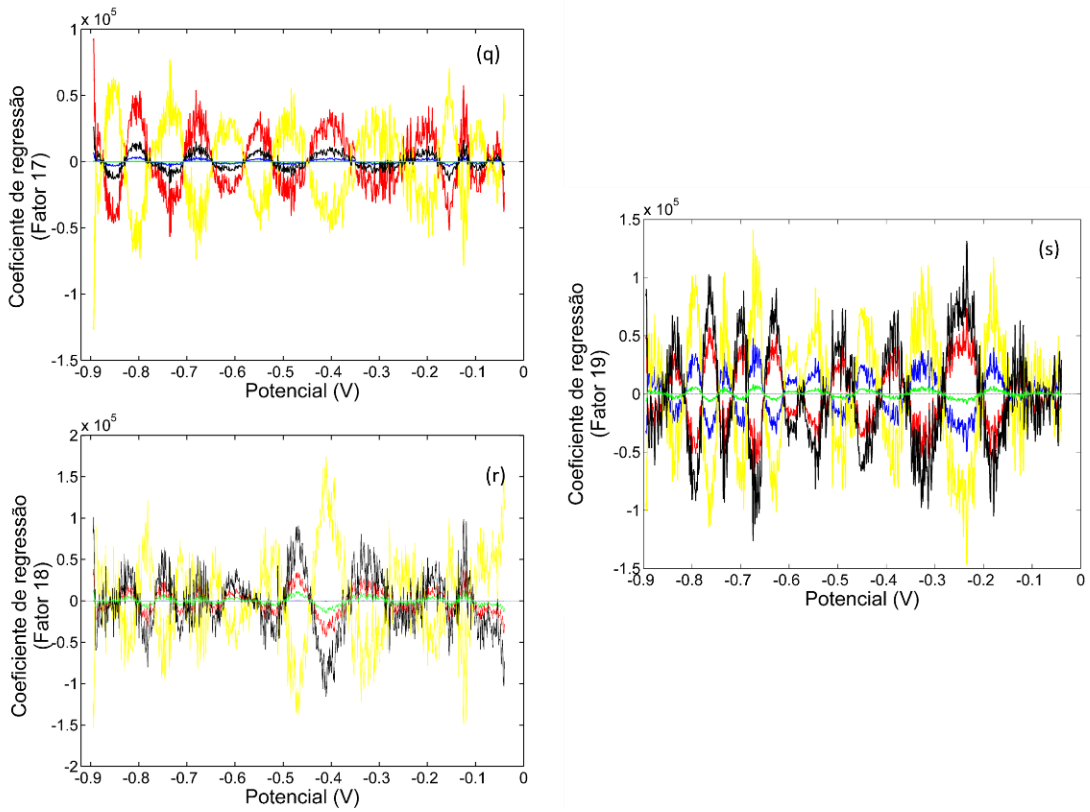


Figura 26 - (q-s) Coeficientes de regressão para o PLS-DA para o estudo: classificação de óleos vegetais com relação ao tipo de matéria-prima e ao prazo de validade obtidos por voltametria de onda quadrada para as 114 amostras de óleos vegetais estudadas (Fonte própria).

Apêndice 2

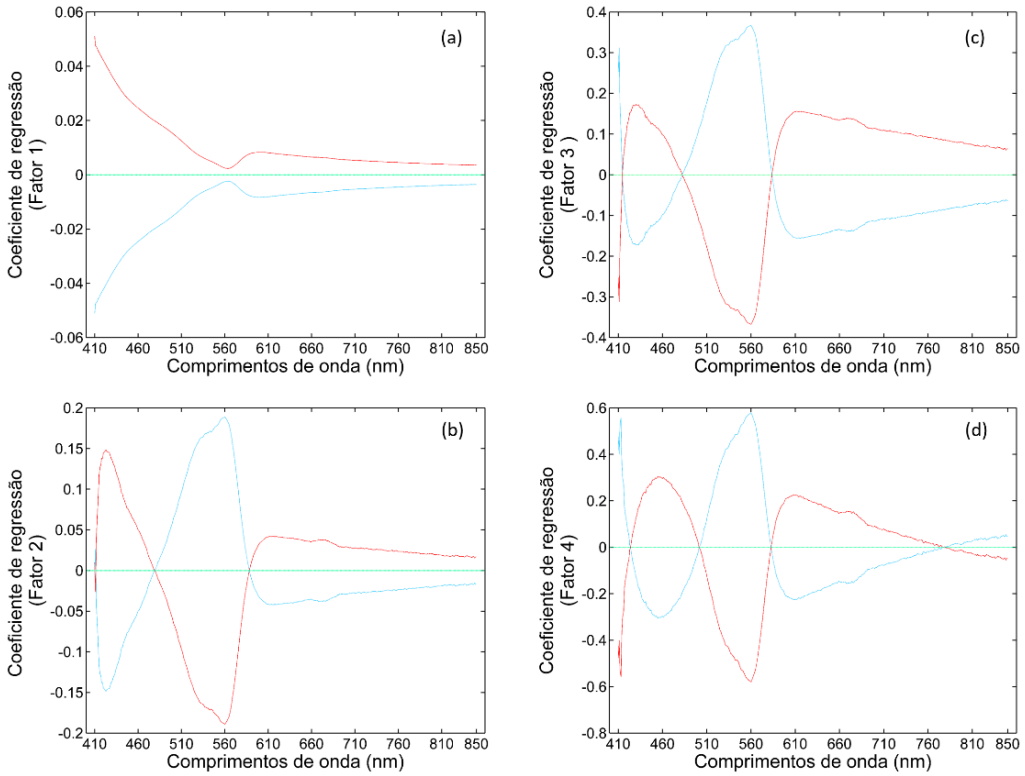


Figura 27 - (a-d) Coeficientes de regressão para o PLS-DA para o estudo: identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal obtidos por espectroscopia UV-vis para as 90 amostras de misturas biodiesel/diesel estudadas (Fonte própria).

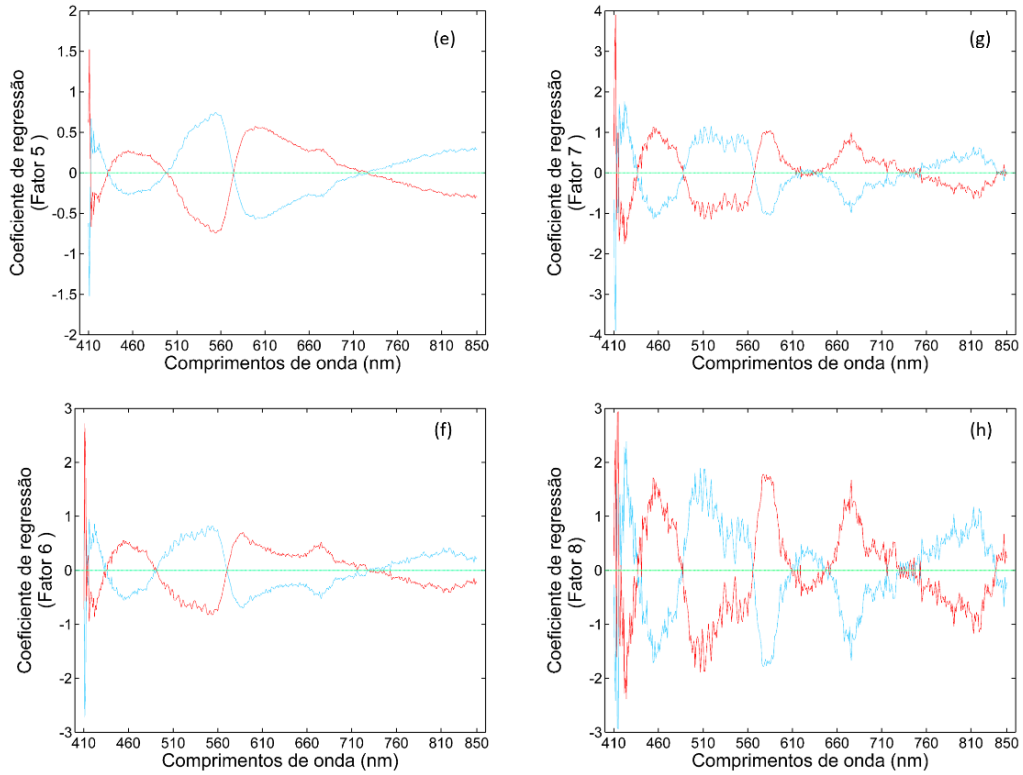


Figura 28 - (e-h) Coeficientes de regressão para o PLS-DA para o estudo: identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal obtidos por espectroscopia UV-vis para as 90 amostras de misturas biodiesel/diesel estudadas (Fonte própria).

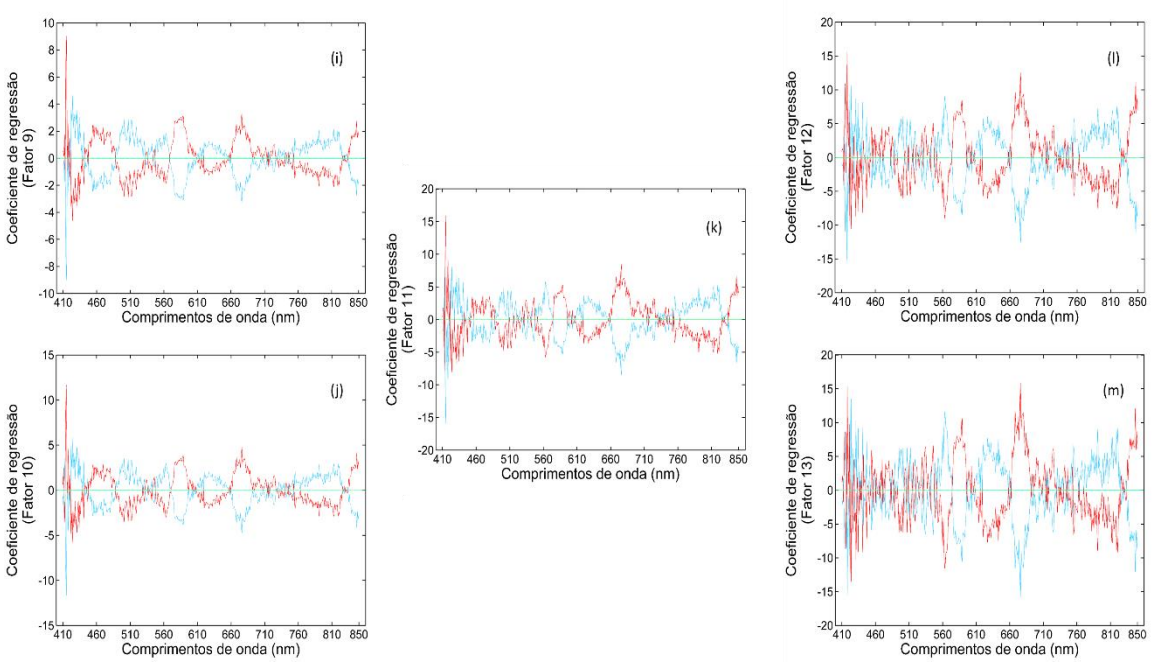


Figura 29 - (i-m) Coeficientes de regressão para o PLS-DA para o estudo: identificação de adulteração de misturas de biodiesel/diesel (B5) com óleo vegetal obtidos por espectroscopia UV-vis para as 90 amostras de misturas biodiesel/diesel estudadas (Fonte própria).

Apêndice 3

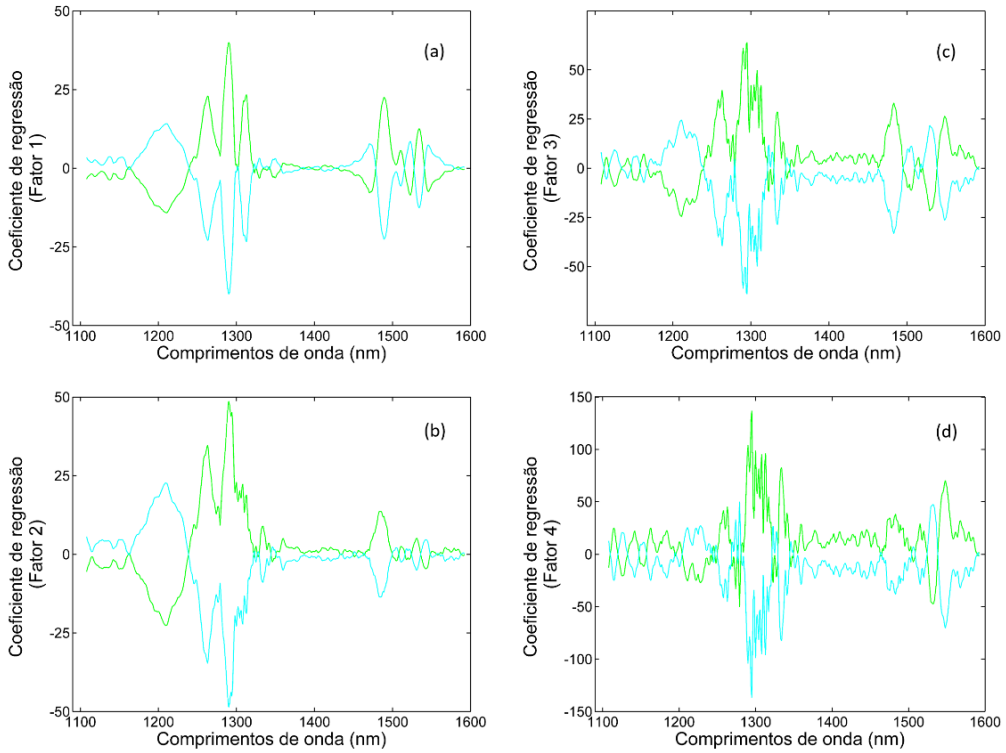


Figura 30 - (a-d) Coeficientes de regressão para o PLS-DA para o estudo de caso 3: classificação de oloes vegetais com relação ao prazo de validade obtidos por espectroscopia do infravermelho para as 50 amostras de óleos vegetais (Fonte própria).

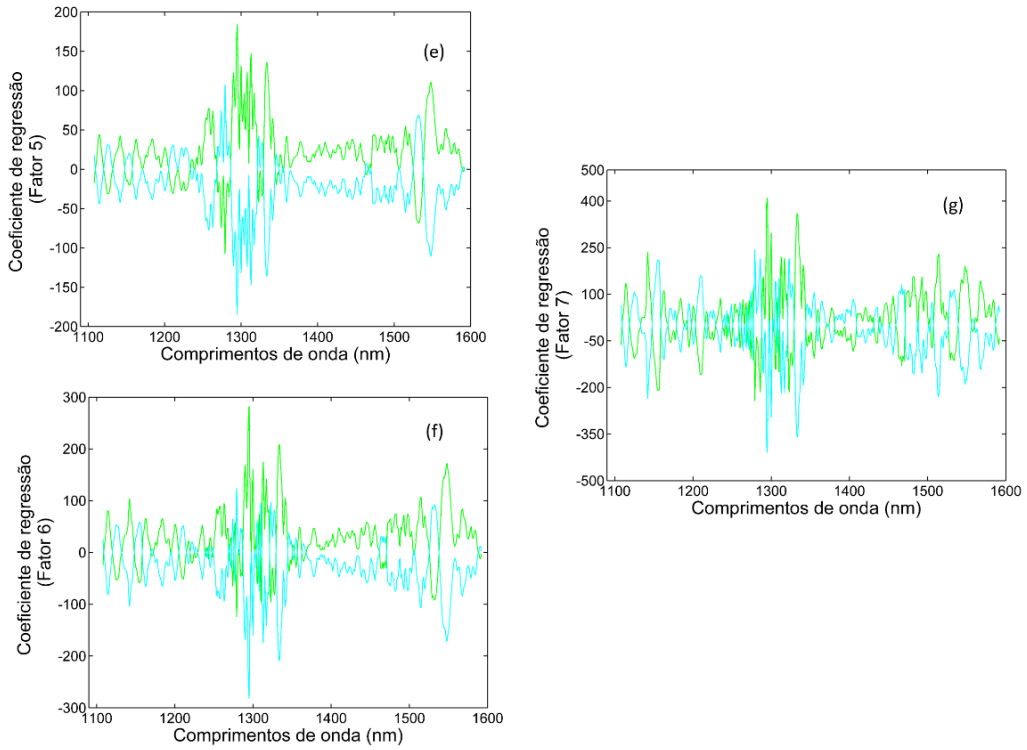


Figura 31 - (e-g) Coeficientes de regressão para o PLS-DA para o estudo de caso 3: classificação de oloes vegetais com relação ao prazo de validade obtidos por espectroscopia do infravermelho para as 50 amostras de óleos vegetais (Fonte própria).

Rotina principal do programa *i*SPA-PLS-DA é apresentado a seguir.

Algoritmo Das Projeções Sucessivas Aplicado a Seleção de Variáveis em Analise Discriminante Linear Por Minimos Quadrados Parciais

O Algoritmo iSPA_PLS_DA Utiliza a seleção de intervalos baseado no citerio de projeções do APS convencional acoplado a modelos de analise discriminante linear por minimos quadrados parciais

DADOS DE ENTRADA:
Train: Matriz (MxJ) das variáveis independentes do conjunto de Treinamento.
Group_Train: Vetor coluna (Mx1) contendo os valores referente as classes do conjunto de Treinamento.
Val:Matriz (NxJ)das variáveis independentes do conjunto de Validação.
Group_Val:Vetor coluna (Nx1) contendo os valores referente as classes do conjunto de Validação.
Test:Matriz (TxJ) das variáveis independentes do conjunto de Predição.
Group_Test:Vetor coluna (Tx1) contendo os valores referente as classes do
intervalos:quantidade de intervalos que o espectro deve ser dividido.
I_max: número máximo de intervalos que devem ser selecionados.
VL: número de variáveis latentes que devem ser usados no calculo do modelo PLS_DA global.
Essa versão estar disponivel apenas validação cruzada, assim Val e Group_Val devem ser subtituidos por atrizes vazias. (Ex.: Val=[]; Group_Val=[];).
Versão 1.0.
Desenvolvido por: David Douglas de Sousa Fernandes.
daviddsf013@gmail.com

REFERENCIAS CONSULTADAS.
[1] S. F. C. Soares, A. A. Gomes, A. R. Galvão Filho, M. C. U. Araujo, R. K. Harrop. Galvão, The Successive Projections Algorithm. Trends in Analytical Chemistry, 42 (2013) 84-98.
[2] A. A. Gomes, R. K. H. Galvão, M. C. U. Araújo, G. Véras, E. C. Silva, The successive projections algorithm for interval selection in PLS. Microchem. J. 110 (2013) 202–208.

```
functionModelo_iAPS_PLS_DA=iAPS_PLS_DA(Train,Group_Train,Val,Group_Val,T
est,Group_Test,intervalos,I_max,VL);
warning off
intervals=intervalos;
N1=1;
N2=I_max;
[nsampval]=size(Val,1);
%validação cruzada
if nsampval==0
   [Nmis_cal,Nlambdas]=size(Train);
end
```

```
metodo = 0;
while (metodo ~= 1) & (metodo ~= 2)
   disp('   ')
   metodo = input('Tecle 1 caso conheça o nº ótimo LVs e 2 caso não conheça,
determine: ');
end

if metodo == 1
   disp('   ')
   disp('Qual o nº ótimo e tecle <ENTER>' )

  Fatores_sugeridos=input('Qual o número de fatores deve ser usado?  ');
  comp=Fatores_sugeridos;
  disp('Escolha o método dimensionamento de dados e tecle <ENTER>')
disp('1 - Centrado na média')
disp('2 - Autoescalamento')
disp('3 - Sem escala')
pret_type = input('Opcao: ');
switch pret_type
   case 1
B=('cent');
   case 2
B=('auto');
   case 3
B=('none');
otherwise
disp('cent')
end
pret_type=B;
disp('------------------------------------------------------------------------')
disp(' ')
disp('Escolha tipo de validação cruzada e tecle <ENTER>')
disp('1 - venetian blinds')
disp('2 - contiguous blocks')
cv_type = input('Opcao: ');
switch cv_type
   case 1
C=('vene');
   case 2
C=('cont');
otherwise
disp('vene')
end
cv_type=C;
disp('------------------------------------------------------------------------')
disp(' ')
cv_group=size(Train,1);
disp(['Número total de amostras de treinamento ',num2str(cv_group)])
Numero_de_Amostra_de_Treinamento = input('Qual o número de cv-groups deseja
usar?    ');
```

```
cv_groups=Numero_de_Amostra_de_Treinamento;
if  cv_groups > cv_group
   error('cv_groups deve ser menor ou igual ao numero total de amostras de
treinamento')
end
disp('------------------------------------------------------------------')
disp(' ')
disp('Escolha o critério de atribuição e tecle <ENTER>')
disp('1 - Baseados no teorema de Bayes')
disp('2 - Máximo')
assign_method = input('Opcao: ');
switch assign_method
    case 1
E=('bayes');
    case 2
E=('max');
 otherwise
    disp('max')
end
P=assign_method;
assign_method=E;
class=Group_Train;
doqtlimit=1;
   if size(Fatores_sugeridos)==0
    comp=Fatores_sugeridos;
    end
%close all
%Particionando o espectro em I intervalos

[X]=fun_part(Train,intervals);
normas=[];
norm_max=[];
for i=1:size(X,2)
   a=X{1,i};
    x=Train(:,a);
   for j=1:size(x,2)
       b=norm(x(:,j));
       normas=[normas b];
       [A index_norm_max]=max(normas);
    end

   norm_max=[norm_max  index_norm_max];
end
iXcal=Train(:,norm_max);
%Aplicando o SPA
[L] = cadeias(iXcal,N1,N2);
 if nsampval==0
    [iNmis_cal,iNlambdas]=size(iXcal);
    R = zeros(1,N2);
    rmsep = [];
```

```
    Lopt = zeros(N2,N2);
     temporizador = waitbar(0,'Calculo dos modelos PLS para as cedeias geradas pelo
SPA ...');
   loopStart = now;
     for N = N1:N2
       loopEnd = loopStart + (now-loopStart)*N2/N;
      waitbar(N/N2,temporizador,['Aplicando PLS-DA nos intervalos. '
datestr(loopEnd)]);
       for i = 1:iNlambdas
          lambdas = L(1:N,i);
          z=[];
          for b = 1:size(lambdas,1);
             g=lambdas(b,1);
             z=[z,X{1,g}];
          end
          Xcal2=Train(:,z);
          cv = plsdacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);

          min_erros=cv.class_param.er;
          rmsep(N,i)=min_erros;% escolha do menor conjunto de erro.

       end
[R(N) imin] = min(rmsep(N,:));
       Lopt(1:N,N)=L(1:N,imin);
     end
    close (temporizador)
    [Rbest,Nbest] = min(R(N1:N2));
    Nbest = Nbest+N1-1;
    rmsepopt = rmsep(Nbest,:);
    l = (Lopt(1:Nbest,Nbest))';
    EC=rmsep;
end
    %construção do modelo PLS para os "l" intervalos selecionados e previsão
h=[];
for f = 1:size(l,2);
    e=l(1,f);
    h=[h,X{1,e}];
Xcal2=Train(:,h);
end
model=plsdafit(Xcal2,class,comp,pret_type,assign_method,doqtlimit);%model ajustado
iSPA_PLS_DA
cv=plsdacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);

%Previsão para um conjunto externo de amostras
Xcal2=Train(:,h);
Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Group_Train-ones(Nmis_cal,1)*mean(Group_Train);
Xpred2=Test(:,h);
Xpred_cm=Xpred2-ones(size(Xpred2,1),1)*mean(Xcal2);
pred = plsdapred(Xpred2,model);%Predição das amostras
```

```
class_param = calc_class_param (pred.class_pred,Group_Test);%matrix de confusão
para conjunto de predição

[n,m]=size(Train);% determinando o número de variáveis
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];% Inicio de cada intervalo
endint=[startint(2:intervals)-1; m]; % Final de Intervalo
X=cell(1,intervals); %Iniciando X
for i=1:intervals
   x=startint(i,1):endint(i,1);
   X{i}=x;
end
[nint,mint]=size(intervals);
Modelo_iAPS_PLS_DA.resultado=intervals;
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
%
int=l;
[X]=fun_part(Train,intervals);
m=size(Train,2);
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
X=cell(1,intervals);
for i=1:intervals
   x=startint(i,1):endint(i,1);
   X{i}=x;
end
in_fin_int_sel=zeros(2,length(l));
for i=1:length(l)
   int_sel=X{l(i)};
   in_fin_int_sel(1,i)=min(int_sel);
   in_fin_int_sel(2,i)=max(int_sel);
end
rawX=(Train);
xaxislabels=size(Train,2);
xaxislabels=[1:1:xaxislabels];
Modelo_iAPS_PLS_DA.resultado.rawX=rawX;
Modelo_iAPS_PLS_DA.resultado.allint=[(1:intervals+1)' [startint;1] [endint;m]];
Modelo_iAPS_PLS_DA.resultado.intervalsequi=1;
Modelo_iAPS_PLS_DA.resultado.intcom=in_fin_int_sel;
Modelo_iAPS_PLS_DA.resultado.Train=model;
Modelo_iAPS_PLS_DA.resultado.Xtrain_sel=Xcal2;
```

```
Modelo_iAPS_PLS_DA.resultado.selected_intervals=l;
Modelo_iAPS_PLS_DA.resultado.Test=pred;
Modelo_iAPS_PLS_DA.resultado.xaxislabels=xaxislabels;
Modelo_iAPS_PLS_DA.resultado.Fatores_sugeridos=comp;
Modelo_iAPS_PLS_DA.resultado.prepro_method='mean';
Modelo_iAPS_PLS_DA.resultado.Train.cv=cv;
Modelo_iAPS_PLS_DA.resultado.Test.class_param=class_param;
Modelo_iAPS_PLS_DA.resultado.erro_rate=EC;
% %Saida gráfica
ispaplsplot(Modelo_iAPS_PLS_DA.resultado,Modelo_iAPS_PLS_DA.resultado.Fatore
s_sugeridos);
%
else
disp('Escolha o método dimensionamento de dados e tecle <ENTER>')
disp('1 - Centrado na média')
disp('2 - Autoescalamento')
disp('3 - Sem escala')
pret_type = input('Opcao: ');

switch pret_type
   case 1
B=('cent');
   case 2
B=('auto');
   case 3
B=('none');
otherwise
disp('cent')
end
pret_type=B;

disp('------------------------------------------------------------------')
disp(' ')
disp('Escolha tipo de validação cruzada e tecle <ENTER>')
disp('1 - venetian blinds')
disp('2 - contiguous blocks')
cv_type = input('Opcao: ');
switch cv_type
   case 1
C=('vene');
   case 2
C=('cont');
otherwise
disp('vene')
end
cv_type=C;

disp('------------------------------------------------------------------')
disp(' ')
cv_group=size(Train,1);
```

```
disp(['Número total de amostras de treinamento ',num2str(cv_group)])
Numero_de_Amostra_de_Treinamento = input('Qual o número de cv-groups deseja
usar?    ');
cv_groups=Numero_de_Amostra_de_Treinamento;
if  cv_groups > cv_group
  error('cv_groups deve ser menor ou igual ao numero total de amostras de
treinamento')
end
disp('-----------------------------------------------------------------')
disp(' ')
disp('Escolha o critério de atribuição e tecle <ENTER>')
disp('1 - Baseados no teorema de Bayes')
disp('2 - Máximo')
assign_method = input('Opcao: ');
switch assign_method
   case 1
E=('bayes');
   case 2
E=('max');
 otherwise
   disp('max')
end
P=assign_method;
assign_method=E;
class=Group_Train;
doqtlimit=1;
%Resultados para o número de fatores ótimo!
res = plsdacompsel(Train,Group_Train,pret_type,cv_type,cv_groups,assign_method);
Modelo_iAPS_PLS_DA.resultado.full_PLS_DA=res;
%Atribuição de classe
if P==2
erro_rate=Modelo_iAPS_PLS_DA.resultado.full_PLS_DA.er;
fatores=[1:1:19];
figure1= figure('Color',[1 1 1]);plot(fatores',erro_rate,'-k'),hold on,
plot(fatores',erro_rate,'ob');
xlabel('Fatores');
ylabel('Taxa de erro');
title('Número ótimo de fatores')
hold off;
[min_er min_ind]=min(erro_rate);

else
%Atribuição de classe
S_erro_rate=Modelo_iAPS_PLS_DA.resultado.full_PLS_DA.er;
fatores=[1:1:19];
figure1= figure('Color',[1 1 1]);subplot(2,1,1);plot(fatores',S_erro_rate,'-k'),hold on,
plot(fatores',S_erro_rate,'ob');
xlabel('Fatores');
ylabel('Taxa de erro');
title('Número ótimo de fatores')
```

```
n_erro_rate=Modelo_iAPS_PLS_DA.resultado.full_PLS_DA.not_ass;
fatores=[1:1:34];
subplot(2,1,2);plot(fatores',n_erro_rate,'-k'),hold on, plot(fatores',n_erro_rate,'or');
xlabel('Fatores');
ylabel('Amostras não atribuidas');
title('Número ótimo de fatores')
[min_n_er min_ind]=min(n_erro_rate);
end

%TREINAMENTO
  disp('------------------------------------------------------------------')
   disp(' ')
   disp(['Número de fatores sugeridos: ',num2str(min_ind)])
   disp('  ')
   Fatores_sugeridos=input('Qual o número de fatores deve ser usado?  ');
   comp=Fatores_sugeridos;


if size(Fatores_sugeridos)==0
    comp=Fatores_sugeridos;
end
%close all
%Particionando o espectro em I intervalos
[X]=fun_part(Train,intervals);
normas=[];
norm_max=[];
for i=1:size(X,2)
   a=X{1,i};
    x=Train(:,a);
   for j=1:size(x,2)
       b=norm(x(:,j));
       normas=[normas b];
       [A index_norm_max]=max(normas);
    end

   norm_max=[norm_max  index_norm_max];
end
iXcal=Train(:,norm_max);
%Aplicando o SPA
[L] = cadeias(iXcal,N1,N2);

if nsampval==0
    [iNmis_cal,iNlambdas]=size(iXcal);
    R = zeros(1,N2);
    rmsep = [];
    Lopt = zeros(N2,N2);

    temporizador = waitbar(0,'Calculo dos modelos PLS para as cedeias geradas pelo
SPA ...');
   loopStart = now;
```

```
   for N = N1:N2

      loopEnd = loopStart + (now-loopStart)*N2/N;
      waitbar(N/N2,temporizador,['Aplicando PLS-DA nos intervalos. '
datestr(loopEnd)]);
      for i = 1:iNlambdas
         lambdas = L(1:N,i);
         z=[];
         for b = 1:size(lambdas,1);
            g=lambdas(b,1);
            z=[z,X{1,g}];
         end
         Xcal2=Train(:,z);
         cv = plsdacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);

         min_erros=cv.class_param.er;
         rmsep(N,i)=min_erros;% escolha do menor conjunto de erro.

      end

      [R(N) imin] = min(rmsep(N,:));
      Lopt(1:N,N)=L(1:N,imin);

   end
   %teste
   ER=rmsep;
%     figure,surf(ER),figure,contourf(ER),pause
   close (temporizador)
   [Rbest,Nbest] = min(R(N1:N2));
   Nbest = Nbest+N1-1;
   rmsepopt = rmsep(Nbest,:);
   l = (Lopt(1:Nbest,Nbest))';
end
   %construção do modelo PLS para os "l" intervalos selecionados e previsão
h=[];
for f = 1:size(l,2);
   e=l(1,f);
   h=[h,X{1,e}];
   Xcal2=Train(:,h);
end
model=plsdafit(Xcal2,class,comp,pret_type,assign_method,doqtlimit);
%model ajustado iSPA_PLS_DA
cv=plsdacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method;
%Previsão para um conjunto externo de amostras
Xcal2=Train(:,h);
Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Group_Train-ones(Nmis_cal,1)*mean(Group_Train);
Xpred2=Test(:,h);
Xpred_cm=Xpred2-ones(size(Xpred2,1),1)*mean(Xcal2);
```

```
pred = plsdapred(Xpred2,model);%Predição das amostras
class_param = calc_class_param (pred.class_pred,Group_Test);%matrix de confusão
para conjunto de predição

[n,m]=size(Train);% determinando o número de variáveis
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];% Inicio de cada intervalo
endint=[startint(2:intervals)-1; m]; % Final de Intervalo
X=cell(1,intervals); %Iniciando X
for i=1:intervals
   x=startint(i,1):endint(i,1);
   X{i}=x;
end
[nint,mint]=size(intervals);
Modelo_iAPS_PLS_DA.resultado=intervals;
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
%
int=l;
[X]=fun_part(Train,intervals);
m=size(Train,2);
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
X=cell(1,intervals);
for i=1:intervals
   x=startint(i,1):endint(i,1);
   X{i}=x;
end
in_fin_int_sel=zeros(2,length(l));
for i=1:length(l)
   int_sel=X{l(i)};
   in_fin_int_sel(1,i)=min(int_sel);
   in_fin_int_sel(2,i)=max(int_sel);
end

rawX=(Train);
xaxislabels=size(Train,2);
xaxislabels=[1:1:xaxislabels];
Modelo_iAPS_PLS_DA.resultado.rawX=rawX;
Modelo_iAPS_PLS_DA.resultado.allint=[(1:intervals+1)' [startint;1] [endint;m]];
Modelo_iAPS_PLS_DA.resultado.intervalsequi=1;
Modelo_iAPS_PLS_DA.resultado.intcom=in_fin_int_sel;
```

```
Modelo_iAPS_PLS_DA.resultado.Train=model;
Modelo_iAPS_PLS_DA.resultado.Xtrain_sel=Xcal2;
Modelo_iAPS_PLS_DA.resultado.selected_intervals=l;
Modelo_iAPS_PLS_DA.resultado.Test=pred;
Modelo_iAPS_PLS_DA.resultado.xaxislabels=xaxislabels;
Modelo_iAPS_PLS_DA.resultado.Fatores_sugeridos=comp;
Modelo_iAPS_PLS_DA.resultado.prepro_method='mean';
Modelo_iAPS_PLS_DA.resultado.Train.cv=cv;
Modelo_iAPS_PLS_DA.resultado.Test.class_param=class_param;
Modelo_iAPS_PLS_DA.resultado.erro_rate=ER;
ispaplsplot(Modelo_iAPS_PLS_DA.resultado,Modelo_iAPS_PLS_DA.resultado.Fatores_sugeridos);
end
```

Rotina principal do programa *i*SPA-SIMCA, é apresentado a seguir.

```
ALGORITMO DAS PROJEÇÕES SUCESSIVAS APLICADO A SELEÇÃO DE VARIÁVEIS EM MODELAGEM SUAVE INDEPENDENTE POR ANALOGIA DE CLASSE

O Algoritmo iAPS-SIMCA Utiliza a seleção de intervalos baseado no critério de projeções do APS convencional acoplado a modelos SIMCA

DADOS DE ENTRADA:
Train: Matriz (MxJ) das variáveis independentes do conjunto de Treinamento.
Group_Train: Vetor coluna (Mx1) contendo os valores referente as classes do conjunto de Treinamento.
Val:Matriz (NxJ)das variáveis independentes do conjunto de Validação.
Group_Val:Vetor coluna (Nx1) contendo os valores referente as classes do conjunto de Validação.
Test:Matriz (TxJ) das variáveis independentes do conjunto de Predição.
Group_Test:Vetor coluna (Tx1) contendo os valores referente as classes do
intervalos:quantidade de intervalos que o espectro deve ser dividido
I_max: número máximo de intervalos que devem ser selecionados.
VL: número de variáveis latentes que devem ser empregado no calculo do modelo PLS_DA global.

Essa versão estar disponivel apenas validação cruzada, assim Val e Group_Val devem ser subtituidos por matrizes vazias. (Ex.: Val=[]; Group_Val=[];).

Versão 1.0.

Desenvolvido por: David Douglas de Sousa Fernandes.
daviddsf013@gmail.com

REFERENCIAS CONSULTADAS.

[1] S. F. C. Soares, A. A. Gomes, A. R. Galvão Filho, M. C. U. Araujo, R. K. Harrop. Galvão, The Successive Projections Algorithm. Trends in Analytical Chemistry,  42 (2013) 84-98.
```

[2] A. A. Gomes, R. K. H. Galvão, M. C. U. Araújo, G. Véras, E. C. Silva, The successive projections algorithm for interval selection in PLS. Microchem. J. 110 (2013) 202–208.

```
FunctionModelo_iAPS_SIMCA=iAPS_SIMCA(Train,Group_Train,Val,Group_Val,Tes
t,Group_Test,intervalos,I_max);

clc, close all, warning off
intervals=intervalos;
N1=1;
N2=I_max;

[nsampval]=size(Val,1);

if nsampval==0
   %
   [Nmis_cal,Nlambdas]=size(Train);
end

  metodo = 0;
while (metodo ~= 1) & (metodo ~= 2)
   disp('   ')
   metodo = input('Tecle 1 caso conheça o nº ótimo PCs para cada classe ou 2 caso não
conheça, determine: ');
end
 if metodo == 1
   disp('   ')
  for i=1:max(Group_Train)
  disp('     ')
   PC(i)=input([' Número de PCs para a classe ',     num2str(i),' ?---->>>']);
  end
 comp=PC;
num_comp=PC;
disp('Escolha o método dimensionamento de dados e tecle <ENTER>')
disp('1 - Centrado na média')
disp('2 - Autoescalamento')
disp('3 - Sem escala')
pret_type = input('Opcao: ');
 switch pret_type
   case 1
B=('cent');
   case 2
B=('auto');
   case 3
B=('none');
otherwise
disp('cent')
end
pret_type=B;
```

```
 disp('----------------------------------------------------------------------')
disp(' ')
disp('Escolha tipo de validação cruzada e tecle <ENTER>')
disp('1 - venetian blinds')
disp('2 - contiguous blocks')
cv_type = input('Opcao: ');
switch cv_type
    case 1
C=('vene');
    case 2
C=('cont');
otherwise
disp('vene')
end
cv_type=C;
 disp('----------------------------------------------------------------------')
disp(' ')
cv_group=size(Train,1);
disp(['Número total de amostras de treinamento ',num2str(cv_group)])
Numero_de_Amostra_de_Treinamento = input('Qual o número de cv-groups deseja
usar?    ');
cv_groups=Numero_de_Amostra_de_Treinamento;
if  cv_groups > cv_group
   error('cv_groups deve ser menor ou igual ao numero total de amostras de
treinamento')
end
disp('----------------------------------------------------------------------')
disp(' ')
disp('Escolha o critério de atribuição e tecle <ENTER>')
disp('1 - probabilidade')
disp('2 - distancia')
assign_method = input('Opcao: ');
switch assign_method
    case 1
E=('prob');
    case 2
E=('dist');
 otherwise
 disp('dist')
end
assign_method=E;

class=Group_Train;

%Particionando o espectro em I intervalos

[X]=fun_part(Train,intervals);
normas=[];
norm_max=[];
for i=1:size(X,2)
```

```
a=X{1,i};
x=Train(:,a);
for j=1:size(x,2)
b=norm(x(:,j));
normas=[normas b];
[A index_norm_max]=max(normas);
end
norm_max=[norm_max  index_norm_max];
end
iXcal=Train(:,norm_max);
  %Aplicando o SPA
  [L] = cadeias(iXcal,N1,N2);

if nsampval==0
[iNmis_cal,iNlambdas]=size(iXcal);
R = zeros(1,N2);
rmsep = [];
Lopt = zeros(N2,N2);
 temporizador = waitbar(0,'Calculo dos modelos SIMCA para as cedeias geradas pelo
SPA ...');
loopStart = now;

for N = N1:N2
% %
loopEnd = loopStart + (now-loopStart)*N2:N;
waitbar(N/N2,temporizador,['Aplicando SIMCA nas cadeias de intervalos. '
datestr(loopEnd)]);
%
for i = 1:iNlambdas
lambdas = L(1:N,i);
z=[];
for b = 1:size(lambdas,1);
g=lambdas(b,1);
z=[z,X{1,g}];
end
%
Xcal2=Train(:,z);
%
cv=simcacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);
%
min_erros=cv.class_param.er;
rmsep(N,i)=min_erros;% escolha do menor conjunto de erro.

end
[R(N) imin] = min(rmsep(N,:));
Lopt(1:N,N)=L(1:N,imin);
% %
end
close (temporizador)
[Rbest,Nbest] = min(R(N1:N2));
```

```
Nbest = Nbest+N1-1;
rmsepopt = rmsep(Nbest,:);
l = (Lopt(1:Nbest,Nbest))';
EC=rmsep;
end
h=[];
for f = 1:size(l,2);
e=l(1,f);
h=[h,X{1,e}];
Xcal2=Train(:,h);
end
model = simcafit(Xcal2,class,num_comp,pret_type,assign_method);% model ajustado
iSPA_SIMCA
cv = simcacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);

 %Previsão para um conjunto externo de amostras
Xcal2=Train(:,h);
Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Group_Train-ones(Nmis_cal,1)*mean(Group_Train);
Xpred2=Test(:,h);
Xpred_cm=Xpred2-ones(size(Xpred2,1),1)*mean(Xcal2);
pred = simcapred(Xpred2,model);
class_param = calc_class_param (pred.class_pred,Group_Test);%matrix de confusão
para conjunto de predição

[n,m]=size(Train);% determinando o número de variáveis
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];% Inicio de cada intervalo
endint=[startint(2:intervals)-1; m]; % Final de Intervalo
X=cell(1,intervals); %Iniciando X
for i=1:intervals
    x=startint(i,1):endint(i,1);
    X{i}=x;
end
[nint,mint]=size(intervals);
Modelo_iAPS_SIMCA.resultado=intervals;
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
   int=l;
   [X]=fun_part(Train,intervals);
   m=size(Train,2);
   vars_left_over=mod(m,intervals);
   N=fix(m/intervals);
   startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
```

```
  endint=[startint(2:intervals)-1; m];
  X=cell(1,intervals);
for i=1:intervals
    x=startint(i,1):endint(i,1);
    X{i}=x;
end
  in_fin_int_sel=zeros(2,length(l));
for i=1:length(l)
   int_sel=X{l(i)};
   in_fin_int_sel(1,i)=min(int_sel);
   in_fin_int_sel(2,i)=max(int_sel);
end

rawX=(Train);
xaxislabels=size(Train,2);
xaxislabels=[1:1:xaxislabels];
Modelo_iAPS_SIMCA.resultado.rawX=rawX;
Modelo_iAPS_SIMCA.resultado.allint=[(1:intervals+1)' [startint;1] [endint;m]];
Modelo_iAPS_SIMCA.resultado.intervalsequi=1;
Modelo_iAPS_SIMCA.resultado.intcom=in_fin_int_sel;
Modelo_iAPS_SIMCA.resultado.Train=model;
Modelo_iAPS_SIMCA.resultado.Xtrain_sel=Xcal2;
Modelo_iAPS_SIMCA.resultado.selected_intervals=l;
Modelo_iAPS_SIMCA.resultado.Test=pred;
Modelo_iAPS_SIMCA.resultado.xaxislabels=xaxislabels;
Modelo_iAPS_SIMCA.resultado.Fatores_sugeridos=comp;
Modelo_iAPS_SIMCA.resultado.prepro_method='mean';
Modelo_iAPS_SIMCA.resultado.Train.cv=cv;
Modelo_iAPS_SIMCA.resultado.Test.class_param=class_param;
Modelo_iAPS_SIMCA.resultado.erro_rate=EC;
% Saida gráfica
ispaplsplot(Modelo_iAPS_SIMCA.resultado,Modelo_iAPS_SIMCA.resultado.Fatores_
sugeridos);
%figure,surf(Modelo_iAPS_SIMCA.resultado.erro_rate)

else
disp('-----------------------------------------------------------------')
disp('Escolha o método dimensionamento de dados e tecle <ENTER>')
disp('1 - Centrado na média')
disp('2 - Autoescalamento')
disp('3 - Sem escala')
pret_type = input('Opcao: ');
 switch pret_type
    case 1
B=('cent');
    case 2
B=('auto');
    case 3
B=('none');
otherwise
```

```
disp('cent')
end
pret_type=B;
 disp('------------------------------------------------------------------')
disp(' ')
disp('Escolha tipo de validação cruzada e tecle <ENTER>')
disp('1 - venetian blinds')
disp('2 - contiguous blocks')
cv_type = input('Opcao: ');
switch cv_type
    case 1
C=('vene');
    case 2
C=('cont');
otherwise
disp('vene')
end
cv_type=C;
 disp('------------------------------------------------------------------')
disp(' ')
cv_group=size(Train,1);
disp(['Número total de amostras de treinamento ',num2str(cv_group)])
Numero_de_Amostra_de_Treinamento = input('Qual o número de cv-groups deseja
usar?    ');
cv_groups=Numero_de_Amostra_de_Treinamento;
if  cv_groups > cv_group
   error('cv_groups deve ser menor ou igual ao numero total de amostras de
treinamento')
end
disp('------------------------------------------------------------------')
disp(' ')
disp('Escolha o critério de atribuição e tecle <ENTER>')
disp('1 - probabilidade')
disp('2 - distancia')
assign_method = input('Opcao: ');
switch assign_method
    case 1
E=('prob');
    case 2
E=('dist');
 otherwise
 disp('dist')
end
P=assign_method;
assign_method=E;
class=Group_Train;
 %Resultados para o número de fatores ótimo!
res=simcacompsel(Train,class,pret_type,cv_type,cv_groups,assign_method);
Modelo_iAPS_SIMCA.resultado.full_SIMCA=res;
```

```
for i=1:size(Modelo_iAPS_SIMCA.resultado.full_SIMCA.er,1)
index=1:length(Modelo_iAPS_SIMCA.resultado.full_SIMCA.er);figure,plot(index,Mo
delo_iAPS_SIMCA.resultado.full_SIMCA.er(i,:))
 title(['Class' , num2str(i)])
 [min_er min_ind]=min(Modelo_iAPS_SIMCA.resultado.full_SIMCA.er(i,:));
 comp(i)=min_ind;
 num_comp=comp;
 Modelo_iAPS_SIMCA.resultado.full_SIMCA.fatores=comp;
 disp(['Número de fatores sugeridos para a classe '  num2str(i),'  .....>>>>>  '
,num2str(min_ind)])
 end
 disp('   ')
for i=1:max(Group_Train)
disp('     ')
PC(i)=input([' Número de PCs para a classe ',     num2str(i),' ?---->>>']);
  end
 comp=PC;
num_comp=PC;
[X]=fun_part(Train,intervals);
normas=[];
norm_max=[];
for i=1:size(X,2)
a=X{1,i};
x=Train(:,a);
for j=1:size(x,2)
b=norm(x(:,j));
normas=[normas b];
[A index_norm_max]=max(normas);
end
norm_max=[norm_max  index_norm_max];
end
iXcal=Train(:,norm_max);
   %Aplicando o SPA
[L] = cadeias(iXcal,N1,N2);
if nsampval==0
[iNmis_cal,iNlambdas]=size(iXcal);
R = zeros(1,N2);
rmsep = [];
Lopt = zeros(N2,N2);
temporizador = waitbar(0,'Calculo dos modelos PLS para as cedeias geradas pelo SPA
...');
loopStart = now;
for N = N1:N2
loopEnd = loopStart + (now-loopStart)*N2:N;
waitbar(N/N2,temporizador,['Aplicando SIMCA nas cadeias de intervalos. '
datestr(loopEnd)]);
for i = 1:iNlambdas
lambdas = L(1:N,i);
z=[];
for b = 1:size(lambdas,1);
```

```
g=lambdas(b,1);
z=[z,X{1,g}];
end
Xcal2=Train(:,z);
cv = simcacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);
min_erros=cv.class_param.er;
rmsep(N,i)=min_erros;% escolha do menor conjunto de erro.
end
[R(N) imin] = min(rmsep(N,:));
Lopt(1:N,N)=L(1:N,imin);
end
close (temporizador)
[Rbest,Nbest] = min(R(N1:N2));
Nbest = Nbest+N1-1;
rmsepopt = rmsep(Nbest,:);
l = (Lopt(1:Nbest,Nbest))';
ER=rmsep;
end
construção do modelo PLS para os "l" intervalos selecionados e previsão
h=[];
for f = 1:size(l,2);
e=l(1,f);
h=[h,X{1,e}];
Xcal2=Train(:,h);
end
model = simcafit(Xcal2,class,num_comp,pret_type,assign_method);% model ajustado
iSPA_SIMCA
cv = simcacv(Xcal2,class,comp,pret_type,cv_type,cv_groups,assign_method);

Previsão para um conjunto externo de amostras
Xcal2=Train(:,h);
Xcm=Xcal2-ones(Nmis_cal,1)*mean(Xcal2);
Ycm=Group_Train-ones(Nmis_cal,1)*mean(Group_Train);
Xpred2=Test(:,h);
Xpred_cm=Xpred2-ones(size(Xpred2,1),1)*mean(Xcal2);
pred = simcapred(Xpred2,model);
class_param = calc_class_param (pred.class_pred,Group_Test);%matrix de confusão
para conjunto de predição

[n,m]=size(Train);% determinando o número de variáveis
vars_left_over=mod(m,intervals);
N=fix(m/intervals);startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)';
((vars_left_over1)*(N+1)+1+1+N:N:m)'];% Inicio de cada intervalo
endint=[startint(2:intervals)-1; m]; % Final de Intervalo
X=cell(1,intervals); %Iniciando X
for i=1:intervals
x=startint(i,1):endint(i,1);
X{i}=x;
end
[nint,mint]=size(intervals);
```

```
Modelo_iAPS_SIMCA.resultado=intervals;
 vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
int=l;
[X]=fun_part(Train,intervals);
m=size(Train,2);
vars_left_over=mod(m,intervals);
N=fix(m/intervals);
startint=[(1:(N+1):(vars_left_over-1)*(N+1)+1)'; ((vars_left_over-
1)*(N+1)+1+1+N:N:m)'];
endint=[startint(2:intervals)-1; m];
X=cell(1,intervals);
for i=1:intervals
x=startint(i,1):endint(i,1);
X{i}=x;
end
in_fin_int_sel=zeros(2,length(l));
for i=1:length(l)
int_sel=X{l(i)};
in_fin_int_sel(1,i)=min(int_sel);
in_fin_int_sel(2,i)=max(int_sel);
end
rawX=(Train);
xaxislabels=size(Train,2);
xaxislabels=[1:1:xaxislabels];
Modelo_iAPS_SIMCA.resultado.rawX=rawX;
Modelo_iAPS_SIMCA.resultado.allint=[(1:intervals+1)' [startint;1] [endint;m]];
Modelo_iAPS_SIMCA.resultado.intervalsequi=1;
Modelo_iAPS_SIMCA.resultado.intcom=in_fin_int_sel;
Modelo_iAPS_SIMCA.resultado.Train=model;
Modelo_iAPS_SIMCA.resultado.Xtrain_sel=Xcal2;
Modelo_iAPS_SIMCA.resultado.selected_intervals=l;
Modelo_iAPS_SIMCA.resultado.Test=pred;
Modelo_iAPS_SIMCA.resultado.xaxislabels=xaxislabels;
Modelo_iAPS_SIMCA.resultado.Fatores_sugeridos=comp;
Modelo_iAPS_SIMCA.resultado.prepro_method='mean';
Modelo_iAPS_SIMCA.resultado.Train.cv=cv;
Modelo_iAPS_SIMCA.resultado.Test.class_param=class_param;
Modelo_iAPS_SIMCA.resultado.errro_rate=ER;
ispaplsplot(Modelo_iAPS_SIMCA.resultado,Modelo_iAPS_SIMCA.resultado.Fatores_
sugeridos);
end
```