**UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL**
**ESCOLA DE ENGENHARIA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO**

Felipe Soares

# ABORDAGENS DE SELEÇÃO DE VARIÁVEIS PARA CLASSIFICAÇÃO E REGRESSÃO EM QUÍMICA ANALÍTICA

Porto Alegre
2017

Felipe Soares

# Abordagens de seleção de variáveis para classificação e regressão em química analítica

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito parcial à obtenção do título de Mestre em Engenharia de Produção, modalidade Acadêmica, na área de concentração em Sistemas de Produção.

Orientador: Michel José Anzanello, *Ph.D.*

Porto Alegre

2017

Felipe Soares

**Abordagens de seleção de variáveis para classificação e regressão em química analítica**

Esta dissertação foi julgada adequada para a obtenção do título de Mestre em Engenharia de Produção na modalidade Acadêmica e aprovada em sua forma final pelo Orientador e pela Banca Examinadora designada pelo Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul.

_____

**Prof. Michel José Anzanello, *Ph.D.***
Orientador PPGEP/UFRGS

_____

**Prof. Flávio Sanson Fogliatto, *Ph.D*.**
Coordenador PPGEP/UFRGS

**Banca Examinadora:**

Professor Flávio Sanson Fogliatto, *Ph.D.* (PPGEP/UFRGS)

Professor Giovani da Silveira, *Ph.D.* (Haskayne School of Business/University of Calgary)

Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

*"If one were to bring ten of the wisest men in the world together and ask them what was the most stupid thing in existence, they would not be able to discover anything so stupid as astrology."*

David Hilbert

SOARES, Felipe. ***Abordagens de seleção de variáveis para classificação e regressão em química analítica***, 2017. Dissertação (Mestrado em Engenharia) – Universidade Ferderal do Rio Grande do Sul, Brasil.

## RESUMO

A utilização de técnicas analíticas para classificação de produtos ou predição de propriedades químicas tem se mostrado de especial interesse tanto na indústria quanto na academia. Através da análise da concentração elementar, ou de técnicas de espectroscopia, é possível obter-se um grande número de informações sobre as amostras em análise. Contudo, o elevado número de variáveis disponíveis (comprimentos de onda, ou elementos químicos, por exemplo) pode prejudicar a acurácia dos modelos gerados, necessitando da utilização de técnicas para seleção das variáveis mais relevantes com vistas a tornar os modelos mais robustos. Esta dissertação propõe métodos para seleção de variáveis em química analítica com propósito de classificação de produtos e predição via regressão de propriedades químicas. Para tal, inicialmente propõe-se um método de seleção de intervalos não equidistantes de comprimentos de onda em espectroscopia para classificação de combustíveis, o qual baseia-se na distância entre espectros médios de duas classes distintas; os intervalos são então utilizados em técnicas de classificação. Ao ser aplicado em dois bancos de dados de espectroscopia, o método foi capaz de reduzir o número de variáveis utilizadas para somente 23,19% e 4,95% das variáveis originais, diminuindo o erro de 13,90% para 11,63% e de 4,71% para 1,21%. Em seguida é apresentado um método para seleção dos elementos mais relevantes para classificação de vinhos provenientes de quatro países da América do Sul, baseado nos parâmetros da análise discriminante linear. O método possibilitou atingir acurácia média de 99,9% retendo em média 6,82 elementos químicos, sendo que a melhor acurácia média atingida utilizando todos os 45 elementos disponíveis foi de 91,2%. Por fim, utiliza-se o algoritmo *support vector regression – recursive feature elimination* (SVR-RFE) para seleção dos comprimentos de onda mais importantes na regressão por vetores de suporte. Ao serem aplicado em 12 bancos de dados juntamente com outros métodos de seleção e regressão, o SVR e o SVR-RFE obtiveram os melhores resultados em 8 deles, sendo que o SVR-RFE foi significativamente superior dentre os algoritmos de seleção. A aplicação dos métodos de

seleção de variáveis propostos na presente dissertação possibilitou a realização de classificações e regressões mais robustas, bem como a redução do número de variáveis retidas nos modelos.

**Palavras-chave:** Seleção de variáveis, Classificação, Regressão, Química analítica, Espectroscopia, Análise elementar.

SOARES, Felipe. ***Feature selection approaches for classification and regression in analytical chemistry***, 2017. Dissertation (Master in Engineering) - Federal University of do Rio Grande do Sul, Brazil.

## ABSTRACT

The use of analytical techniques in product classification or chemical properties estimation has been of great interest in both industry and academy. The employment of spectroscopy techniques, or through elemental analysis, provides a great amount of information about the samples being analyzed. However, the large number of features (e.g.: wavelengths or chemical elements) included in the models may jeopardize the accuracy, urging the employment of feature selection techniques to identify the most relevant features, producing more robust models. This dissertation presents feature selection methods focused on analytical chemistry, aiming at product classification and chemical property estimation (regression). For that matter, the first proposed method aims at identifying the most relevant wavelength intervals for fuel classification based on the distance between the average spectra of the two classes being analyzed. The identified intervals are then used as input for classifiers. When applied to two spectroscopy datasets, the proposed framework reduced the number of features to just 23.19% and 4.95% of the original ones, also reducing the misclassification error to 4.71% and 1.21%. Next, a method for identifying the most important elements for wine classification is presented, which is based on the parameters from linear discriminant analysis and aims at classifying wine samples produced in four south American countries. The method achieved average accuracy of 99.9% retaining average 8.82 chemical elements; the best accuracy using all 45 available chemical elements was 91.2%. Finally, the use of the support vector regression – recursive feature elimination (SVR-RFE) algorithm is proposed to identify the most relevant wavelengths for support vector regression. The proposed framework was applied to 12 datasets with other feature selection approaches and regression algorithms. SVR and SVR-RFE achieved the best results in 8 out of 12 datasets; SVR-RFE when compared to other feature selection algorithms proved have significantly better performance. The employment of the proposed feature selection methods

in this dissertation yield more robust classifiers and regression models, also reducing the number of features needed to produce accurate results.

**Keywords:** Feature selection, Classification, Regression, Analytical chemistry, Spectroscopy, Elemental analysis.

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE SIGLAS

| | |
|---|---|
| ACP | Análise de componentes principais |
| ADL | Análise discriminante linear |
| BE | *Backward elimination* |
| CDO | *Controlled denomination of origin* |
| FS | *Forward Selection* |
| FTIR | *Fourier transform infrared spectroscopy* |
| ICP-MS | *Inductively coupled plasma mass spectrometry* |
| ICP-OES | *Inductively coupled plasma optical emission spectrometry* |
| KNN | *K-nearest neighbors* |
| LDA | *Linear discriminant analysis* |
| LOOCV | *Leave one out cross validation* |
| NB | *Naïve Bayes* |
| NN | *Nearest neighbor* |
| PCA | *Principal component analysis* |
| PLS | *Partial least squares* |
| PNN | *Probabilistic neural networks* |
| RMSEP | *Root mean square error of prediction* |
| SPA | *Successive projection algorithm* |
| SVM | *Support vector machine* |
| SVR | *Support vector regression* |

**SUMÁRIO**

# 1. Introdução

## 1.1 Considerações Iniciais

A utilização de técnicas multivariadas e de inteligência artificial em análises químicas tem se mostrado de extremo interesse e aplicabilidade, especialmente na indústria (FERREIRA; TOBYN, 2015; LIU; SUN; ZENG, 2014). A aplicação de ferramental analítico para controle de qualidade é imperativo em diversos processos produtivos, seja para garantir a segurança de consumo do produto final, ou para garantir a adequação do produto ou processo à legislação vigente (KELLY, 2003). A combinação de ferramentas multivariadas e de inteligência artificial com tradicionais métodos analíticos torna possível a obtenção de modelos mais precisos para classificação de produtos ou predição de propriedades químicas, facilitando a tomada de decisão em diversos setores e aplicações (MUÑOZ-OLIVAS, 2004).

Além de sua utilização em ambientes industriais, as técnicas analíticas são também de especial interesse aos órgãos de controle ambiental ou sanitário, como a rAgência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). A ANP é o órgão brasileiro responsável pelo controle da produção e comercialização de combustíveis, sendo sua responsabilidade fiscalizar e garantir a qualidade e o atendimento das especificações dos combustíveis presentes no mercado, de modo a evitar prejuízos ambientais e riscos à saúde da população (AUGUSTO HORTA NOGUEIRA; SILVA CAPAZ, 2013). De modo similar, a indústria de bebidas tem potencial interesse em certificar a procedência e autenticidade de seus produtos, de modo a obter vantagem competitiva e garantir a proteção de suas marcas (DINIZ et al., 2014; KAROUI; DE BAERDEMAEKER, 2007).

A construção de modelos preditivos muitas vezes pode ser prejudicada pelo grande número de variáveis sendo analisadas, podendo reduzir sua acurácia ou produzindo resultados não confiáveis (XIAOBO et al., 2010). A redução da dimensionalidade dos dados pode ser obtida através da combinação das variáveis originais (extração), ou pela identificação de um subconjunto de variáveis originais que sejam mais significativas para o problema analisado (BISHOP, 2006; WEBB; COPSEY, 2011). A seleção de um reduzido número de variáveis informativas pode proporcionar a obtenção de modelos mais robustos, precisos e

interpretáveis, além de possibilitar a criação de aparatos mais compactos (ANDERSEN; BRO, 2010).

A presente dissertação é composta por três artigos abordando a seleção de variáveis com o propósito de classificar produtos e predizer informações químicas utilizando técnicas analíticas. No primeiro artigo é proposta uma sistemática para seleção de comprimentos de onda em espectroscopia para classificação de combustíveis. O método é baseado na distância entre os espectros médios de cada classe, identificando os intervalos mais relevantes, os quais são apresentados a três classificadores: *Linear Discriminant Analysis* (LDA), *k-Nearest Neighbors* (KNN) e *Probabilistic Neural Networks* (PNN). O segundo artigo aborda a identificação dos elementos químicos mais relevantes para a classificação de amostras de vinho de acordo com a região geográfica de produção. Para tanto é proposta uma sistemática de seleção em duas etapas; o teste Kruskal-Wallis é utilizado para preliminarmente eliminar as variáveis menos informativas, seguido da utilização dos parâmetros da LDA para ordenar as variáveis remanescentes; na segunda etapa as variáveis são iterativamente testadas em quatro classificadores. O terceiro artigo propõe a utilização da regressão por vetores de suporte (SVR) para predição de informações químicas utilizando espectroscopia no infravermelho próximo (NIR), bem como a identificação dos comprimentos de onda mais relevantes. O algoritmo de seleção de variáveis SVM-RFE, originalmente utilizado para classificação, é adaptado para o caso de regressão, produzindo um índice ordenado de variáveis para posterior seleção.

## 1.2 Objetivos

O objetivo principal da dissertação é propor sistemáticas de seleção de variáveis com vistas à classificação de produtos e predição de propriedades químicas.

Os seguintes objetivos específicos são apresentados:

- Selecionar intervalos de comprimentos de onda não equidistantes provenientes de espectroscopia para classificar amostras de diesel e misturas de diesel/biodiesel;

- Aplicar o teste de Kruskal-Wallis como ferramenta de seleção preliminar, possibilitando o emprego de métodos mais sofisticados de seleção em etapa subsequente;
- Criar um Índice de Importância de Variáveis baseado na LDA que possibilite a posterior inserção ordenada em algoritmos de classificação;
- Empregar a regressão por vetores de suporte como ferramenta para calibração multivariada de dados espectroscópicos;
- Adaptar o algoritmo SVM-RFE para seleção de comprimentos de onda em modelos de regressão.

## 1.3 Justificativa do Tema e dos Objetivos

O constante avanço e aperfeiçoamento de técnicas analíticas possibilitam extrair um grande número de informações relacionadas a propriedades químicas de forma rápida e precisa. Porém, o grande volume de informações gerado por tais técnicas pode conter porções de dados não relevantes e ruidosos para o problema em questão, podendo diminuir a habilidade preditiva ou exploratória dos modelos gerados (XIAOBO et al., 2010). Por conseguinte, o pré-tratamento dos dados obtidos constitui-se em importante etapa para a obtenção de modelos mais precisos (MEHMOOD et al., 2012). De tal forma, as sistemáticas propostas no presente trabalho possuem respaldo prático na potencial redução do número de variáveis analisadas pelos algoritmos multivariados na predição de propriedades químicas e classificação de produtos.

Na literatura acadêmica observa-se o constante interesse pelo desenvolvimento de métodos para identificação das variáveis mais relevantes para construção de modelos preditivos confiáveis e parcimoniosos. Portanto, a realização desta pesquisa justifica-se em função do estudo da combinação de técnicas multivariadas e sistemáticas de identificação das variáveis mais relevantes para construção de modelos robustos com vistas à classificação e predição de propriedades químicas com base em dados oriundos de técnicas analíticas.

## 1.4 Procedimentos Metodológicos

Em relação aos objetivos, a presente dissertação é classificada como pesquisa exploratória, uma vez que permite conhecer o problema e possibilita a construção de hipóteses para sua solução. Quanto à natureza, é tida como pesquisa aplicada, dado que a fundamentação teórica é explorada e utilizada na solução de problemas genéricos (CRESWELL, 2010). A dissertação apresenta abordagem quantitativa, pois utiliza análises estatísticas e modelagem matemática para solução dos problemas apresentados.

No primeiro artigo o método proposto para seleção de intervalos de comprimentos de onda para classificação de amostras de diesel/biodiesel é dividido em 4 etapas. Inicialmente são computados os espectros médios para cada uma das duas classes, sendo em seguida computada a diferença absoluta entre as médias, obtendo-se um perfil de distância entre os espectros médios. No terceiro passo são identificados os picos no perfil de distância, com seus mínimos locais à esquerda e direita definindo os intervalos de comprimentos de onda. Após a identificação, os intervalos são apresentados a três ferramentas de classificação com três abordagens diferentes de seleção: seleção *forwards*, eliminação *backwards*, e todas as combinações de 1, 2, 3 e 4 intervalos. A melhor combinação de intervalos e ferramenta de classificação é identificada para cada banco de dados baseada na taxa de erro na classificação.

Com o objetivo de classificar amostras de vinho de acordo com o país de produção, o segundo artigo apresenta um método para identificação dos elementos químicos mais informativos. O método é dividido em duas fases principais: a filtragem e ordenação das variáveis, seguido por um processo iterativo para identificação do melhor subconjunto de acordo com diferentes classificadores. Na primeira fase é utilizado o teste Kruskal-Wallis como primeira etapa de seleção, seguido pelo emprego da LDA para derivação de um índice de importância para ordenação das variáveis de acordo com sua relevância na classificação. Na segunda fase as variáveis são apresentadas ordenadamente a quatro classificadores, sendo identificado o melhor subconjunto de variáveis para cada classificador.

O terceiro artigo tem por objetivo empregar a regressão por vetores de suporte, bem como o emprego de um reduzido subconjunto de comprimentos de onda para calibração multivariada de dados espectroscópicos. O algoritmo de eliminação recursiva baseado em máquinas de vetores de suporte (SVM) é adaptado para o caso de regressão, utilizando os parâmetros da solução SVR como índice de importância. A cada iteração o comprimento de onda menos relevante é eliminado, até que reste somente um comprimento de onda. Após a

ordenação de todos os comprimentos de onda, o subconjunto com melhor desempenho é selecionado.

## 1.5 Estrutura da Dissertação

A dissertação encontra-se dividida em 5 capítulos. O primeiro capítulo introduz o trabalho, apresentando os objetivos e as justificativas, bem como o método de pesquisa adotado. A estrutura do trabalho e a delimitação do estudo finalizam o capítulo.

No segundo capítulo é apresentado o primeiro artigo, que propõe a identificação de intervalos de comprimentos de onda baseando-se na distância entre os espectros médios das classes sendo analisadas. Os intervalos identificados são apresentados a três tradicionais classificadores (LDA, KNN e PNN) em três diferentes abordagens de seleção: para frente (*forwards*), eliminação para trás (*backwards*) e todas as combinações de 1, 2, 3 e 4 intervalos. O método utilizado pretende classificar amostras de diesel em *premium* ou *standard* e amostras de mistura biodiesel/diesel em rural ou metropolitano.

O terceiro capítulo da dissertação apresenta o segundo artigo, o qual visa classificar amostras de vinho de acordo com o país de produção baseando-se na concentração de diferentes elementos químicos. É proposto um método para seleção dos elementos utilizando o teste Kruskal-Wallis e a LDA para seleção e ordenamento dos elementos. Quatro classificadores, *Naïve Bayes* (NB), LDA, SVM e vizinho mais próximo (NN), são utilizados para obter subconjuntos de variáveis através da abordagem de seleção *forwards* (FS). O melhor subconjunto de elementos químicos identificado pelo classificador mais preciso é analisado qualitativamente através de gráficos *box plot*.

O quarto capítulo traz o terceiro artigo, que apresenta a adaptação do algoritmo SVM-RFE para o caso de regressão e a utilização do SVR para calibração multivariada de dados espectroscópicos. O algoritmo de seleção proposto, SVR-RFE, utiliza os pesos da regressão por vetores de suporte como critério de importância para os comprimentos de onda. A cada iteração do algoritmo, um comprimento de onda é eliminado e o desempenho de predição é computado; tal procedimento é repetido até que todos os comprimentos de onda tenham sido eliminados. O subconjunto de comprimentos de onda com menor erro de calibração é selecionado. O método proposto é comparado com métodos tradicionais de seleção em 12

bancos de dados, bem como a utilização do espectro completo para regressão SVR e por mínimos quadrados parciais (PLS).

O quinto e último capítulo traz a conclusão do trabalho, na qual são avaliados os principais resultados frente aos objetivos almejados e as delimitações citadas. Essa seção traz ainda sugestões para desdobramentos futuros.

## 1.6 Delimitações do Estudo

Constituem-se em restrições do presente estudo:

- O trabalho não irá apresentar novas ferramentas de classificação ou regressão, restringindo-se a combinar tais ferramentas de forma a gerar novas abordagens para seleção de variáveis;
- As variáveis são selecionadas somente através de métodos supervisionados, sendo necessária prévia informação sobre a classe de cada amostra, ou variável dependente a ser modelada;
- A influência de métodos de pré-tratamento no desempenho dos algoritmos de classificação e regressão não são testadas; e
- Os bancos de dados estudados restringem-se a espectroscopia e concentração de elementos químicos obtidas por ICP-MS e ICP-OES.

## 1.7 Referências

ANDERSEN, C. M.; BRO, R. Variable selection in regression-a tutorial. **Journal of Chemometrics**, v. 24, n. 11-12, p. 728–737, 2010.

AUGUSTO HORTA NOGUEIRA, L.; SILVA CAPAZ, R. Biofuels in Brazil: Evolution, achievements and perspectives on food security. **Global Food Security**, v. 2, n. 2, p. 117–125, 2013.

BISHOP, C. M. **Pattern Recognition and Machine Learning** . [s.l.] Springer, 2006.

CRESWELL, J. W. Projeto de pesquisa métodos qualitativo, quantitativo e misto. In: **Projeto de pesquisa métodos qualitativo, quantitativo e misto**. [s.l.] Artmed, 2010.

DINIZ, P. H. G. D. et al. Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA. **Food Analytical Methods**, v. 7, n. 8, p. 1712–1718, 2014.

FERREIRA, A. P.; TOBYN, M. Multivariate analysis in the pharmaceutical industry: enabling process understanding and improvement in the PAT and QbD era. **Pharmaceutical Development and Technology**, v. 20, n. 5, p. 513–527, 4 jul. 2015.

KAROUI, R.; DE BAERDEMAEKER, J. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. **Food Chemistry**, v. 102, n. 3, p. 621–640, 2007.

KELLY, S. D. **Using stable isotope ratio mass spectrometry (IRMS) in food authentication and traceability**. [s.l: s.n.].

LIU, D.; SUN, D.; ZENG, X. Recent Advances in Wavelength Selection Techniques for Hyperspectral Image Processing in the Food Industry. **Food Bioprocess Technology**, v. 7, p. 307–323, 2014.

MEHMOOD, T. et al. A review of variable selection methods in Partial Least Squares Regression. **Chemometrics and Intelligent Laboratory Systems**, v. 118, p. 62–69, ago. 2012.

MUÑOZ-OLIVAS, R. Screening analysis: An overview of methods applied to environmental, clinical and food analyses. **TrAC - Trends in Analytical Chemistry**, v. 23, n. 3, p. 203–216, 2004.

WEBB, A. R.; COPSEY, K. D. **Statistical Pattern Recognition**. 3. ed. [s.l.] Wiley, 2011.

XIAOBO, Z. et al. Variables selection methods in near-infrared spectroscopy. **Analytica chimica acta**, v. 667, n. 1-2, p. 14–32, 14 maio 2010.

## 2. Primeiro artigo: Seleção de intervalos de comprimentos de onda não equidistantes para classificação de amostras de diesel/biodiesel

**Felipe Soares**

**Michel José Anzanello**

**Marcelo Caetano Alexandre Marcelo**

**Marco Flôres Ferrão**

### Resumo

Técnicas de espectroscopia apoiadas em infravermelho por transformada de Fourier (FTIR) no infravermelho médio (MIR) ou infravermelho próximo (NIR) têm sido largamente adotadas como ferramentas analíticas em diferentes áreas e com propósitos distintos de análise. Dados de MIR e NIR usualmente apoiam-se em centenas ou milhares de comprimentos de onda (representando variáveis) altamente correlacionados, o que pode comprometer a acurácia de várias técnicas estatísticas. Em vista disso, a seleção de comprimentos de onda é tida como uma importante etapa na construção de modelos de classificação baseados em dados espectroscópicos. O presente artigo propõe um novo método para seleção de comprimentos de onda para classificar amostras em classes binárias, o qual é empregado em dois bancos de dados do setor de petróleo. O método consiste de duas etapas principais: determinação dos intervalos através da distância entre os espectros médios de cada classe e seleção dos intervalos mais apropriados através da validação cruzada. O método proposto reduziu a taxa de classificações incorretas para o banco de dados NIR de diesel de 13,90% para 11,63%, retendo 23,19% dos comprimentos de onda originais. Para o banco de dados MIR de biodiesel/diesel, o método alcançou uma taxa de erro de classificação de 1,21%, retendo 4,95% dos comprimentos de onda originais; um erro de 4,71% é gerado ao utilizar-se todo o espectro (1738 comprimentos de ondas). O método proposto também foi comparado com técnicas tradicionais para seleção de intervalos, tendo obtido desempenho superior.

**Palavras-chave:** FTIR, MIR, NIR, classificação de combustíveis, seleção de comprimentos de onda, biodiesel/diesel

*Submetido ao periódico Chemometrics and Intelligent Laboratory Systems (Qualis A1)*

## 2.1 Introdução

A Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) é o órgão brasileiro responsável pela regulação da produção e comercialização de combustíveis no Brasil. Os padrões da ANP para biodiesel incluem a concentração de enxofre, a qual varia dependendo do local de comercialização: metropolitano ou rural. Considerando que o dióxido de enxofre pode ocasionar diversos problemas respiratórios [1], é importante assegurar que o biodiesel rural não seja disponibilizado em áreas metropolitanas, pois apresenta maior concentração de enxofre. Outro problema de interesse refere-se à correta classificação de amostras de combustíveis em classes pré-determinadas, as quais podem estar relacionadas a aspectos de qualidade (combustível *premium* ou *standard*, por exemplo) ou autenticidade (como amostras no padrão correto ou com alterações na concentração de enxofre).

Técnicas espectroscópicas vêm sendo largamente utilizadas em diversos segmentos industriais, especialmente devido à sua versatilidade e custo-benefício [2–4]. Além disso, sua facilidade de operação, usualmente não requerendo pré-tratamento, é um dos principais diferenciais [5]. A análise de dados provenientes de espectroscopia através de técnicas multivariadas tem resultado em importantes métodos para predição de propriedades químicas ou classificação de amostras. Pimentel *et al*. [6] apresentaram um modelo de calibração para determinar o conteúdo de biodiesel em misturas de diesel utilizando espectroscopia no infravermelho médio (MIR) e próximo (NIR). Os autores utilizaram regressões baseadas em mínimos quadrados parciais (PLS) para predizer o conteúdo de biodiesel com o objetivo de monitorar a qualidade do combustível. Similarmente, Ferrão *et al*. [7] e Anzanello *et al*. [8] aplicaram o PLS para predizer diferentes parâmetros de qualidade em misturas de biodiesel, tais como ponto de fulgor, densidade relativa e conteúdo de enxofre.

Embora a ANP recomende o uso do padrão ASTM D5453 para mensuração do conteúdo total de enxofre [9], tal método não declara diretamente se a amostra pertence à classe de biodiesel metropolitano ou não. Segundo Pontes *et al*. [10], métodos laboratoriais convencionais que fornecem informações detalhadas sobre as amostras têm sido sistematicamente substituídos por métodos analíticos que fornecem respostas binárias do tipo sim/não dado um limiar de concentração pré-determinado. Tais métodos tipicamente

apresentam melhor relação custo/benefício, são mais simples e podem ser empregados diretamente em campo, fornecendo uma resposta imediata para tomada de decisão [11]. Alinhadas com tal propósito, abordagens integrando espectroscopia e técnicas multivariadas para classificação de produtos (como gasolina e misturas de diesel/biodiesel) em categorias de interesse [10,12–14] têm sido propostas. Kim *et al*. [14] construíram um classificador em tempo real (RTC) baseado na análise de componentes principais (PCA) e um classificador bayesiano para categorizar espectros NIR associados a seis diferentes produtos provenientes do petróleo. Li e Dai [12] e Balabin *et al*. [13] aplicaram ferramentas de classificação para dados de espectroscopia Raman e NIR para classificar gasolina por marca e origem. Já Pontes *et al*. [10] utilizaram análise discriminante baseada em mínimos quadrados parciais (PLS-DA) e análise discriminante linear (LDA) para construir classificadores, utilizando espectroscopia NIR, que detectam adulteração em diesel/biodiesel. Com objetivos similares, Silva *et al*. [3] propuseram um método analítico para detectar adulterações em álcool etílico hidratado utilizando espectros MIR e NIR em conjunto com a LDA. Os modelos construídos foram baseados em um subconjunto reduzido de comprimentos de onda selecionados por três diferentes técnicas.

Um dos principais empecilhos na análise multivariada de dados espectroscópicos é a sua alta dimensionalidade, isto é, o elevado número de comprimentos de onda em relação ao número de amostras. A alta dimensionalidade usualmente reduz a capacidade exploratória ou preditiva dos modelos construídos, podendo inclusive impossibilitar a utilização de determinados algoritmos [5,15]. A seleção de um subconjunto de comprimentos de onda é uma das formas de mitigar o problema da dimensionalidade, sendo fundamental obter um reduzido número de comprimentos de onda sem que haja perda substancial da informação de interesse presente nos dados originais [16,17]. Abordagens especificamente desenvolvidas para seleção de comprimentos de onda são usualmente aplicadas no pré-processamento de dados espectroscópicos, podendo aumentar a qualidade e a velocidade das análises subsequentes [5,18].

Diversos métodos para identificação das bandas espectrais mais relevantes foram propostos na área do aprendizado supervisionado, especialmente em análises químicas [5,19,20]. Xiaobo *et al*. [5] efetuaram uma detalhada revisão sobre os principais métodos de seleção em NIR, com especial atenção aos de seleção de intervalos espectrais. Ao encontro

com os objetivos do presente artigo, o método *interval PLS (iPLS)* para seleção de intervalos espectrais foi proposto por Norgaard *et al*. [20], e posteriormente adaptado por Dyrby *et al*. [21] para otimizar a qualidade de predição e interpretação dos modelos. O iPLS é uma abordagem graficamente orientada que divide o espectro em diversos intervalos equidistantes, os quais são individualmente modelados por regressões locais. Os autores Ferrão *et al*. [7] compararam o desempenho do iPLS e do *synergy interval PLS (siPLS)* com a utilização do espectro completo para predição de parâmetros de qualidade de misturas de biodiesel/diesel. O método siPLS é implementado de forma similar ao iPLS, porém todas as combinações de dois, três ou quatro intervalos são testadas, sendo escolhida a combinação que produz o menor erro. Embora diferentes *frameworks* para seleção de comprimentos de onda (e intervalos) tenham sido propostos, ainda há oportunidade para abordagens mais simples e eficientes.

O presente artigo apresenta um novo método de seleção das regiões espectrais mais relevantes para classificação de amostras de diesel em classes binárias. O método consiste em inicialmente dividir o espectro em intervalos, os quais são baseados na distância entre o espectro médio de cada uma das classes. Maior capacidade discriminativa é esperada em regiões onde a distância entre os espectros médios é elevada. O método proposto diferencia-se de abordagens já existentes, pois os intervalos obtidos não são equidistantes (como usualmente empregado na seleção de intervalos em *iPLS*), o que possibilita focar a análise em regiões de diferentes tamanhos responsáveis pela maior discriminação entre classes. Os intervalos selecionados são então utilizados em diferentes ferramentas de classificação, sendo o erro de classificação calculado em cada cenário. Quando aplicado em dois bancos de dados de diesel e mistura diesel/biodiesel, o método proposto reduziu substancialmente o percentual de comprimentos de onda necessários para classificação, reduzindo também o percentual de erro. Por fim o desempenho do método é comparado com abordagens tradicionais para seleção de comprimentos de onda presentes na literatura.

As principais contribuições do artigo estão relacionadas à nova metodologia para definição de intervalos não equidistantes para classificação. Dado que métodos tradicionais de seleção por intervalos podem dividir picos do espectro em duas regiões distintas, fazendo com que sejam necessários dois intervalos para que a classificação obtenha melhor resultado.

## 2.2 Materiais e métodos

Nessa seção são apresentados os fundamentos das técnicas multivariadas utilizadas no presente estudo, os dois bancos de dados testados e os passos para operacionalização do método proposto.

### 2.2.1 Análise de componentes principais

A análise de componentes principais (PCA) é uma ferramenta largamente utilizada para redução da dimensionalidade de bancos de dados e visualização. Na PCA as variáveis originais que descrevem os dados são projetadas em um novo sistema de coordenadas ortogonais que melhor representam os dados, com o objetivo de maximizar a variância em cada componente. A PCA é operacionalizada através do cálculo dos autovalores e autovetores da matriz de covariância dos dados, procurando maximizar a variância explicada em cada componente e minimizar a redundância. Informações detalhadas sobre PCA podem ser encontradas em [17,22–24].

### 2.2.2 Análise discriminante linear

LDA é um método multivariado para classificação de amostras, introduzido por Fisher em 1936 [25]. A LDA pode ser interpretada como uma redução de dimensionalidade que projeta os dados originais em um novo sistema de coordenadas de modo a maximizar a separabilidade entre classes, sendo a projeção então utilizada para classificar novas amostras. Maiores detalhes sobre a LDA estão disponíveis em [26–28].

### 2.2.3 *K*-vizinhos mais próximos

O método dos *k*-vizinhos mais próximos (KNN) é largamente utilizado como ferramenta de classificação, sendo proposto por Fix e Hodges [29] e adaptado por diferentes autores [30–34]. Para classificar uma nova amostra em uma das *d* possíveis classes, o KNN computa a distância da nova amostra para as amostras presentes na partição de treino utilizando uma métrica de distância pré-definida, como Euclidiana ou Mahalanobis, por exemplo. A nova amostra é inserida na classe predominante entre os *k*-vizinhos mais

próximos [22]. O número *k* pode ser determinado através de validação cruzada na partição de treino, ou aproximado pela raiz quadrada do número de amostras de treino [27,28].

### 2.2.4 Redes neurais probabilísticas

As redes neurais têm sua origem na tentativa de replicar matematicamente o processamento de informações em sistemas biológicos, imitando a atividade cerebral. Uma rede neural usualmente consiste de uma camada de entrada, camadas intermediárias, ou ocultas, e uma camada de saída. As amostras são apresentadas para as unidades na camada de entrada, sendo linearmente combinadas nas unidades ocultas e então transformadas por uma função de ativação, usualmente não-linear (logística ou tangente hiperbólica) [22]. Os valores transformados nas unidades ocultas são novamente combinados linearmente e transformados na camada de saída, levando ao resultado final. A escolha da função de ativação depende da natureza da saída sendo modelada [22,23]. Uma partição de treino é utilizada para determinar os pesos utilizados nas combinações lineares.

Uma adaptação das redes neurais para classificação, a rede neural probabilística, ou PNN, foi desenvolvida por Specht [35]. Na PNN a função de ativação é substituída por uma função estatística, aproximando assintoticamente a superfície ótima de decisão de Bayes. A PNN tem a vantagem de ser facilmente retreinada quando novos dados são disponibilizados, embora o procedimento de treino possa ser computacionalmente caro [36]. Detalhes sobre as redes neurais probabilísticas estão presentes em [37–39].

### 2.2.5 Banco de dados de mistura biodiesel/diesel

Um total de 85 amostras de misturas de biodiesel/diesel brasileiro foram disponibilizadas por Ferrão *et al* [7]. As amostras foram preparadas com biodiesel constituído de éster metílico de óleo de soja e dois tipos de diesel: metropolitano e rural. As concentrações de biodiesel variaram entre 0,2% (v/v) e 30,0% (v/v).

Espectros das amostras foram obtidos através de espectroscopia de infravermelho por transformada de Fourier (FTIR) por Ferrão *et al*. [7] em temperatura ambiente, com espectro compreendido entre 4000 a 650 $cm^{-1}$ e resolução de 4 $cm^{-1}$. O banco de dados resultante

consiste em 1738 comprimentos de onda e 85 amostras, as quais se referem a duas classes diferentes de diesel: metropolitano (56 amostras) e rural (29 amostras).

### 2.2.6 Banco de dados de diesel

O *Southwest Research Institute*, patrocinado pelo exército americano, coletou espectros NIR de amostras de diesel, além de outras propriedades do combustível. Uma delas é o número de cetano, análogo ao índice de octano da gasolina, o qual está relacionado à velocidade de ignição, sendo uma medida da qualidade do combustível. O banco de dados completo está disponível na página da *web* do *Eigenvector Research Institute (*http://www.eigenvector.com/)

No Brasil, a Petrobras usa o número de cetano como métrica para classificar o diesel em comum ou *premium* [40]. O diesel comum apresenta um número de cetano mínimo de 42, enquanto o diesel *premium* deve ter no mínimo 51. Com base nessa informação, as amostras foram divididas nas mesmas classes utilizadas pela empresa. O banco de dados resultante contém 401 comprimentos de onda e 222 amostras, das quais 172 são do tipo comum e 50 do tipo *premium*.

### 2.2.7 Método proposto

O método proposto no presente estudo assemelha-se ao iPLS, proposto por Norgaard *et al*. [20], porém com vistas à classificação e com diferente abordagem para definição dos intervalos. Enquanto o iPLS divide o espectro em diversos intervalos equidistantes, o método proposto baseia-se em encontrar picos que sejam relacionados a uma maior diferença entre as duas classes. As regiões ao redor dos picos também são incluídas na modelagem, pois podem conter informações de interesse para a classificação. Tal procedimento possibilita selecionar regiões do espectro de diferentes tamanhos que sejam responsáveis por uma maior separabilidade.

Inicialmente considera-se $\mathbf{X}$ a matriz com o espectro das $N$ amostras de combustível pertencentes a duas classes distintas. A matriz $\mathbf{X}$ é dividida em duas matrizes, $\mathbf{X_{C1}}$ e $\mathbf{X_{C2}}$, sendo a primeira composta pelas amostras pertencentes à classe 1 (diesel rural, por exemplo), e a segunda pelas amostras da classe 2 (diesel metropolitano, por exemplo). O método

proposto para seleção dos comprimentos de onda é operacionalizado de modo a encontrar as regiões espectrais onde a diferença entre o espectro médio de cada classe esteja acima de um determinado limiar. O espectro médio é computado através da média aritmética de cada classe, ou seja, a média das colunas das matrizes $\mathbf{X_{C1}}$ e $\mathbf{X_{C2}}$, as quais são armazenadas nos vetores $\overline{\boldsymbol{x_1}}$ e $\overline{\boldsymbol{x_2}}$. O valor absoluto da diferença entre as médias é então calculado, conforme a Eq (1).

$$D = |\overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}| \quad (1)$$

O próximo passo consiste em encontrar os *p* picos em *D* que estejam acima de um limiar pré-definido, como a média ou mediana de *D*, por exemplo. Tal etapa tem como objetivo identificar os pontos associados às maiores diferenças entre os espectros das duas classes. Dado que os *p* picos tenham sido encontrados, os intervalos $I_p$ ao redor dos picos são definidos pelos mínimos locais à direita e à esquerda de cada ponto (como exemplificado nos intervalos $I_1$ e $I_2$ da Figura 2.1), ou até o limite de outro intervalo, não devendo haver sobreposição. Essa etapa deve retornar *p* intervalos de comprimentos de onda, como exemplificado na Figura 2.1. Uma visão esquemática do método de determinação dos intervalos é mostrada na Figura 2.2.

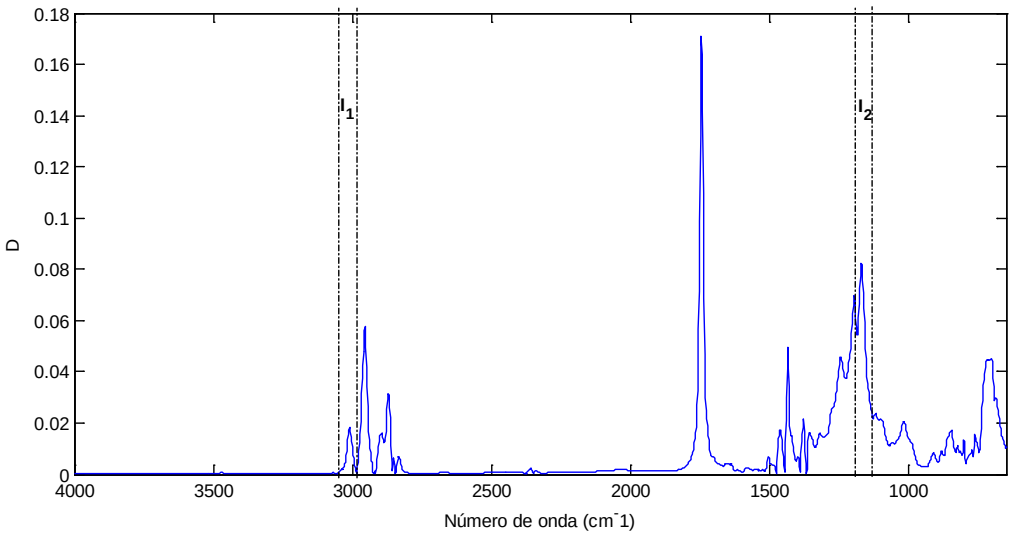


**Figura 2.1 Exemplo de dois picos em D e seus intervalos correspondentes ($I_1$ e $I_2$)**

Uma vez definidos os *p* intervalos, um processo de seleção é aplicado para definir os intervalos mais relevantes para determinada ferramenta de classificação. Três diferentes abordagens de seleção são utilizadas no presente estudo: seleção *forwards* (FS), eliminação

*backwards* (BE) e todas as combinações possíveis de 1, 2, 3 e 4 intervalos (COMB). O desempenho é avaliado através da taxa de erro de classificação em uma validação cruzada de 10 porções.

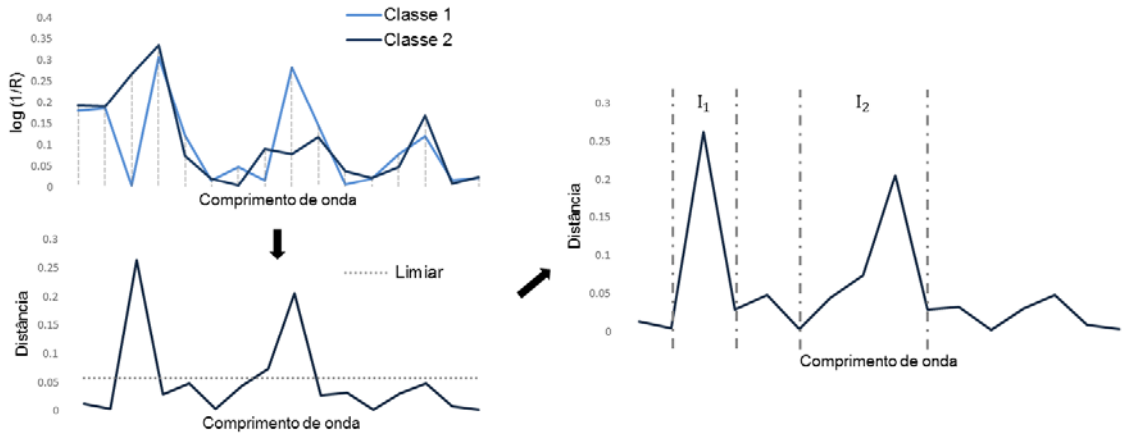


**Figura 2.2 Visão esquemática do método proposto para seleção de intervalos**

Na primeira abordagem de seleção, FS [27], a taxa de erro de classificação (MR) é avaliada para cada intervalo; o intervalo com a menor MR é inserido no subconjunto que fará parte do modelo. Na iteração seguinte, os intervalos restantes são individualmente combinados com os já selecionados; o que apresentar o menor erro é adicionado ao subconjunto dos selecionados. O algoritmo continua até que não seja observada melhoria na MR ou todos os intervalos tenham sido selecionados. A eliminação *backwards* é similar ao FS [27], porém inicia com todos os intervalos no subconjunto dos selecionados. Em cada iteração um intervalo é omitido, sendo eliminado o que apresentar maior decrescimento da MR. O algoritmo é interrompido quando não é observada melhoria no desempenho ou reste somente um intervalo. Na última abordagem (COMB), a qual é baseada no algoritmo siPLS [20], todas as combinações de 1, 2, 3 ou 4 intervalos são testadas, sendo a combinação com a menor MR a escolhida.

As abordagens de seleção descritas anteriormente são integradas a três diferentes algoritmos de classificação: LDA, KNN e PNN. Dado que o desempenho do classificador LDA é severamente prejudicado quando o número de amostras é inferior ao número de variáveis, aplicou-se PCA nos dados como pré-processamento em todos os classificadores. Todas as combinações são estudadas, sendo escolhida a que proporcionar o menor erro de classificação.

## 2.3 Resultados e discussão

As análises foram realizadas utilizando o software Matlab R2012b, em um computador com processador AMD Quad-Core A8-4500M e 8 GB de memória RAM.

### 2.3.1 Definição de parâmetros

Para melhor comparar as abordagens de seleção e ferramentas de classificação, os parâmetros requeridos por cada técnica foram otimizados independentemente. Similarmente ao procedimento utilizado por Balabin e Smirnov [15], o erro de classificação em todas as combinações de abordagem de seleção e ferramentas foi minimizado baseando-se na validação cruzada de 10 porções. Através da otimização individual o viés da configuração de cada ferramenta é diminuído, melhorando a comparabilidade, embora o custo computacional seja aumentado [15].

Os parâmetros otimizados para o KNN foram o número de componentes principais (PC) e o número *k* de vizinhos utilizados; para a PNN foi otimizado o parâmetro de *spread* (referente a função de ativação), além do número de componentes principais. Para a LDA o único parâmetro variável foi o número de componentes, o qual variou entre 1 e 7, devido ao risco da não convergência pela alta dimensionalidade. A Tabela 2.1 sumariza os parâmetros utilizados em cada um dos bancos de dados nas respectivas combinações.

**Tabela 2.1 Parâmetros otimizados para cada abordagem de seleção e classificador**

| Abordagem de seleção | | | Biodiesel/Diesel | Diesel |
|---|---|---|---|---|
| | | Parâmetro | Valor | Valor |
| BE | KNN | # PC | 12 | 6 |
| | | k | 3 | 20 |
| | LDA | # PC | 6 | 6 |
| | PNN | # PC | 6 | 3 |
| | | *Spread* | 0,02833 | 0,00666 |
| FS | KNN | # PC | 3 | 6 |
| | | k | 3 | 20 |
| | LDA | # PC | 6 | 6 |
| | PNN | # PC | 7 | 3 |
| | | *Spread* | 0,00888 | 0,00666 |
| COMB | KNN | # PC | 3 | 6 |
| | | k | 1 | 20 |
| | LDA | # PC | 6 | 6 |
| | PNN | # PC | 7 | 8 |
| | | *Spread* | 0,00888 | 0,00666 |

### 2.3.2 Resultados numéricos para o banco de dados de mistura biodiesel/diesel

O método proposto foi aplicado no banco de dados de biodiesel/diesel com o objetivo de classificar as amostras como diesel metropolitano ou rural. Para tanto, os espectros médios de cada classe foram computados e utilizados para estimar o perfil de distância *D*, conforme mostrado na Figura 2.3, sendo o limiar para identificação dos picos definido como o dobro da média aritmética de *D*. Regiões espectrais com picos satisfazendo o limiar estabelecido foram utilizadas nas diferentes abordagens de seleção e técnicas de classificação, tendo sido coletada a MR em cada combinação.

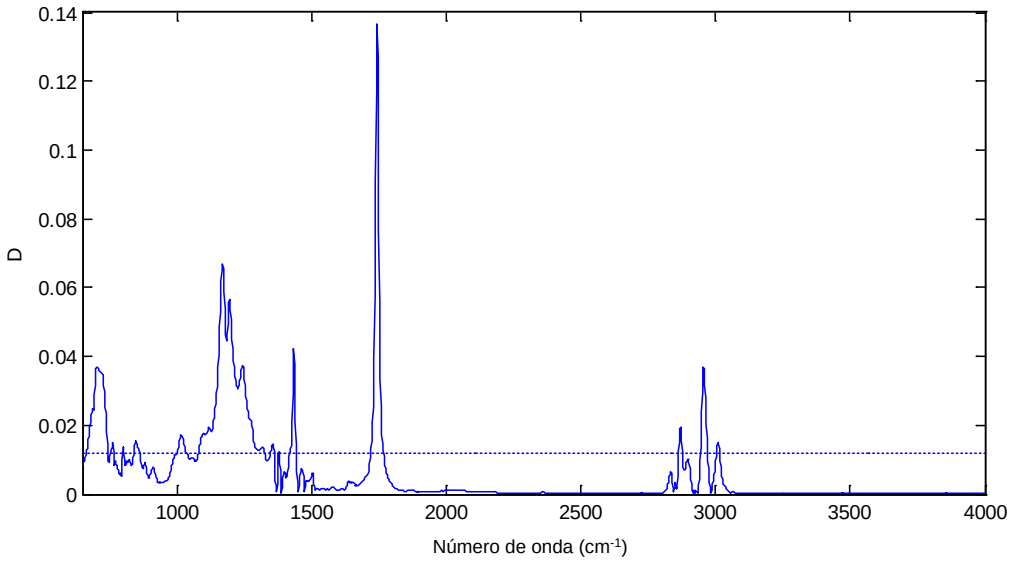


**Figura 2.3 Valores de D (linha sólida), e limiar (linha pontilhada), para o banco de dados de biodiesel/diesel**

O método reduziu substancialmente a taxa de erro na classificação, como demonstrado na Tabela 2.2. A menor MR média ao utilizar-se todos os comprimentos de onda no modelo de classificação foi de 4,62% utilizando LDA para classificação, seguindo pela PNN, com 7,53% e KNN, com 13,94%.

A abordagem de seleção COMB apresentou os melhores resultados, atingindo em média 1,21% classificações incorretas com LDA. A técnica de classificação KNN combinada com a seleção *forwards* apresentou o segundo melhor resultado, com MR média de 1,61%. Os números apresentados entre parênteses na Tabela 2.2 referem-se ao desvio padrão da MR na validação cruzada. A melhor combinação de classificador com abordagem de seleção (LDA

com COMB) também apresentou menor dispersão, com desvio padrão de 0,61, corroborando a robustez do *framework* proposto.

**Tabela 2.2 Taxa de erro na classificação (MR) após a seleção de comprimentos de onda para o banco de dados de biodiesel/diesel**

| Taxa de erro média na classificação (MR) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Forwards* (FS) | | *Backwards* (BE) | | Todas as combinações (COMB) | | Espectro completo | |
| kNN | 1,61% | (+- 0.62) | 3,82% | (+- 2.83) | 2,15% | (+- 0.77) | 13,94% | (+- 0.50) |
| LDA | 2,16% | (+- 0.94) | 4,62% | (+- 0.90) | 1,21% | (+- 0.61) | 4,62% | (+- 1.49) |
| PNN | 2,35% | (+- 0.55) | 4,24% | (+- 1.21) | 2,35% | (+- 0.63) | 7,53% | (+- 1.33) |

O subconjunto de comprimentos de onda selecionados contém intervalos na faixa de 1275-1373 $cm^{-1}$ e 2933-2984 $cm^{-1}$. Tais resultados são consistentes com os encontrados por Ferrão *et al*. [7] para a concentração de enxofre no mesmo banco de dados. Os intervalos retidos referem-se à deformação angular do $CH_2$ e $CH_3$ e dos modos de alongamento do $CH_2$ e $CH_3$, respectivamente, os quais podem estar indiretamente relacionados à concentração de enxofre. A Figura 2.4 mostra o espectro médio das duas classes e os intervalos espectrais selecionados. O subconjunto dos comprimentos de onda retidos é composto por 86 dos 1738 comprimentos originais (4,95% do total), representando uma redução substancial no número de comprimentos utilizados na classificação.

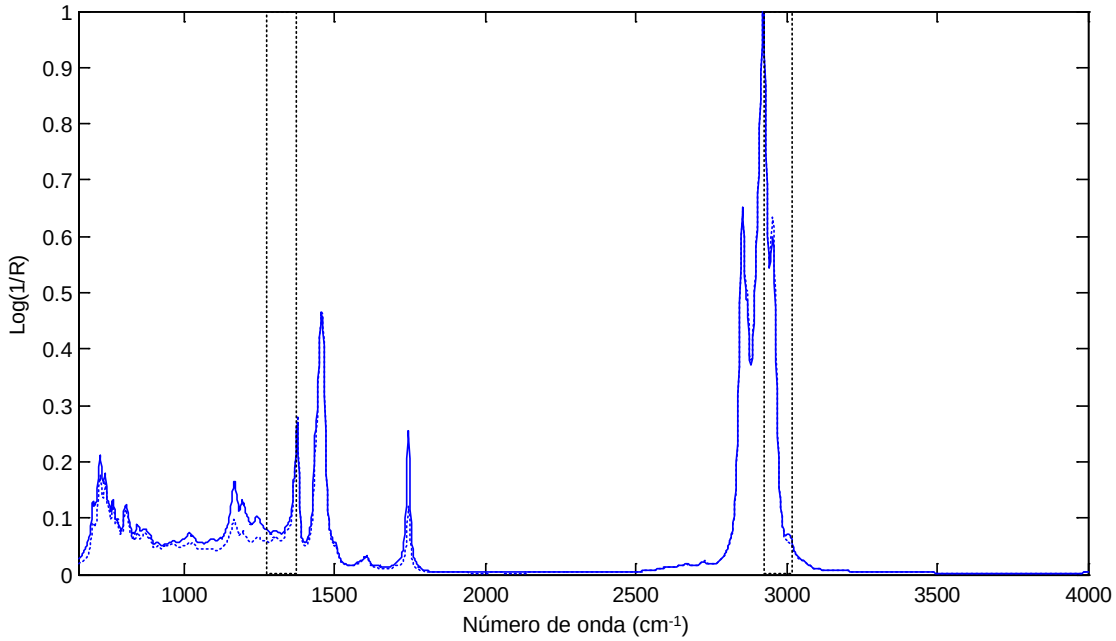


**Figura 2.4 Espectro médio das duas classes (linha sólida e pontilhada) e duas regiões espectrais selecionadas pela LDA no banco de dados de biodiesel/diesel (linhas verticais)**

### 2.3.3 Resultados numéricos para o banco de dados de diesel

O método proposto foi aplicado no banco de dados de diesel descrito na seção 2.3.1 utilizando a validação cruzada de 10 porções. Na Tabela 2.3 encontram-se a taxa de classificação incorreta e seu desvio padrão para os dados em questão. O perfil da curva *D* para o banco de dados é ilustrado na Figura 2.5, tendo o limiar sido estabelecido como a média aritmética de *D*.

**Tabela 2.3 Taxa de erro na classificação (MR) após a seleção de comprimentos de onda para o banco de dados de diesel**

| Taxa de erro média na classificação (MR) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Forwards* (FS) | | *Backwards* (BE) | | Todas as combinações (COMB) | | Espectro completo | |
| kNN | 12,00% | (+- 0,49) | 13,18% | (+- 0,77) | 11,63% | (+- 0,34) | 13,90% | (+- 0,46) |
| LDA | 16,28% | (+- 0,67) | 16,16% | (+- 0,66) | 15,54% | (+- 0,57) | 17,70% | (+- 0,77) |
| PNN | 13,79% | (+- 0,43) | 13,67% | (+- 0,40) | 13,75% | (+- 0,51) | 13,96% | (+- 0,50) |

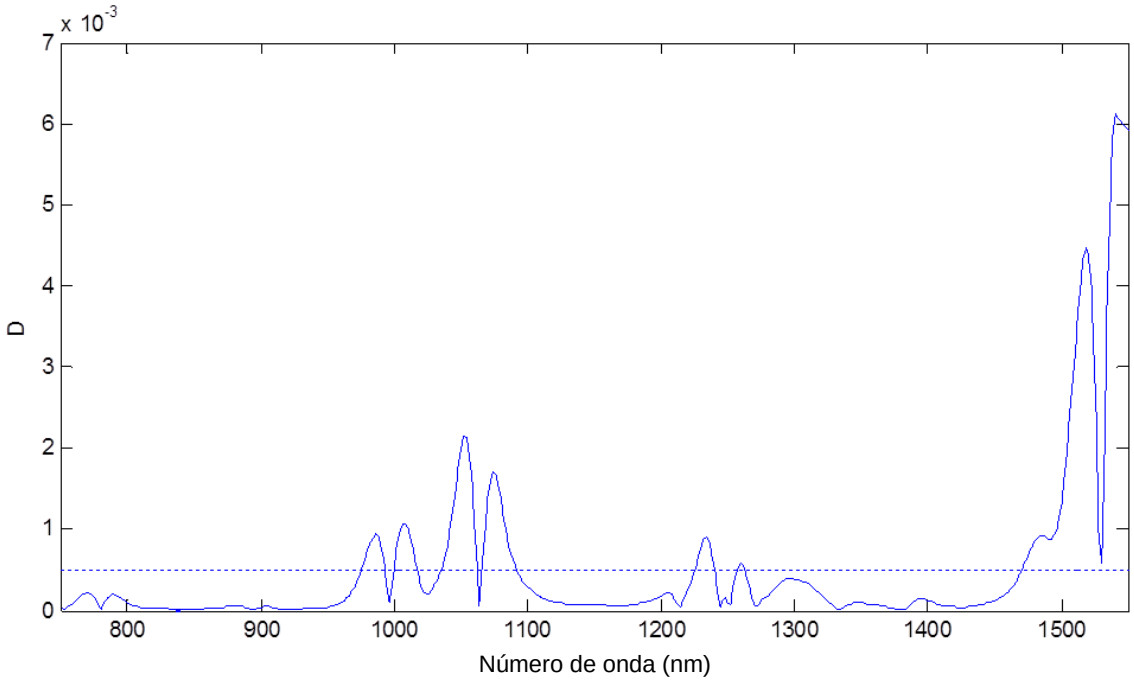


**Figura 2.5 Valores de D (linha sólida), e limiar (linha pontilhada), para o banco de dados de diesel**

A menor MR, 11,63%, foi atingida utilizando a abordagem COMB de seleção juntamente com o algoritmo KNN para classificação. A MR média anterior ao processo de seleção era de 13,90% para o KNN, 13,96% para a PNN e 17.70% para a LDA. A utilização de COMB com KNN também originou o menor desvio padrão (0,34), demonstrando satisfatória robustez.

Comprimentos de onda nas regiões de 922 – 996 nm e 1240 – 1270 nm foram retidos, conforme mostrado na Figura 2.6. Tais regiões estão relacionadas ao terceiro e segundo harmônicos de CH, $CH_2$ e $CH_3$, os quais podem estar associados aos grupos alifáticos que contribuem para aumentar ou diminuir o número de cetano.

Dos 401 comprimentos de onda originais do banco de dados, 93 foram utilizados pelo melhor modelo para classificar as amostras, representando uma redução de 76,81% no número de variáveis. Embora a taxa de erro na classificação não tenha apresentado decrescimento substancial como no banco de dados anterior, o método ainda foi capaz de melhorar o desempenho na classificação através da eliminação dos comprimentos de onda irrelevantes para o problema. Um dos possíveis motivos para tal diferença em relação ao banco de dados anterior é de que a variável utilizada pra estratificar os dados em duas classes

apresenta continuidade próxima ao limiar de separação, diferentemente do anterior, onde havia maior espaçamento.

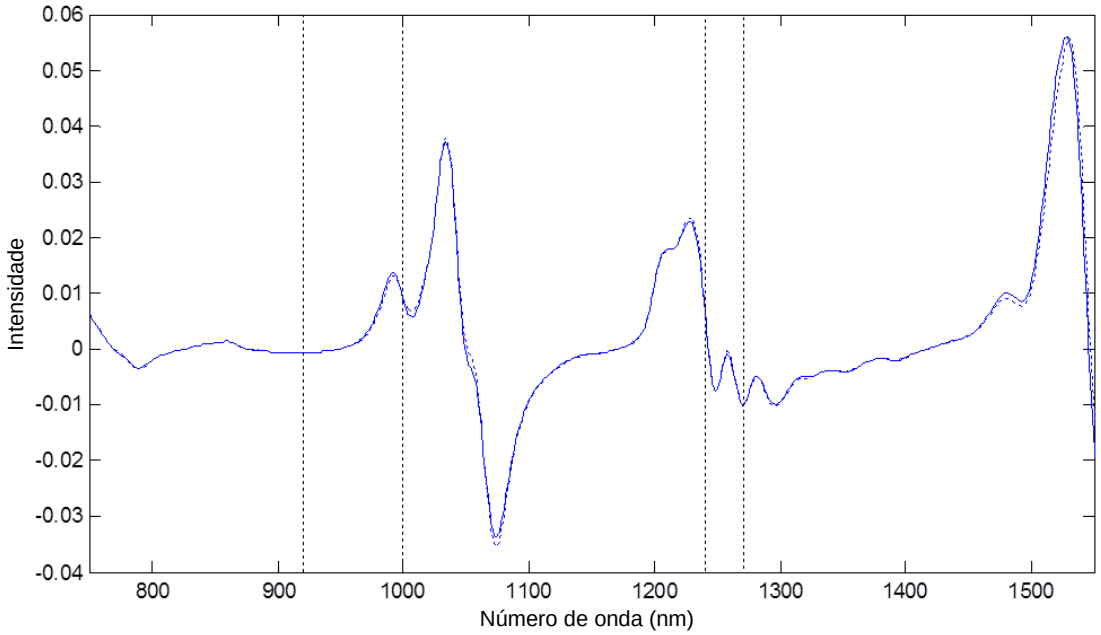


**Figura 2.6 Espectro médio das duas classes (linha sólida e pontilhada) e duas regiões espectrais no banco de dados de diesel (linhas verticais)**

### 2.3.4 Comparação com outros métodos de seleção

O método proposto foi comparado com dois métodos tradicionais de seleção de comprimentos de onda: seleção *forwards* e eliminação *backwards* baseadas em intervalos (referidas como Divisão Equidistante). O procedimento seguido é similar ao iPLS: o espectro foi dividido em intervalos equidistantes, então as abordagens FS e BE foram utilizadas em conjunto com LDA, KNN e PNN como classificadores. Os parâmetros dos algoritmos de classificação foram otimizados conforme descrito na seção 2.3.1. Foram testados diferentes números de intervalos na divisão: 2, 4, 8, 16 e 32.

A Tabela 2.4 compara as abordagens FS e BE baseadas na divisão equidistante (DE) com os melhores resultados apresentados pelo método proposto em ambos os bancos de dados. O desempenho é verificado utilizando a MR média na validação cruzada de 10 porções, sendo o desvio padrão apresentado entre parênteses.

O método proposto apresentou resultados superiores para o banco de dados de Biodiesel/Diesel, com menor MR média e menor variabilidade. Um resultado representativo é

para o algoritmo LDA, onde a abordagem BE apresentou MR de 3,20%, enquanto o método proposto alcançou 1,21%. Para o banco de dados de Diesel, o método proposto apresentou melhores resultados para 2 das 3 ferramentas de classificação, sendo o KNN a exceção, onde a abordagem FS com intervalos equidistantes foi superior ao método proposto. Embora o método proposto não tenha apresentado menor MR em todas as ferramentas de classificação do segundo banco de dados, o desvio padrão foi menor quando comparado ao DE.

**Tabela 2.4 Comparação de desempenho entre o método proposto e outros métodos**

| Biodiesel/Diesel | ED | | | | Método proposto | |
|---|---|---|---|---|---|---|
| | FS | | BE | | | |
| KNN | 1,73% | (+- 0,70) | 3,87% | (+- 1,22) | 1,61% | (+- 0,62) |
| LDA | 3,00% | (+- 0,89) | 3,20% | (+- 0,82) | 1,21% | (+- 0,61) |
| PNN | 2,60% | (+- 0,73) | 5,09% | (+- 1,18) | 2,35% | (+- 0,55) |
| Diesel | ED | | | | Método proposto | |
| | FS | | BE | | | |
| KNN | 11,41% | (+- 0,45) | 11,99% | (+- 0,68) | 11,63% | (+- 0,34) |
| LDA | 15,96% | (+- 0,92) | 16,44% | (+- 0,74) | 15,54% | (+- 0,57) |
| PNN | 13,92% | (+- 0,55) | 13,86% | (+- 0,55) | 13,67% | (+- 0,40) |

## 2.4 Conclusão

No presente artigo um novo método para seleção de comprimentos de onda foi proposto, sendo a classificação de amostras de combustível em classes binárias o objetivo de interesse. O método assemelha-se ao iPLS [20], porém é capaz de selecionar intervalos não equidistantes, sendo os limites dos intervalos derivados da distância entre os espectros médios das duas classes. Inicialmente o espectro médio de cada classe é calculado, sendo a distância entre os espectros médios calculada como o valor absoluto da diferença entre eles. Os picos de distância acima de um limiar são utilizados para identificar os potenciais intervalos, os quais são limitados pelos mínimos locais à esquerda e direita de cada pico de distância. Os intervalos são então iterativamente inseridos em ferramentas de classificação, sendo o melhor modelo determinado pela menor taxa de erro na classificação.

O método proposto foi aplicado em dois bancos de dados (biodiesel e diesel) com diferentes números de amostras e comprimentos de ondas. Para avaliar o desempenho do

método foram testadas três ferramentas de classificação (KNN, LDA e PNN), bem como diferentes abordagens de seleção: eliminação *backwards*, seleção *forwards* e todas as combinações de 1, 2, 3 e 4 intervalos. O método proposto para seleção de intervalos provou ser robusto nas ferramentas de classificação utilizadas, reduzindo a taxa de erro na classificação quando comparado com a utilização de todos os comprimentos de onda. A comparação com técnicas tradicionais de seleção de intervalos baseadas na divisão equidistante do espectro confirmou a superioridade do método proposto.

Oportunidades para pesquisas futuras incluem a extensão do método proposto para o caso de múltiplas classes. Também é de potencial interesse a avaliação do impacto de diferentes valores de limiar para identificação dos picos, o que pode influenciar o número de intervalos selecionados. Por fim, a extensão do método para cenários de predição também é de interesse, visto que o monitoramento de parâmetros de qualidade é tipicamente observado em aplicações de química analítica.

## 2.5 Referências

[1] World Health Organization, WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment, Geneva World Heal. Organ. (2006) 1–22. http://whqlibdoc.who.int/hq/2006/WHO_SDE_PHE_OEH_06.02_eng.pdf?ua=1 (accessed March 13, 2016).

[2] J.C.L. Alves, R.J. Poppi, Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)., Talanta. 104 (2013) 155–61. doi:10.1016/j.talanta.2012.11.033.

[3] A.C. Silva, L.F.B. Lira Pontes, M.F. Pimentel, M.J.C. Pontes, Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, Talanta. 93 (2012) 129–134. doi:10.1016/j.talanta.2012.01.060.

[4] F. Been, Y. Roggo, K. Degardin, P. Esseiva, P. Margot, Profiling of counterfeit medicines by vibrational spectroscopy., Forensic Sci. Int. 211 (2011) 83–100. doi:10.1016/j.forsciint.2011.04.023.

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

[6] M. Fernanda Pimentel, G.M.G.S. Ribeiro, R.S. da Cruz, L. Stragevitch, J.G. a. Pacheco Filho, L.S.G. Teixeira, Determination of biodiesel content when blended with mineral diesel fuel using infrared spectroscopy and multivariate calibration, Microchem. J. 82 (2006) 201–206. doi:10.1016/j.microc.2006.01.019.

[7] M.F. Ferrão, M.D.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. doi:10.1016/j.fuel.2010.09.016.

[8] M.J. Anzanello, K. Fu, F.F. Fogliatto, M.F. Ferrao, M.F. Ferr??o, HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point, Chemom. Intell. Lab. Syst. 145 (2015) 1–6. doi:10.1016/j.chemolab.2015.04.008.

[9] Brazilian National Agency for Petroleum Natural Gas and Biofuel (ANP), Resolution no 45. de 25.08.2014, (n.d.). http://www.anp.gov.br/ (accessed March 13, 2016).

[10] M.J.C. Pontes, C.F. Pereira, M.F. Pimentel, F.V.C. Vasconcelos, A.G.B. Silva, Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification., Talanta. 85 (2011) 2159–65. doi:10.1016/j.talanta.2011.07.064.

[11] R. Muñoz-Olivas, Screening analysis: An overview of methods applied to environmental, clinical and food analyses, TrAC - Trends Anal. Chem. 23 (2004) 203–216. doi:10.1016/S0165-9936(04)00318-8.

[12] S. Li, L. Dai, Classification of gasoline brand and origin by Raman spectroscopy and a novel R-weighted LSSVM algorithm, Fuel. 96 (2012) 146–152. doi:10.1016/j.fuel.2012.01.001.

[13] R.M. Balabin, R.Z. Safieva, Gasoline classification by source and type based on near infrared (NIR) spectroscopy data, Fuel. 87 (2008) 1096–1101. doi:10.1016/j.fuel.2007.07.018.

[14] M. Kim, Y.-H. Lee, C. Han, Real-time classification of petroleum products using near-infrared spectra, Comput. Chem. Eng. 24 (2000) 513–517. doi:10.1016/S0098-1354(00)00522-6.

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

[16] M. Goodarzi, W. Saeys, Selection of the most informative near infrared spectroscopy wavebands for continuous glucose monitoring in human serum, Talanta. 146 (2015) TALD1501704. doi:10.1016/j.talanta.2015.08.033.

[17] M. Khanmohammadi, A. Bagheri Garmarudi, M. De La Guardia, Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear

discriminant analysis, Talanta. 104 (2013) 128–134. doi:10.1016/j.talanta.2012.11.032.

[18] O.Y. Rodionova, a. L. Pomerantsev, NIR-based approach to counterfeit-drug detection, TrAC Trends Anal. Chem. 29 (2010) 795–803. doi:10.1016/j.trac.2010.05.004.

[19] M.J. Anzanello, R.S. Ortiz, R.P. Limbergerb, P. Mayorga, A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes., J. Pharm. Biomed. Anal. 83 (2013) 209–14. doi:10.1016/j.jpba.2013.05.004.

[20] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares egression (iPLS): A comparative chemometric study with an sxample from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[21] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra, Appl. Spectrosc. 56 (2002) 579–585. doi:10.1366/0003702021955358.

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

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

[24] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, 2000.

[25] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

[26] J. Zhao, P. Yu, L. Shi, S. Li, Separable linear discriminant analysis, Comput. Stat. Data Anal. (2012).

[27] K. Fukunaga, Introduction to statistical pattern recognition, 2nd ed., Academic Press, 1990.

[28] R. Balabin, R. Safieva, E. Lomakina, Gasoline classification using near infrared (NIR) spectroscopy data: comparison of multivariate techniques, Anal. Chim. Acta. 671 (2010) 27–35. doi:10.1016/j.aca.2010.05.013.

[29] E. Fix, J.L. Hodges Jr, Discriminatory analysis-nonparametric discrimination: consistency properties, 1951.

[30] R. Timofte, L. Van Gool, Iterative Nearest Neighbors, Pattern Recognit. 48 (2015) 60–72. doi:10.1016/j.patcog.2014.07.011.

[31] Y. Lin, J. Li, M. Lin, J. Chen, A new nearest neighbor classifier via fusing neighborhood information, Neurocomputing. 143 (2014) 164–169. doi:10.1016/j.neucom.2014.06.009.

[32] X. Sun, C.M. Zimmermann, G.P. Jackson, C.E. Bunker, P.B. Harrington, Classification

of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography--fast scanning quadrupole ion trap mass spectrometry., Talanta. 83 (2011) 1260–8. doi:10.1016/j.talanta.2010.05.063.

[33] Z. Liu, Q. Pan, J. Dezert, A new belief-based K-nearest neighbor classification method, Pattern Recognit. 46 (2012) 834–844. doi:10.1016/j.patcog.2012.10.001.

[34] J. Derrac, F. Chiclana, S. García, F. Herrera, Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets, Inf. Sci. (Ny). 329 (2016) 144–163. doi:10.1016/j.ins.2015.09.007.

[35] D.F. Specht, Probabilistic neural networks, Neural Networks. 3 (1990) 109–118. doi:10.1016/0893-6080(90)90049-Q.

[36] A. V Savchenko, Probabilistic neural network with homogeneity testing in recognition of discrete patterns set., Neural Netw. 46 (2013) 227–41. doi:10.1016/j.neunet.2013.06.003.

[37] H. Lin, T. Liang, S. Chen, Estimation of Battery State of Health Using Probabilistic Neural Network, Ind. Informatics, IEEE …. 9 (2013) 679–685. doi:10.1109/TII.2012.2222650.

[38] A.T. Azar, S.A. El-Said, Probabilistic neural network for breast cancer classification, Neural Comput. Appl. 23 (2012) 1737–1751. doi:10.1007/s00521-012-1134-8.

[39] N. Huang, D. Xu, X. Liu, L. Lin, Power quality disturbances classification based on S-transform and probabilistic neural network, Neurocomputing. 98 (2012) 12–23. doi:10.1016/j.neucom.2011.06.041.

[40] Petrobras Distribuidora, Produtos automotivos - oleo diesel, (n.d.). http://www.br.com.br/pc/produtos-e-servicos/para-seu-veiculo/oleo-diesel (accessed February 3, 2017).

## 3. Segundo artigo: Classificação da origem de amostras de vinhos sul-americanos através da análise de concentração elementar

**Felipe Soares**

**Michel José Anzanello**

**Flávio Sanson Fogliatto**

**Marcelo Caetano Alexandre Marcelo**

**Marco Flôres Ferrão**

### Resumo

O presente artigo apresenta o estudo de ferramentas de mineração de dados associadas à informações sobre a concentração de elementos químicos para classificação de amostras de vinho quanto a origem geográfica de produção. Cinquenta e quatro amostras provenientes da Argentina, Brasil, Chile e Uruguai, são analisadas em relação à concentração de quarenta e cinco elementos químicos (Al, Ag, As, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cu, Dy, Er, Eu, Fe, Gd, Ho, K, La, Li, Lu, Mg, Mn, Mo, Na, Nd, Ni, P, Pb, Pr, Rb, Sb, Sn, Se, Sm, Sr, Tb, Ti, Tl, Tm, U, V, Yb e Zn), obtidas através de ICP-MS e ICP-OES. É proposta uma técnica para seleção dos elementos químicos mais importantes na classificação, a qual utiliza o teste Kruskal-Wallis e a análise discriminante linear (LDA) para filtrar os elementos (variáveis) mais importantes e ordená-los. Os elementos ordenados foram apresentados a quatro classificadores, *Naive Bayes* (NB), *Support Vector Machine* (SVM) com *kernel* linear, vizinho mais próximo (NN) e LDA, sendo o subconjunto composto pelos elementos com maior capacidade discriminante determinado para cada classificador. O melhor desempenho foi observado para o SVM, o qual obteve acurácia média de 99,9% retendo em média 6,82 elementos químicos. O melhor resultado ao utilizar todos os elementos químicos foi o NB, com acurácia média de 91,2%. As concentrações dos seis elementos químicos mais frequentemente selecionados pelos classificadores (Mg, RB, V, Li, Tl e Ce) foram comparados através de gráfico *box plot*. A combinação do método proposto de seleção com o

classificador SVM provou ser uma técnica robusta e útil para a verificação da autenticidade de vinhos oriundos dos principais países produtores na América do Sul.

**Palavras-Chave:** Classificação de vinhos, concentração elementar, controle de qualidade, aprendizado de máquina.

## 3.1 Introdução

O crescimento do comércio internacional e o desenvolvimento de mercados potenciais para alimentos e bebidas têm motivado regiões produtoras a desenvolver e aplicar leis ou regulamentações para assegurar o rastreamento da origem dos alimentos [1]. A associação de marcas com o local de origem tende a aumentar a aceitação de tais produtos, gerando vantagem comercial e garantindo maiores preços na comercialização [2,3]. Neste contexto, produtores de alimentos e bebidas têm apresentado crescente interesse em garantir uma precisa classificação dos produtos de acordo com o local de origem, bem como otimizar os mecanismos de confirmação de autenticidade dos produtos [4,5].

A qualidade de vinhos e seus atributos dependem fortemente das características da uva utilizada, das propriedades do solo e das condições climáticas. Tais fatores, quando combinados com técnicas específicas de cultivo, produção e preservação, tornam-se fundamentais para promoção de produtos e sua distinção [6]. Vinhos originados em específicas regiões geográficas e sujeitos a restritas regulações são certificados com uma distinção CDO (*Controlled Denomination of Origin*), a qual assegura sua qualidade superior e o comprometimento com as melhores práticas de produção [7]. Em virtude de tais aspectos, o desenvolvimento de técnicas confiáveis, simples e eficientes que reconheçam precisamente a autenticidade do vinho de acordo com o CDO é um relevante problema para garantir a reputação de tal nicho [6].

Uma técnica analítica para rastrear a origem de produtos alimentícios consiste em analisar a sua composição elementar e concentrações químicas [8]. A análise da concentração dos elementos pode ser determinada por espectrometria de emissão óptica por plasma acoplado indutivamente (ICP-OES) ou espectrometria de massa por plasma acoplado indutivamente (ICP-MS), técnicas essas largamente utilizadas para determinar a qualidade de

produtos como café orgânico [9], ovos [10], arroz [11] e chás [3,12]. Em virtude de sua alta sensibilidade e capacidade em medir isótopos, ICP-MS é tido como um dos métodos mais apropriados para determinar o traço de elementos em vinhos [13]. Tal técnica quantifica a presença de elementos químicos altamente concentrados (como Cd, Cu, Fe, Mn, Sn e Zn) que podem impactar na estabilidade do vinho em termos de cor, sabor e aspectos organolépticos. A concentração de tais elementos pode ser determinada por aspectos geoquímicos, como características do solo no qual as uvas foram cultivadas, bem como por variações no processo produtivo de vinhos. Em vista disso, a mensuração da concentração de elementos químicos é um importante recurso na identificação da origem de vinhos, bem como para corroborar sua autenticidade. Posto isso, identificar os elementos químicos com maior capacidade discriminativa torna-se uma etapa crucial para garantir a correta classificação de amostras de vinho de acordo com a região ou país de produção.

Diversos estudos têm aplicado técnicas estatísticas ou de aprendizado de máquina na classificação de vinhos de acordo com características organolépticas ou origem geográfica [6,14,15], embora poucos tenham focado na seleção das variáveis mais relevantes para discriminar e classificar amostras de vinho. A identificação das variáveis mais importantes torna-se um tópico relevante para a indústria vinícola e autoridades reguladoras, uma vez que um subconjunto reduzido de variáveis, composto pelos elementos químicos mais relevantes, produz modelos de fácil interpretação, além de reduzir o esforço necessário em análises laboratoriais.

O presente artigo propõe uma nova abordagem para seleção de variáveis (elementos químicos) com foco na classificação de amostras de vinho de acordo com a região de origem. O método combina técnicas de seleção de variáveis baseadas em filtragem e *wrapping* em duas etapas operacionais. Na primeira etapa, chamada de filtragem, é empregado o método não-paramétrico Kruskal-Wallis (KW) em cada variável; as que apresentarem *p*-valor maior que determinado limiar *h* são removidas da análise. O objetivo desse procedimento é remover prontamente as variáveis que não apresentam capacidade discriminante significativa para classificar amostras de vinho de acordo com a região de origem, reduzindo o esforço computacional das próximas etapas e potencialmente aumentando o desempenho na classificação. As variáveis remanescentes são utilizadas para desenvolver um índice de importância para as variáveis a partir dos pesos provenientes da *Linear Discriminant Analysis*

(LDA). Tal índice é então utilizado para orientar a seleção das variáveis na etapa seguinte do método proposto. Na segunda etapa, de *wrapping*, uma seleção incremental para frente, ou *forward selection* (FS), é aplicada baseada na ordem das variáveis estabelecida pelo índice de importância anteriormente desenvolvido. As variáveis são inseridas uma a uma, da mais importante para a menos importante, na modelagem; após cada inserção, o desempenho da classificação é verificado. O número de variáveis selecionadas é determinado de acordo com a máxima acurácia na validação cruzada. Buscando verificar a qualidade do conjunto de variáveis selecionadas, diferentes ferramentas de classificação listadas na seção 3.2.3 são testadas.

Existem três principais contribuições no presente estudo. A primeira é a utilização de um teste de significância simples, porém eficiente, o KW, para inicialmente podar as variáveis candidatas, o que reduz o custo computacional dos procedimentos subsequentes, além de melhorar o condicionamento dos dados para o classificador LDA. A segunda é a proposição de um novo índice de importância derivado dos parâmetros da LDA, sendo intuitivo e de simples implementação. Por fim, até o momento a maioria dos estudos para classificação de vinhos foram desenvolvidos em vinhos europeus, portanto o presente trabalho busca estender a análise para os principais países produtores de vinhos da América do Sul.

## 3.2 Materiais e métodos

Esta seção é dividida em quatro subseções. Inicialmente são descritos os instrumentos utilizados para obter as informações químicas das amostras utilizadas nas análises. Em seguida são caracterizadas as amostras de vinho e os procedimentos empregados para obter as informações sobre os elementos químicos. Os métodos multivariados usados no estudo são brevemente descritos, sendo a seção encerrada com a apresentação do procedimento proposto para seleção de variáveis.

### 3.2.1 Instrumentação

Um espectrômetro Optima 2000 DV ICP OES (PerkinElmer, Sheltoon, CT, EUA) foi utilizado para determinação dos elementos principais e secundários (Al, Ba, Ca, Fe, K, Mg, Mn, Na, P, Rb, Sr, Ti e Zn). As seguintes linhas espectrais (comprimentos de onda, em

nanômetros) foram monitoradas: Al (396.153), Ba (233.527), Ca (317.933), Fe (238.204), K (766.490), Mg (285.213), Mn (257.610), Na (589.592), P (213.617), Rb (780.023), Sr (407.771), Ti (334.940) and Zn (206.200). A solução da amostra foi introduzida no plasma através de um nebulizador pneumático acoplado a uma câmara ciclônica. O traço de elementos foi determinado utilizando o instrumento ELAN DRC II (PerkinElmer/SCIEX, Thornhill, Canada). Os parâmetros instrumentais (utilizando o instrumento ICP-MS em modo padrão), tais como fluxo de gás no nebulizador, potência de radiofrequência e voltagem das lentes foram otimizados para obter a intensidade máxima de $115In^{+}$ e intensidade mínima de $Ba^{++}/Ba^{+}$ e $LaO^{+}/La^{+}$. Os seguintes isótopos foram monitorados: 7Li, 9Be, 51V, 53Cr, 58Ni, 59Co, 65Cu, 75As, 82Se, 98Mo, 107Ag, 111Cd, 120Sn, 121Sb, 205Tl, 208Pb, 209Bi, 238U, 139La, 140Ce, 141Pr, 146Nd, 147Sm, 151Eu, 157Gd, 159Tb, 163Dy, 165Ho, 167Er, 169Tm, 172Yb, e 175Lu. Os elementos Be, Ag, Cd, Sn, Sb, Tl, Bi, U, La, Ce, Pr, Nd, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, e Lu foram determinados utilizando um nebulizador ultrassônico. As temperaturas de aquecimento e resfriamento do nebulizador foram ajustadas em 140 ºC e -4 ºC, respectivamente. Um nebulizador concêntrico (Meinhard®, Golden, CO, EUA) acoplado a uma câmara ciclônica foi utilizado para determinação de Li, V, Cr, Ni, Co, Cu, As, Mo, Se e Pb. As câmaras de spray e o nebulizador MicroMist utilizados foram produzidos pela Glass Expansion (Melbourne, Australia), o nebulizador ultrassônico pela CETAC (Omaha, NE, EUA), enquanto o nebulizador Meinhard foi produzido pela Meinhard Associates.

### 3.2.2 Amostras

Cinquenta e quatro (54) amostras de vinho tinto provenientes de regiões distintas de quatro países produtores da América do Sul (Argentina, Brasil, Chile e Uruguai) foram adquiridas em supermercados locais. A origem geográfica e os cultivares foram obtidos dos rótulos das garrafas de vinho. O número de amostras para os quatro países não foi o mesmo, dado que nem todos os cultivares são produzidos em todos os quatro países, ou não são facilmente encontrados no mercado.

As amostras de vinho foram decompostas de acordo com o seguinte procedimento: 1 mL de vinho foi transferido para frascos de politetrafluoretileno (PTFE), seguido da adição de 3mL de $HNO_3$. Após 15 horas, os frascos foram fechados e aquecidos em três etapas: 50

ºC por 1 hora, 100 ºC por 1 hora e 150 ºC por 3 horas. Uma vez resfriada, as soluções obtidas foram transferidas para viais de polipropileno graduados e o volume completado com água até 25 mL. As soluções foram então diluídas 10 vezes com $HNO_3$ a 5% v/v para determinações por ICP-MS ou diretamente analisados por ICP-OES. Todas as amostras foram analisadas em triplicata. Testes de recuperação de analitos e comparação dos resultados obtidos por ICP-MS e ICP-OES foram utilizados para avaliar possíveis interferências e checagem de precisão e acurácia. Para avaliação do procedimento de preparação das amostras, uma amostra de vinho tinto foi diluída em uma solução de ácido nítrico (para obter $HNO_3$ a 5% v/v), deixada em contato com $HNO_3$ (1 mL de vinho + 3 mL de $HNO_3$) ou decomposta na forma descrita anteriormente.

### 3.2.3 Técnicas multivariadas

Foram utilizadas cinco técnicas estatísticas e de aprendizado de máquina no presente estudo: o teste Kruskal-Wallis, com o objetivo de promover uma rodada inicial de descarte das variáveis menos relevantes, e as quatro restantes utilizadas na etapa de classificação. Os principais fundamentos de tais técnicas são brevemente explicados a seguir.

O Kruskal-Wallis (KW) é um teste não paramétrico utilizado para comparar três ou mais populações quando as suposições da análise de variância (ANOVA) não são satisfeitas. A hipótese nula do KW é de que as médias dos postos (*rank*) dos grupos são as mesmas, ou seja, de que as amostras são provenientes de populações com medianas iguais. Para um dado número de *k* grupos, a estatística KW é comparada à distribuição Chi-quadrado. Quando $k = 2$, ou seja, somente dois grupos são considerados, o teste de Wilcoxon pode ser utilizado em substituição ao Kruskal-Wallis. Maiores detalhes podem ser encontrados em [16–18].

A análise discriminante linear, ou *linear discriminant analisys* (LDA), é uma técnica de classificação multivariada introduzida por Fisher em 1936 [19]. A LDA reduz a dimensão dos dados projetando as variáveis originais em um novo sistema de coordenadas que maximiza a separabilidade das *K* classes. O novo subespaço pode ter no máximo $(K - 1)$ projeções, ou componentes, os quais são combinações lineares das variáveis originais na forma $\mathbf{W^T X}$, onde $\mathbf{X}$ é a matriz de dados com amostras em linhas. A matriz de projeção $\mathbf{W}$ na LDA é obtida de tal forma que a distância normalizada entre a média dos grupos, ou a razão

entre a covariância interclasse e intraclasse, seja maximizada. A maximização da razão anterior, conhecida como critério de Fisher, pode ser obtida através da decomposição em autovalores e autovetores, onde cada autovetor projeta os dados originais em um componente. O autovalor associado a cada autovetor simboliza a discriminabilidade de cada componente. Se a matriz de covariância intraclasse é aproximadamente singular, a análise pode ser comprometida ou até mesmo impossível de ser realizada [20].

O método do vizinho mais próximo, ou *nearest neighbor* (NN), é um algoritmo não-paramétrico de classificação muito utilizado em aprendizado de máquina e análise de padrões. A família de vizinhos mais próximos foi primeiramente apresentada por Fix e Hodges (1951) [21], e classifica uma nova observação baseando-se na classe dos *k* vizinhos mais próximos. Diversas variações do algoritmo original foram formuladas, também considerando diferentes métricas de distância. Contudo, como constatado por Cover e Hart (1697) [22], a regra do vizinho mais próximo ($k$ = 1) é a que apresenta a menor probabilidade de erro na classificação. Além disso, no caso de mais de duas classes, a probabilidade de empate é mais provável de acontecer; portanto, ao definir $k$ = 1, o risco de empate é mitigado.

*Naïve Bayes* é um classificador probabilístico utilizado em diferentes cenários, dada sua simplicidade e acurácia usualmente alta [23]. O classificador utiliza a regra de Bayes para estimar a probabilidade posterior de uma amostra pertencer a uma classe dado um conjunto de parâmetros, ou variáveis. O algoritmo computa a probabilidade posterior para todas as classes, e então insere a amostra na classe que apresenta a maior probabilidade posterior. O classificador baseia-se na forte suposição de que as variáveis analisadas sejam independentes entre si, fato que origina seu nome [23]. Maiores detalhes podem ser encontrados em [24–26].

Por fim, a máquina de vetores de suporte, ou *support vector machine* (SVM), é um algoritmo de aprendizagem de máquina supervisionado, introduzido na sua forma atual em 1992 por Boser *et al.* [27] e Cortes e Vapnik em 1995 [28], e posteriormente desenvolvido por diferentes autores [29,30]. A ideia do SVM binário é encontrar um hiperplano que possa separar os dados em duas classes, assumindo que elas sejam linearmente separáveis. Considerando que o conjunto de treino $\{\mathbf{x}_i, y_i\}, i = 1, \dots, N$, tal que $y_i \in \{-1,1\}$ e $\mathbf{x}_i \in \mathcal{R}^D$, onde $\mathbf{x}_i$ é o $i$-ésimo vetor contendo as $D$ variáveis que descrevem um ponto, e $y_i$ a classe de tal ponto, um hiperplano de separação na forma $(\mathbf{w} \cdot \mathbf{X} + b) = 0$ pode ser definido, onde $\mathbf{w}$ é o vetor normal ao hiperplano, $b/\|\mathbf{w}\|$ é a distância perpendicular do plano até a origem do

sistema de coordenadas, e ||**w**|| é a norma euclidiana de **w**. Todos os pontos que estão no hiperplano irão satisfazer a equação. Sendo $d_+$ e $d_-$ as menores distâncias do hiperplano até os pontos mais próximos das classes positivas e negativas, respectivamente, a margem do hiperplano de separação é definida como $(d_+ + d_-)$. No caso binário e linearmente separável, o SVM encontra os valores de **w** e *b* que maximiza a margem. Quando a suposição de separabilidade linear não é verificada, variáveis de folga podem ser adicionadas ao problema, permitindo certo grau de erro, o qual é controlado pela constante *c* [31].

O SVM clássico é definido para classificação binária, porém pode ser facilmente adaptado para tarefas multiclasse. No trabalho de Hsu e Lin [32], diferentes métodos para SVM multiclasse são apresentados e avaliados. Um dos métodos é o um-versus-todos, o qual baseia-se na construção de $K(K-1)/2$ classificadores, sendo cada um treinado em duas das *K* classes. Para predizer um novo dado, uma estratégia de voto é utilizada considerando todos os classificadores treinados.

### 3.2.4 Método proposto para seleção de variáveis

O método proposto no presente artigo pode ser visto como uma combinação em duas fases: Filtro e *Wrapper*. Em um método de filtro, métricas simples (estatística F e ganho de informação, por exemplo) [33] são utilizadas para ranquear e eliminar variáveis em uma etapa de pré-processamento, independente do classificador utilizado [34], sendo o número de variáveis mantidas uma decisão não trivial. Métodos *wrapper* utilizam um classificador como parte do processo de seleção, sendo escolhida alguma métrica de desempenho de classificação como critério para a escolha do melhor subconjunto de variáveis, sendo assim dependente do algoritmo utilizado na classificação [34].

A etapa de filtragem do método proposto é responsável pela geração de um ranking de variáveis. Durante a fase *wrapper*, um método de seleção incremental para frente, ou *forward feature selection*, baseado no ranking obtido na fase de filtragem é utilizado. O melhor conjunto de variáveis é obtido para um dado classificador, escolhendo as variáveis que conduzem à melhor acurácia. A Figura 3.1 demonstra a visão esquemática do método proposto.

Fase 1 – Filtragem de variáveis

O primeiro passo na fase de filtragem é aplicar o teste KW individualmente para cada uma das variáveis originais, retendo somente as quais apresentarem *p*-valor menor do que um limiar *h* pré-definido. O objetivo dessa etapa é remover de imediato as variáveis que não apresentam poder discriminativo satisfatório. Ben Brahim e Limam [35] também propuseram o uso de um filtro simples para reduzir preliminarmente o número de variáveis, seguido de uma abordagem mais sofisticada de seleção. Como apontado pelos autores, o desempenho dos métodos de filtragem é menor do que os *wrappers*, porém proporciona uma redução no custo computacional, o que é desejável na análise de grandes bancos de dados. No método apresentado neste artigo, o uso do KW como fase preliminar objetiva facilitar a execução do método LDA. Dado que a LDA falha quando há mais variáveis do que amostras, a seleção preliminar pelo KW busca mitigar o risco da matriz de covariância intraclasse atingir a singularidade. O subconjunto de variáveis remanescentes da etapa preliminar é referido como subconjunto D, sendo posteriormente utilizado na LDA para gerar uma métrica de importância mais refinada.

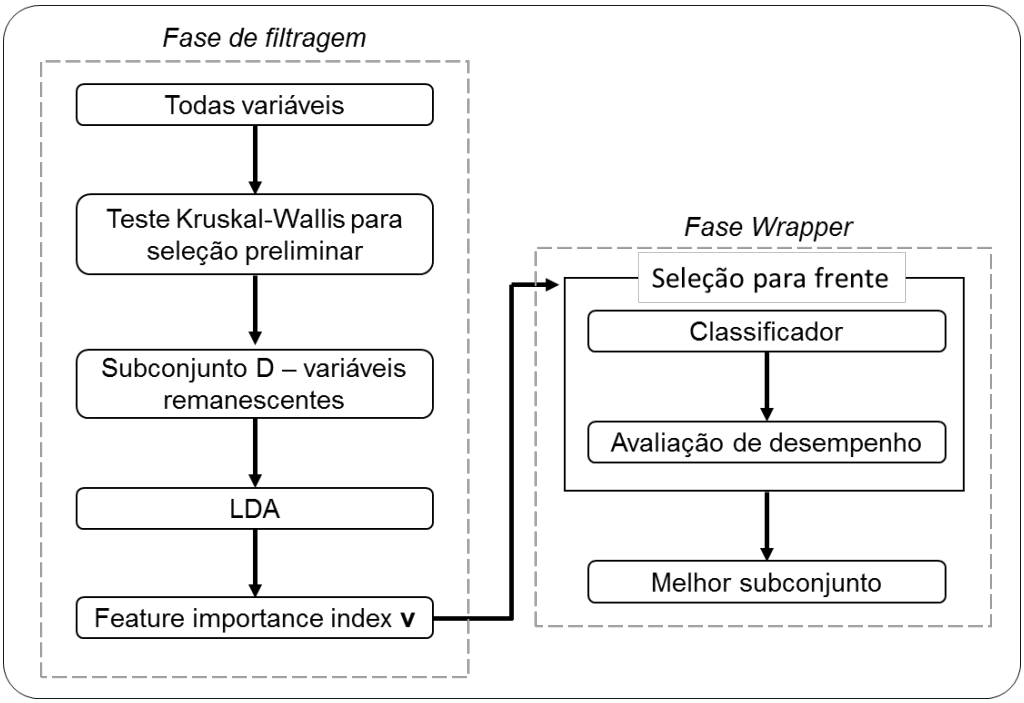


**Figura 3.1 Visão esquemática do método proposto**

No passo seguinte é criado um índice de importância para as variáveis presentes em D, o qual irá guiar o procedimento de seleção *forwards* na fase *wrapper*. O índice é criado aplicando LDA nas variáveis retidas na primeira etapa, armazenando os autovetores e

autovalores, similarmente ao método proposto por Song *et al.* [36]. Define-se $\mathbf{t}_j$ como o *j*-ésimo autovetor ($j = 1, \dots, K-1$) e $\lambda_j$ seu correspondente autovalor. Para cada $\mathbf{t}_j$, os valores absolutos de seus coeficientes são normalizados entre 0 e 1, sendo o vetor normalizado renomeado como $\mathbf{t}_j^N$. O valor $t_{ji}^N$ é a entrada no vetor $\mathbf{t}_j^N$ correspondente a variável *i* ($i = 1, \dots, D$). O índice de importância $c_i$ associado à *i*-ésima variável é calculado da seguinte forma:

$$c_i = \sum_{j=1}^{K-1} \left( t_{ji}^N \lambda_j \right) \text{ para todo } i \text{ em } D \qquad (1)$$

Na Eq (1), os coeficientes dos autovetores fornecem a importância relativa de cada variável na projeção linear, sendo o autovalor a discriminalidade explicada por cada componente. Os valores $c_i$ são então ordenados em forma decrescente em um novo vetor **v**, onde a primeira variável é a mais importante para o propósito de classificação.

Fase 2 – *Wrapper*

Toma-se $B$ como o subconjunto de variáveis retidas no modelo de classificação, as quais serão escolhidas do vetor **v** através de um processo iterativo de seleção *forwards*. O processo inicia-se incluindo a variável da primeira posição de **v**, ou seja, a com maior $c_i$, no subconjunto $B$; é realizada a classificação e sua acurácia armazenada. O processo é iterativamente repetido para as próximas variáveis em **v**, sendo interrompido quando todas as variáveis estão inseridas em $B$. O subconjunto de variáveis responsável pela melhor acurácia é escolhido, passando a ser referenciado como $B^*$.

Alternativamente, diferentes critérios podem ser utilizados para guiar a busca pelas variáveis que farão parte do subconjunto $B^*$. Anzanello *et al*. [37], por exemplo, apresentaram uma abordagem onde o melhor subconjunto de variáveis é o que conduz à menor distância Euclidiana em relação a um cenário ótimo, onde a acurácia de 100% é atingida com somente uma variável. Tal proposição beneficia cenários onde se deseja um reduzido número de variáveis em $B^*$, mesmo que a acurácia de classificação seja levemente comprometida.

O processo iterativo descrito acima é repetido utilizando quatro diferentes técnicas de classificação: vizinho mais próximo (NN), análise discriminante linear (LDA), Naïve Bayes (NB) e máquinas de vetores de suporte (SVM); busca-se o classificador que conduz ao melhor desempenho na classificação. O procedimento apoia-se na validação cruzada de 10 porções

para avaliar a acurácia da classificação. Nesse procedimento, as amostras são aleatoriamente divididas em 10 partes, 9 das quais utilizadas para treinar o classificador e a restante para testar o desempenho, sendo o procedimento de treino/teste repetido de modo que cada uma das 10 partições seja testada. Finalmente, recomenda-se a repetição de todo o procedimento de seleção 100 vezes com diferentes sementes aleatórias, de modo a evitar que um particionamento específico dos dados favoreça o método [38].

## 3.3 Resultados e discussão

Como descrito na seção 3.2.4, o primeiro passo do método proposto é aplicar o teste KW em todas as variáveis e eliminar as que possuírem *p*-valor maior que um determinado limiar *h*. Para melhor estudar a influência do limiar, os valores de 0,01, 0,02 e 0,05 foram testados. Para a fase *wrapper*, quatro classificadores foram testados: vizinho mais próximo (NN), análise discriminante linear (LDA), *Naïve Bayes* e máquina de vetores de suporte (SVM) com núcleo linear. Para o classificador NN foi utilizada a distância euclidiana, enquanto para o SVM o parâmetro *c* foi mantido em 0.35 e as variáveis normalizadas. As análises foram realizadas utilizando o software Matlab R2012b, em um computador com processador AMD Quad-Core A8-4500M e 8 GB de memória RAM.

### 3.3.1 Análise da acurácia de classificação

A Tabela 3.1 apresenta a acurácia média e o número médio de variáveis selecionadas, juntamente com o desvio padrão (entre parênteses), considerando cada combinação de limiar (para o teste KW) e classificador. A acurácia média dos modelos considerando todas as variáveis também é apresentada. É importante notar que o classificador LDA não pôde ser empregado no cenário com todas as variáveis, dado que a alta dimensionalidade resultou em uma matriz de covariância intraclasse aproximadamente singular.

**Tabela 3.1 Acurácia de classificação média e número de variáveis retidas para diferentes combinações de limiar *h* e classificador**

| | *p*-valor = 0,01 | | *p*-valor = 0,02 | | *p*-valor = 0,05 | | Todas variáveis |
|---|---|---|---|---|---|---|---|
| | Acurácia | # variáveis | Acurácia | # variáveis | Acurácia | # variáveis | Acurácia |
| NN | 0,941(0,011) | 10,40(1,70) | 0,958(0,006) | 10,84(0,64) | 0,962(0,010) | 10,13(1,61) | 0,759(0,177) |
| LDA | 0,939(0,014) | 7,79(1,27) | 0,956(0,012) | 8,01(3,25) | 0,951(0,014) | 7,09(0,81) | - |
| NB | 0,936(0,011) | 9,69(1,39) | 0,942(0,008) | 9,1(3,99) | 0,935(0,016) | 8,61(1,47) | 0,912(0,011) |
| SVM | 0,976(0,009) | 7,19(2,44) | 0,999(0,003) | 6,82(1,10) | 0,989(0,010) | 6,73(1,41) | 0,833(0,029) |

Como demonstrado na Tabela 3.1, o classificador SVM superou o desempenho de todos os outros algoritmos em relação à acurácia e número de variáveis retidas. A acurácia média máxima obtida foi 99,9%, retendo em média 6,82 das 45 variáveis originais, utilizando o classificador SVM e *p*-valor de 0,02. O algoritmo SVM obteve resultados similares quando foi utilizado o limiar de 0,05, atingindo uma acurácia média de 98,9%, retendo um número de variáveis ligeiramente menor, porém apresentando um maior desvio padrão em ambas as métricas.

A Figura 3.2 ilustra o perfil de acurácia de um dos modelos SVM conforme as variáveis são inseridas no subconjunto utilizado para classificação, com a seta indicando o ponto de maior acurácia para tal modelo e particionamento dos dados. Para o caso ilustrado, a acurácia máxima pode ser observada ao reter-se 7 ou 8 variáveis, sendo escolhido o modelo com o menor número de variáveis, 7. Os outros classificadores testados (NN, NB e LDA) atingiram acurácias médias no intervalo de 93,46% a 96,26%, também superando o desempenho dos modelos construídos com as 45 variáveis originais. Com base em tais resultados, é recomendado o uso do classificador SVM devido ao seu desempenho superior, simplicidade da fundamentação matemático e disponibilidade em diversos pacotes computacionais.

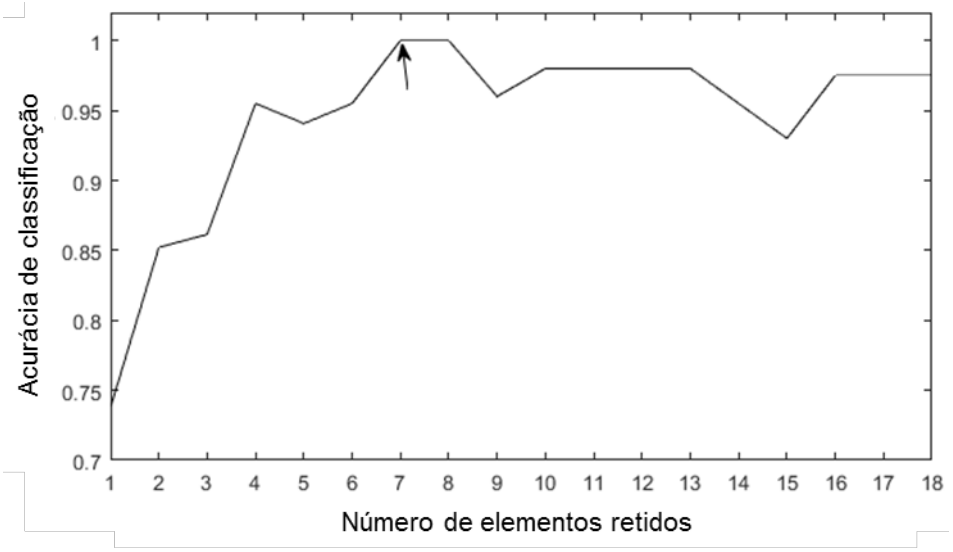


**Figura 3.2 Perfil de acurácia para o SVM conforme as variáveis são inseridas no subconjunto *B***

### 3.3.2 Análise dos elementos selecionados

As Figuras 3.3 a 3.6 demonstram a frequência com que cada uma das 45 variáveis originais foi retida em cada um dos classificadores para as 100 repetições da validação cruzada. Para tal análise, foi considerado somente o melhor *p*-valor para cada classificador. Uma frequência de 1,0 demonstra que dado elemento foi retido em todos os modelos construídos, enquanto uma frequência de 0,0 significa que o elemento não foi utilizado em nenhum dos modelos. É possível notar que os elementos Mg, Rb, V, Li, Tl e Ce foram mais frequentemente selecionados por todos os classificadores para discriminar as amostras de vinho entre as regiões geográficas estudadas. Adicionalmente, é interessante observar que o elemento Hólmio foi consistentemente retido pelos 3 classificadores que obtiveram desempenho inferior, porém nunca retido pelo melhor classificador (SVM com *kernel* linear).

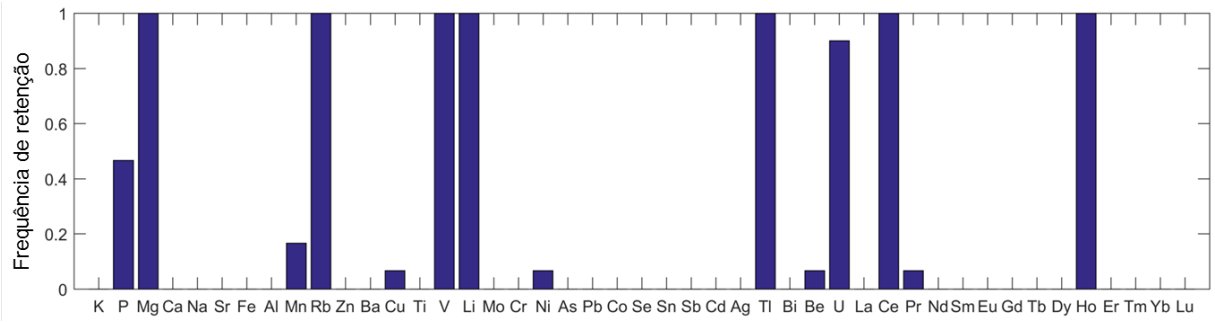


**Figura 3.3 Frequência de retenção para cada elemento utilizando o classificador Naive Bayes**

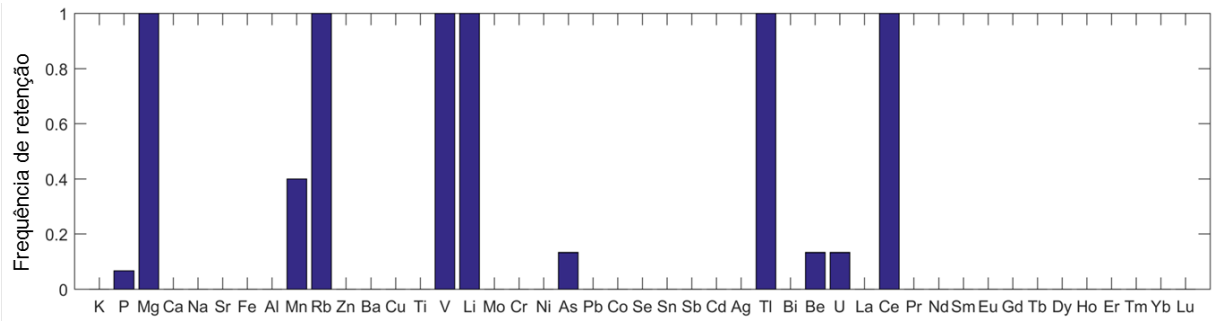


**Figura 3.4 Frequência de retenção para cada elemento utilizando o classificador SVM**

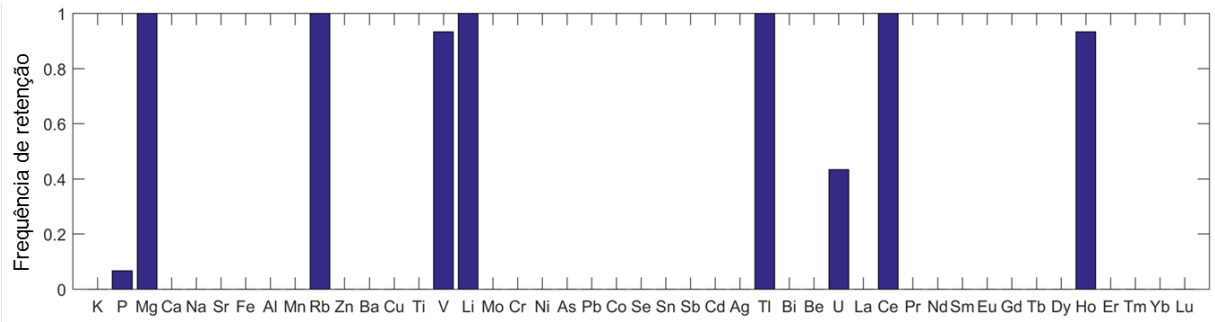


**Figura 3.5 Frequência de retenção para cada elemento utilizando o classificador LDA**

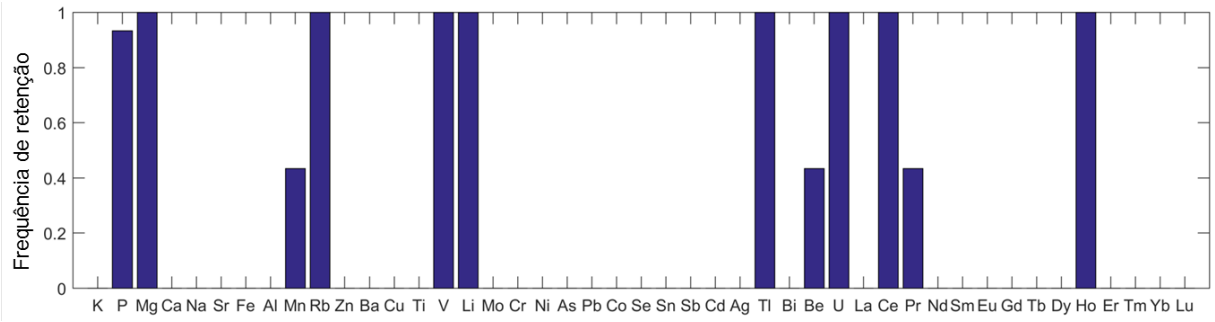


**Figura 3.6 Frequência de retenção para cada elemento utilizando o classificador NN**

A Figura 3.7 apresenta o *box plot* de cada elemento selecionado estratificado nos quatro países produtores. É visível que os vinhos argentinos podem ser discriminados dos vinhos produzidos nos outros países baseando-se nas concentrações de Rb e Li, sendo similares aos vinhos chilenos em termos da concentração de Tl. Vinhos brasileiros podem ser distinguidos dos outros vinhos analisados em relação à concentração de Tl, também sendo possível basear-se no elemento Ce. Os vinhos brasileiros também apresentam em média maiores concentrações de Rb, embora a variabilidade seja grande entre as amostras. Vinhos uruguaios apresentam menores concentrações de Mg, podendo-se utilizar tal característica para identificar os vinhos de tal país. Em relação às porções sobrepostas nos gráficos, é possível notar que os elementos Mg e Rb possuem reduzidas seções de sobreposição. É importante observar que o elemento Tl apresenta uma capacidade discriminante considerável em relação aos quatro países. O único elemento a apresentar grandes porções de sobreposição e concentrações médias similares para os quatro países é o V, porém sua alta frequência de retenção pode ser explicada pelas possíveis interações de tal variável com as outras já incluídas no modelo de classificação. Como apontado por Gheyas e Smith [39], interações entre variáveis que individualmente não apresentam capacidade discriminante podem aumentar o desempenho do modelo como um todo.

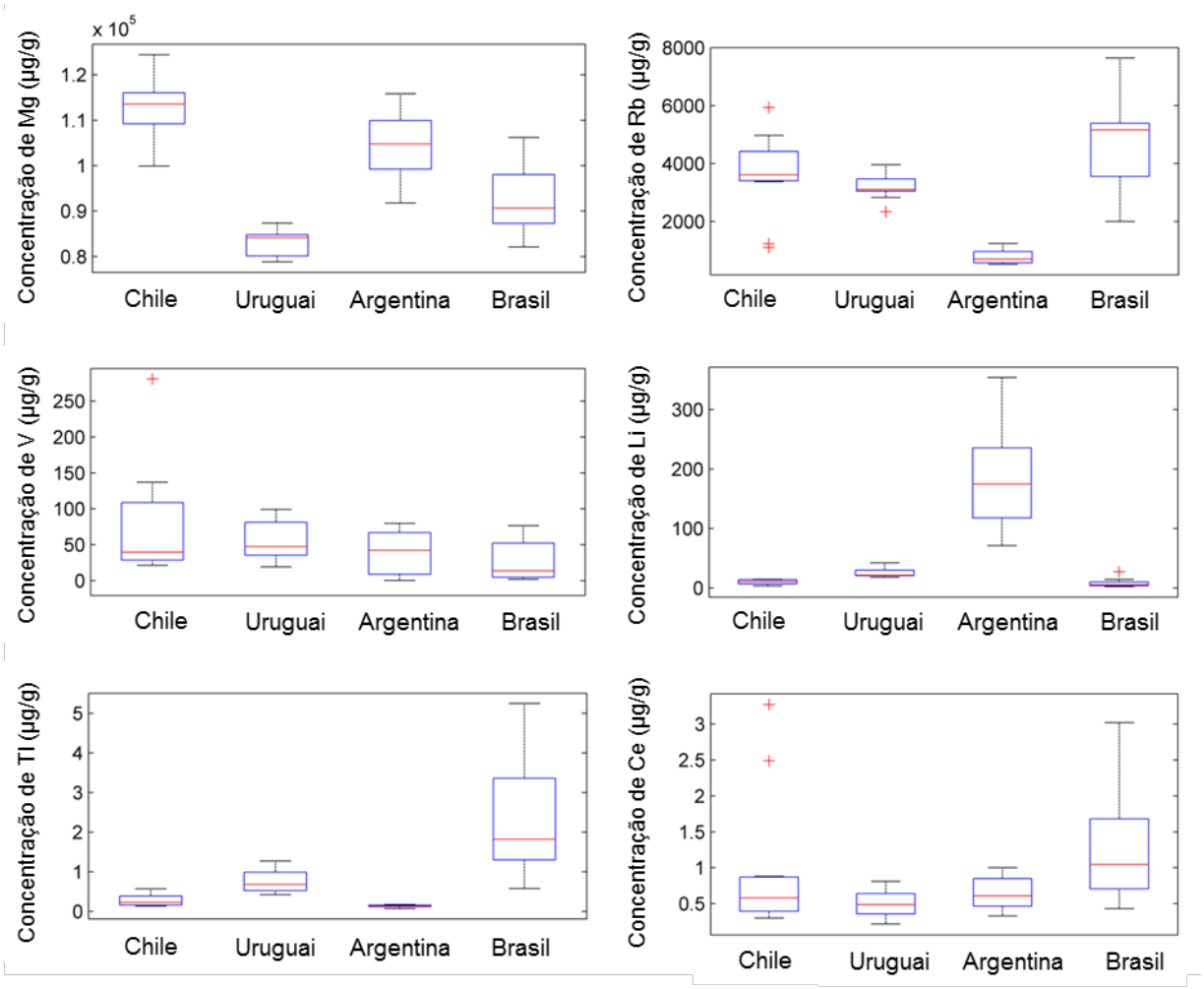


**Figura 3.7 *Boxplots* dos seis elementos selecionados pelo método utilizando SVM (Mg, Rb, V, Li, Tl e Ce)**

## 3.4 Conclusão

O presente artigo propôs a utilização de técnicas multivariadas para classificação de vinhos de quatro países da América do Sul com base na concentração de elementos químicos oriundos destes vinhos. A abordagem de seleção de variáveis proposta se apoia em duas etapas: uma de filtragem inicial através do teste de Kruskal-Wallis e ordenação das variáveis utilizando os coeficientes da LDA, e outra de seleção *wrapper* utilizando quatro classificadores (LDA, NB, SVM e NN).

As proposições foram aplicadas em um banco de dados de vinhos, constituído de 54 amostras e 45 elementos químicos (Al, Ag, As, Ba, Be, Bi, Ca, Cd, Ce, Co, Cr, Cu, Dy, Er,

Eu, Fe, Gd, Ho, K, La, Li, Lu, Mg, Mn, Mo, Na, Nd, Ni, P, Pb, Pr, Rb, Sb, Sn, Se, Sm, Sr, Tb, Ti, Tl, Tm, U, V, Yb e Zn), provenientes de quatro países da América do Sul com grande produção vinícola: Argentina, Brasil, Chile e Uruguai. A utilização do método proposto de seleção de variáveis atrelado à técnica de classificação SVM classificou corretamente 99,9% das amostras teste, retendo em média somente 6,73 dos 45 elementos originais. A classificação com todos os elementos atingiu acurácia média de 91,2% para o classificador Naïve Bayes. Por fim os elementos selecionados foram analisados qualitativamente através de *box plots*. Mg, Rb, V, Li, Tl e Ce foram os elementos mais frequentemente selecionados por todos os classificadores, além do elemento Ho, o qual foi selecionado com frequência pelos classificadores LDA, NB e NN, porém não selecionado pelo SVM.

Diferentes possibilidades para trabalhos futuros podem ser derivadas do presente estudo. A aplicação da técnica de *bootstrapping* durante a etapa de obtenção do índice de importância pode ser utilizada para gerar um intervalo de confiança para a importância de cada variável, fornecendo uma informação mais detalhada para o ranqueamento. Similarmente, o estudo da aplicação de algoritmos estocásticos, como colônia de formigas ou enxame de partículas, para seleção das variáveis pode ser um possível caminho alternativo. Além disso, o estudo de diferentes funções de *kernel* para o SVM pode reduzir ainda mais o número de variáveis necessárias para a classificação.

## 3.5 Referências

[1] H. Zhao, B. Guo, Y. Wei, B. Zhang, Near infrared reflectance spectroscopy for determination of the geographical origin of wheat, Food Chem. 138 (2013) 1902–1907. doi:10.1016/j.foodchem.2012.11.037.

[2] R. Karoui, J. De Baerdemaeker, A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products, Food Chem. 102 (2007) 621–640. doi:10.1016/j.foodchem.2006.05.042.

[3] P.H.G.D. Diniz, A.A. Gomes, M.F. Pistonesi, B.S.F. Band, M.C.U. de Araujo, Simultaneous Classification of Teas According to Their Varieties and Geographical

Origins by Using NIR Spectroscopy and SPA-LDA, Food Anal. Methods. 7 (2014) 1712–1718. doi:10.1007/s12161-014-9809-7.

[4] M.C.A. Marcelo, C.A. Martins, D. Pozebon, M.F. Ferrão, Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate, Anal. Methods. 6 (2014) 7621–7627. doi:10.1039/C4AY01350F.

[5] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment - A review, Anal. Chim. Acta. 891 (2015) 1–14. doi:10.1016/j.aca.2015.04.042.

[6] F. Marini, R. Bucci, A.L. Magrì, A.D. Magrì, Authentication of Italian CDO wines by class-modeling techniques, Chemom. Intell. Lab. Syst. 84 (2006) 164–171. doi:10.1016/j.chemolab.2006.04.017.

[7] S. Gómez-Meire, C. Campos, E. Falqué, F. Díaz, F. Fdez-Riverola, Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques, Food Res. Int. 60 (2014) 230–240. doi:10.1016/j.foodres.2013.09.032.

[8] S.A. Drivelos, C.A. Georgiou, Multi-element and multi-isotope-ratio analysis to determine the geographical origin of foods in the European Union, TrAC - Trends Anal. Chem. 40 (2012) 38–51. doi:10.1016/j.trac.2012.08.003.

[9] R.M. Barbosa, B.L. Batista, R.M. Varrique, V.A. Coelho, A.D. Campiglia, F. Barbosa, The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee, Food Res. Int. 61 (2014) 246–251. doi:10.1016/j.foodres.2013.07.060.

[10] R.M. Barbosa, L.R. Nacano, R. Freitas, B.L. Batista, F. Barbosa, The Use of Decision Trees and Naive Bayes Algorithms and Trace Element Patterns for Controlling the Authenticity of Free-Range-Pastured Hens' Eggs, J. Food Sci. 79 (2014) C1672–C1677. doi:10.1111/1750-3841.12577.

[11] C. Maione, B.L. Batista, A.D. Campiglia, F. Barbosa, R.M. Barbosa, Classification of geographic origin of rice by data mining and inductively coupled plasma mass

spectrometry, Comput. Electron. Agric. 121 (2016) 101–107. doi:10.1016/j.compag.2015.11.009.

[12] A. Moreda-Pineiro, A. Fisher, S.J. Hill, The classification of tea according to region of origin using pattern recognition techniques and trace metal data, J. Food Compos. Anal. 16 (2003) 195–211. doi:10.1016/S0889-1575(02)00163-1.

[13] A. Gonzálvez, A. Llorens, M.L. Cervera, S. Armenta, M. de la Guardia, Elemental fingerprint of wines from the protected designation of origin Valencia, Food Chem. 112 (2009) 26–34. doi:10.1016/j.foodchem.2008.05.043.

[14] P.P. Coetzee, F.P. Van Jaarsveld, F. Vanhaecke, Intraregional classification of wine via ICP-MS elemental fingerprinting, Food Chem. 164 (2014) 485–492. doi:10.1016/j.foodchem.2014.05.027.

[15] S.M. Azcarate, L.D. Martinez, M. Savio, J.M. Camiña, R.A. Gil, Classification of monovarietal Argentinean white wines by their elemental profile, Food Control. 57 (2015) 268–274. doi:10.1016/j.foodcont.2015.04.025.

[16] E. Theodorsson-Norheim, Kruskal-Wallis test: BASIC computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples, Comput. Methods Programs Biomed. 23 (1986) 57–62. doi:10.1016/0169-2607(86)90081-7.

[17] G.D. Ruxton, G. Beauchamp, Some suggestions about appropriate use of the Kruskal-Wallis test, Anim. Behav. 76 (2008) 1083–1087. doi:10.1016/j.anbehav.2008.04.011.

[18] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (1952) 583–621.

[19] R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugen. 7 (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

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

[21] E. Fix, J.L. Hodges Jr, Discriminatory analysis-nonparametric discrimination:

consistency properties, 1951.

[22] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory. 13 (1967) 21–27. doi:10.1109/TIT.1967.1053964.

[23] D.M. Farid, L. Zhang, C.M. Rahman, M.A. Hossain, R. Strachan, Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks, Expert Syst. Appl. 41 (2014) 1937–1946. doi:10.1016/j.eswa.2013.08.089.

[24] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed., Wiley, 2000.

[25] I. Rish, An empirical study of the naive Bayes classifier, in: Int. Jt. Conf. Artif. Intell. - Work. Empir. Methods Artif. Intell., 2001: pp. 41–46.

[26] D.D. Lewis, Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in: C. Nédellec, C. Rouveirol (Eds.), Eur. Conf. Mach. Learn., Springer Berlin Heidelberg, Berlin, Heidelberg, 1998: pp. 4--15. doi:10.1007/BFb0026666.

[27] B.E. Boser, I.M. Guyon, V.N. Vapnik, A Training Algorithm for Optimal Margin Classifiers, Proc. Fifth Annu. ACM Work. Comput. Learn. Theory. (1992) 144–152. doi:10.1.1.21.3818.

[28] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297. doi:10.1023/A:1022627411411.

[29] Y.-J. Lee, O.L. Mangasarian, SSVM: A Smooth Support Vector Machine for Classification, Comput. Optim. Appl. 20 (2001) 5–22. doi:10.1023/A:1011215321374.

[30] S.S. Keerthi, D. Decoste, A Modified Finite Newton Method for Fast Solution of Large Scale Linear SVMs, J. Mach. Learn. Res. 6 (2005) 341–361.

[31] V. Vapnik, Pattern recognition using generalized portrait method, Autom. Remote Control. 24 (1963) 774–780.

[32] C. Hsu, C. Lin, A comparison of methods for multiclass support vector machines, Neural Networks, IEEE Trans. 13 (2002) 415–425. doi:10.1109/TNN.2002.1021904.

[33] D. Du, K. Li, X. Li, M. Fei, A novel forward gene selection algorithm for microarray data, Neurocomputing. 133 (2014) 446–458. doi:10.1016/j.neucom.2013.12.012.

[34] I. Guyon, a Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182. doi:10.1162/153244303322753616.

[35] A. Ben Brahim, M. Limam, A hybrid feature selection method based on instance learning and cooperative subset search, Pattern Recognit. Lett. 69 (2016) 28–34. doi:10.1016/j.patrec.2015.10.005.

[36] F. Song, D. Mei, H. Li, Feature Selection Based on Linear Discriminant Analysis, 2010 Int. Conf. Intell. Syst. Des. Eng. Appl. 1 (2010) 746–749. doi:10.1109/ISDEA.2010.311.

[37] M.J. Anzanello, S.L. Albin, W. a. Chaovalitwongse, Multicriteria variable selection for classification of production batches, Eur. J. Oper. Res. 218 (2012) 97–105. doi:10.1016/j.ejor.2011.10.015.

[38] F. Keynia, A new feature selection algorithm and composite neural network for electricity price forecasting, Eng. Appl. Artif. Intell. 25 (2012) 1687–1697. doi:10.1016/j.engappai.2011.12.001.

[39] I.A. Gheyas, L.S. Smith, Feature subset selection in large dimensionality domains, Pattern Recognit. 43 (2010) 5–13. doi:10.1016/j.patcog.2009.06.009.

# 4 Terceiro artigo: Uso de regressão por vetores de suporte como alternativa robusta para calibração em dados espectroscópicos

**Felipe Soares**

**Michel José Anzanello**

**Resumo**

No presente artigo a regressão por vetores de suporte (SVR) foi estudada como alternativa ao método de mínimos quadrados parciais (PLS) para calibração multivariada em doze bancos de dados de espectroscopia. Foi proposta a utilização do algoritmo *support vector regression – recursive feature elimination* (SVR-RFE) para a seleção dos comprimentos de onda mais informativos para a regressão SVR. Os modelos SVR construídos com espectro completo e por SVR-RFE foram comparados com os modelos PLS com espectro completo, e por seleção através de iPLS, biPLS, siPLS e SPA-PLS. O desempenho foi comparado através da raiz quadrada do erro quadrático médio na partição de teste (RMSEP), particionada pelo algoritmo Kennard-Stone. Os modelos utilizando SVR conduziram aos melhores resultados em oito dos doze bancos de dados, sendo quatro utilizando todo o espectro e quatro através da seleção SVR-RFE. Na comparação entre os algoritmos de seleção de comprimentos de onda, o SVR-RFE apresentou desempenho significativamente superior aos demais métodos, verificado através do teste de Friedman. A utilização do SVR como alternativa ao PLS mostrou-se promissora, especialmente quando é necessário produzir modelos preditivos com reduzido número de comprimentos de onda.

**Palavras-chave:** *Support Vector Regression*, *Partial Least Squares*, Espectroscopia, Seleção de variáveis.

*Artigo submetido o periódico Analytica Chimica Acta (Qualis A1)*

## 4.1 Introdução

Técnicas espectroscópicas têm encontrado aplicação em diversas áreas produtivas, como alimentos, farmacêutica e petróleo [1–3] devido à habilidade em analisar informações químicas de substâncias em estado sólido ou líquido. Além disso, tais técnicas requerem pré-tratamentos simples, ou eventualmente dispensam pré-tratamentos [4]. A combinação de espectroscopia com técnicas multivariadas tem sido um tópico recorrente na literatura, atingindo notáveis resultados em calibrações analíticas, predição de atributos e classificação de amostras [4].

No caso da geração de modelos preditivos (regressões), um grande número de métodos baseados em mínimos quadrados parciais (PLS) vem sendo propostos. O algoritmo da regressão PLS é um dos métodos padrão para calibração multivariada [5], especialmente devido à sua habilidade em lidar com variáveis correlacionadas e com bancos onde existam mais variáveis do que observações [6]. O vasto uso do PLS para calibração multivariada provou sua robustez e versatilidade em diversos cenários, tais como ciência dos alimentos [7], petróleo e combustíveis [1], setores bioquímico [8] e farmacêutico [9]. Apesar da eficiência reconhecida da PLS para tais fins, o uso de técnicas alternativas de calibração multivariada, especialmente a regressão por vetores de suporte (SVR), têm sido estudadas por diferentes autores [8,10–14]. Filgueiras *et al*. [12] utilizaram um comitê de SVR (eSVR) para predição da temperatura de destilação de petróleo bruto utilizando dados de 1H NMR. A acurácia da predição efetuada pelos modelos SVR foi maior do que o PLS em 2 das 3 propriedades analisadas, enquanto os modelos eSVR apresentaram as melhores acurácias nas três propriedades. Zhu *et al*. [15] analisaram o desempenho do SVR na determinação do conteúdo de sólidos solúveis em maçãs através de espectroscopia no infravermelho próximo com o uso de filtro óptico-acústico sintonizável. Os resultados obtidos pelo SVR foram comparados com o PLS e redes neurais artificiais treinadas por retropropagação, tendo o SVR atingido melhores resultados, especialmente em bancos com poucas amostras e dados ruidosos.

Embora um grande número de comprimentos de onda (variáveis obtidas em técnicas espectroscópicas) possa ser obtido rapidamente por equipamentos modernos, tal volume de dados tipicamente tende a reduzir o desempenho exploratório e preditivo de diversas técnicas multivariadas [4,16]. Nesse sentido, torna-se fundamental obter um subconjunto reduzido de comprimentos de onda que revelem as dimensões mais importantes e informativas dos dados

originais [17,18]. Abordagens de seleção de comprimentos de onda também são justificadas na fase de pré-processamento de dados com grande número de variáveis, onde a redução da dimensionalidade pode melhorar a qualidade e a velocidade das análises subsequentes [4,19]. Diferentes métodos para identificação dos comprimentos de onda mais relevantes têm sido propostos no âmbito de aprendizagem supervisionada em análises químicas [4,16,20]. Xiaobo *et al*. [4] trazem em sua detalhada revisão sistemática os métodos mais aplicados na seleção de comprimentos de onda em espectroscopia no infravermelho próximo (NIR), tais como o algoritmo de projeções sucessivas e métodos de seleção de intervalos.

Embora o SVR seja robusto a dados altamente dimensionais [6], o uso de um reduzido número de comprimentos de onda pode reduzir o tempo computacional tanto na criação do modelo, quanto nas futuras predições. Além disso, a facilidade de interpretação dos dados e a redução no custo e tempo de aquisição dos dados são fatores positivos [4]. Zhu et *al*. [15] utilizaram a regressão linear múltipla *stepwise* e um algoritmo genético baseado em PLS (PLS-GA) para selecionar um subconjunto de comprimentos de onda, o qual foi utilizado na construção de modelos SVR. Similarmente, Dashtbozorgi *et al*. [21] também aplicaram o PLS-GA para selecionar os comprimentos de onda usados no SVR. Filgueiras *et al*. [22] dividiram o espectro em intervalos equidistantes e estudaram o desempenho do SVR efetuando combinações dos intervalos (similarmente ao *synergy interval* PLS (si-PLS) [23]). Mais alinhados com as proposições do presente artigo, Lee *et al*. [24] utilizaram os pesos do SVR linear para ranquear os comprimentos de onda. Para tanto, a magnitude dos coeficientes presentes no vetor de pesos é tida como indicador da influência de cada comprimento de onda no modelo linear [25].

O presente artigo baseia-se na aplicação da regressão por vetores de suporte em diversos bancos de dados de calibração multivariada de dados espectroscópicos como uma alternativa ao uso do tradicional PLS. Além disso, também é apresentada uma sistemática de seleção de comprimentos de onda baseada no tradicional algoritmo de seleção de variáveis de eliminação recursiva de variáveis baseada em máquinas de vetores de suporte (SVM-RFE) [25]. Para a etapa de seleção, os pesos do modelo SVR são utilizados para guiar um processo de eliminação de comprimentos de onda. O método inicialmente treina um modelo SVR com todos os comprimentos de onda, sendo eliminado aquele que apresentar o menor valor de importância. Esse processo é repetido recursivamente até restar somente um comprimento de

onda. A ordem de remoção dos comprimentos de onda é tida como o índice de importância, sendo os comprimentos de onda menos importantes eliminados no início do processo iterativo. Os resultados da seleção através do SVR-RFE são comparados com modelos SVR e PLS utilizando todos os comprimentos de onda, além dos resultados obtidos pelos métodos de seleção utilizando PLS.

As principais contribuições do presente artigo são o estudo do desempenho do SVR e do SVR-RFE em diversos bancos de dados de público domínio, facilitando a comparação com outros métodos. Além disso, a aplicação do SVR-RFE e comparação com algoritmos de seleção baseados em PLS permite avaliar o desempenho do método em relação às abordagens padrão utilizadas na área.

O artigo está estruturado como segue. Na seção 4.2 são apresentados os métodos de aprendizado de máquina baseados em vetores de suporte, bem como o método proposto para seleção de comprimento de onda e os bancos de dados utilizados. A seção 4.3 contém os resultados obtidos pelo método SVR-RFE, além dos resultados com espectro completo para PLS e SVR. Também são comparados os métodos de seleção de comprimentos de onda baseados em PLS. Por fim, a seção 4.4 traz as considerações finais sobre o estudo, bem como sugestões de trabalhos futuros.

## 4.2 Materiais e método

Essa seção apresenta a fundamentação do algoritmo de regressão por vetores de suporte, o método de seleção de comprimentos de onda e os bancos de dados utilizados. As análises foram realizadas utilizando os softwares R 3.3.2 e Matlab R2016a, em um computador com processador Intel Core i7 - 6700HQ e 16 GB de memória RAM.

### 4.2.1 Regressão por vetores de suporte

As máquinas de vetores de suporte (SVM) fazem parte de um conjunto de algoritmos de aprendizagem de máquina supervisionada, propostas por Vapnik, em 1995 [26]. No algoritmo SVM de classificação binária, o objetivo é traçar um hiperplano de separação que maximize a margem entre as duas classes. Considerando o conjunto de dados $\{\mathbf{x}_i, y_i\}, i =$

$1, \ldots, N$, tal que $y_i \in \{-1,1\}$ e $\mathbf{x}_i \in \mathcal{R}^D$, onde $\mathbf{x}_i$ é o *i*-ésimo vetor contendo as *D* variáveis que descrevem um ponto, e $y_i$ a classe de tal ponto, o SVM pode ser visto como o seguinte problema de otimização quadrática (QP):

$$\arg\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$y_i(\mathbf{X}_i \cdot \mathbf{w} + b) - 1 + \xi_i \geq 0 \; \forall \; i$$

As variáveis de folga ξ são utilizadas para permitir uma quantidade de erro na classificação quando o problema não é linearmente separável. O grau de erro na classificação é controlado pela constante de compromisso C. Através da solução do problema QP é encontrado o hiperplano de separação definido por $(\mathbf{w} \cdot \mathbf{X} + \mathrm{b}) = 0$, sendo **w** o vetor normal ao hiperplano e *b/*||***w***|| a distância perpendicular do plano até a origem, o qual é baseado nos pontos de treino que satisfazem a restrição do QP, também chamados de vetores de suporte.

O algoritmo SVM também pode ser aplicado em problemas de regressão [27], sendo necessário definir uma função de perda alternativa que penalize valores fora de uma região limite pré-determinada, chamada de ε-*insensitive loss function* (ε-SVR) [28]. A variável independente é estimada por $f(x) = (\mathbf{w} \cdot \mathbf{X}) + b$; quando *f(x)* encontra-se dentro da região de erro aceitável, definida pela constante ε, o modelo não é penalizado. Porém, se o erro é maior que ε, o modelo é penalizado baseado na constante de compromisso C. Portanto, o problema QP pode ser descrito por:

$$\arg\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\frac{1}{N}\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$f(x_i) - y_i \leq \varepsilon + \xi_i$$

$$y_i - f(x_i) \leq \varepsilon + \xi_i^*$$

Similarmente ao classificador SVM, a solução do SVR utiliza os dados de treino, ou vetores, que satisfazem as restrições do QP para construir a função estimadora da variável

independente. Ambas as constantes, ε e C, influenciarão no número de vetores de suporte utilizados, e são usualmente definidas através de validação cruzada durante a fase de treino [29].

Guyon *et al*. [25] propuseram o uso das máquinas de vetores de suporte para selecionar os genes mais importantes na classificação de câncer. SVM-RFE usa os coeficientes do vetor de pesos **w** da solução SVM para ranquear as variáveis em ordem de importância. Um coeficiente de grande magnitude em **w** indica que a variável correspondente tem larga influência na decisão de classificação [25], indicando que as variáveis com pequeno coeficiente possam ser descartadas por possuírem pequena influência. O SVM-RFE recursivamente modela um SVM linear e descarta a variável com menor influência a cada iteração, sendo o algoritmo interrompido quando o número de variáveis desejado seja alcançado, ou reste somente uma variável.

Lee *et al*. [24] usaram os pesos do SVR linear para ordenar os comprimentos de onda utilizando somente uma iteração. No presente artigo é utilizada a mesma abordagem proposta por Guyon *et al*. [25], porém para recursivamente descartar os comprimentos de onda baseado nos pesos da solução SVR a cada iteração. A ordem final de importância dos comprimentos de onda é dada pela ordem de remoção do SVR-RFE, sendo o último comprimento o mais importante.

### 4.2.2 Método proposto para seleção de comprimentos de onda

O ranking gerado através do SVR-RFE é utilizado para guiar a seleção dos comprimentos de onda que serão utilizados no modelo final através de uma abordagem de seleção *forwards*. A cada iteração, baseado no ranking do SVR-RFE, o algoritmo adiciona o comprimento de onda mais importante na modelagem, além de determinar os melhores valores para ε e C baseando-se na validação cruzada *leave-one-out* (LOOCV) na partição de treino. Esse processo é interrompido quando um número máximo pré-definido de comprimentos de onda é atingido. O subconjunto de comprimentos de onda que apresentar a menor raiz quadrada do erro quadrático médio (RMSE) é tido como ótimo. A Figura 4.1 ilustra o procedimento completo do SVR-RFE.

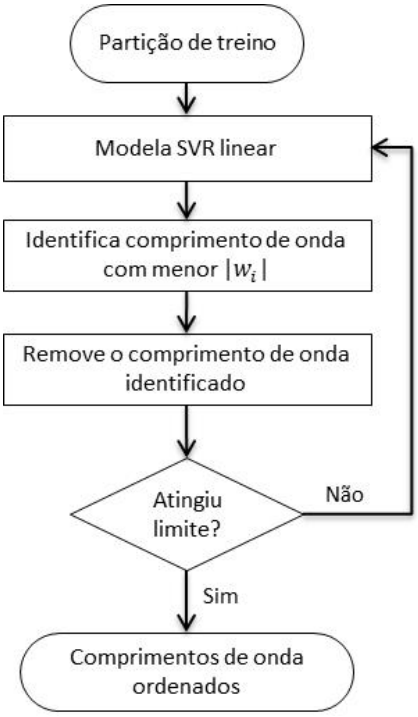


**Figura 4.1 Procedimento de ordenação dos comprimentos de onda por SVR-RFE**

### 4.2.3 Bancos de dados de espectroscopia

Ao todo foram utilizados 12 bancos de dados de espectroscopia no infravermelho próximo (NIR) para avaliar o desempenho da abordagem proposta. Todos os bancos de dados (BD) são de domínio público e provenientes de diferentes campos de aplicação, tais como petróleo, alimentos e biologia. A Tabela 4.1 sumariza os bancos de dados, incluindo a fonte e trabalhos anteriores.

O banco de dados solo 1 é referente à matéria orgânica no solo, enquanto o solo 2 é relacionado à concentração de ergosterol. Os BDs de milho de 1 a 4 são referentes à umidade, óleo, proteína e amido, respectivamente, presentes em amostras de milho. Neste banco de dados estão disponíveis medições em diferentes espectrômetros, sendo que para o presente estudo foram utilizados os dados do instrumento denominado “m5”. Os seis bancos de dados de petróleo são provenientes de um estudo patrocinado pelo exército americano sobre propriedades de óleo diesel. A viscosidade foi medida a 40 ºC, o número de cetano foi determinado de acordo com a norma ASTM D613, o total de aromáticos pela ASTM D5186, a temperatura de solidificação foi determinada em ºC, a densidade determinada de acordo com a norma ASTM D4052, o ponto de ebulição de acordo com a ASTM D86.

**Tabela 4.1 Bancos de dados de espectroscopia NIR utilizados no estudo**

| BD | Campo | Amostras | Comprimentos | Ref. | Fonte |
|---|---|---|---|---|---|
| Solo 1 | Biologia | 108 | 1050 | [30] | http://www.models.life.ku.dk/datasets |
| Solo 2 | Biologia | 108 | 1050 | [30] | http://www.models.life.ku.dk/datasets |
| Milho 1 | Alimentos | 80 | 700 | [31] | http://www.eigenvector.com/data/ |
| Milho 2 | Alimentos | 80 | 700 | [31] | http://www.eigenvector.com/data/ |
| Milho 3 | Alimentos | 80 | 700 | [31] | http://www.eigenvector.com/data/ |
| Milho 4 | Alimentos | 80 | 700 | [31] | http://www.eigenvector.com/data/ |
| Viscosidade | Petróleo | 116 | 401 | [32] | http://www.eigenvector.com/data/ |
| Cetano | Petróleo | 113 | 401 | [32] | http://www.eigenvector.com/data/ |
| Aromáticos totais | Petróleo | 118 | 401 | [32] | http://www.eigenvector.com/data/ |
| Solidificação | Petróleo | 116 | 401 | [32] | http://www.eigenvector.com/data/ |
| Densidade | Petróleo | 122 | 401 | [32] | http://www.eigenvector.com/data/ |
| Ebulição | Petróleo | 113 | 401 | [32] | http://www.eigenvector.com/data/ |

O particionamento dos dados em treino e teste foi realizado utilizando o algoritmo Kennard-Stone [33], de modo que aproximadamente 25% dos dados sejam utilizados para teste. Todos os bancos de dados foram previamente normalizados, tendo sido a média removida e os dados transformados para desvio-padrão unitário.

## 4.3 Resultados e discussão

Nesta seção são apresentados os resultados obtidos pelo método apresentado para seleção de comprimentos de onda baseado em SVR, além da comparação com outros quatro métodos tradicionais em calibração multivariada: *interval partial least squares (iPLS)*, *backward interval partial least squares (biPLS)*, *synergy partial least squares (siPLS)* e o algoritmo de projeções sucessivas (SPA) com PLS. Também são apresentados os resultados utilizando o espectro completo para SVR e PLS.

### 4.3.1 Resultados para o SVR-RFE

O algoritmo SVR-RFE foi utilizado como método de seleção de comprimentos de onda em todos os bancos de dados apresentados na seção 4.2.3. O número máximo de comprimentos de onda a serem retidos foi definido como 30 na etapa de seleção para *forwards*. A Tabela 4.2 contém a raiz quadrada do erro quadrático médio na validação cruzada (RMSECV) da partição de treino e de teste (RMSEP), como o número de comprimentos de onda retidos. É possível verificar que o número de comprimentos de onda retidos na maioria dos modelos demonstrados na Tabela 4.2 não está próximo ao limite máximo de 30, evidenciando a robustez do SVR-RFE.

**Tabela 4.2 Resultados para o SVR-RFE com *kernel* linear**

| Banco de dados | RMSECV | RMSEP | #Comprimentos de onda |
|---|---|---|---|
| Solo 1 | 2,8490 | 1,1375 | 16 |
| Solo 2 | 32,8750 | 32,4142 | 11 |
| Milho 1 | 0,1266 | 0,1192 | 23 |
| Milho 2 | 0,0929 | 0,0942 | 12 |
| Milho 3 | 0,1463 | 0,1754 | 18 |
| Milho 4 | 0,4572 | 0,3565 | 17 |
| Viscosidade | 0,0822 | 0,0691 | 23 |
| Cetano | 2,1250 | 1,7906 | 4 |
| Aromáticos totais | 0,5428 | 0,5120 | 15 |
| Solidificação | 2,3338 | 1,4644 | 7 |
| Densidade | 0,0010 | 0,0009 | 12 |
| Ebulição | 3,1871 | 2,6462 | 16 |

### 4.3.2 Comparação com outros métodos

Para comparar a proposição de utilizar a regressão por vetores de suporte como alternativa ao PLS, quatro tradicionais métodos de seleção de comprimentos de onda baseados em PLS foram estudados. O método iPLS [23] divide o espectro em um número pré-determinado de regiões e constrói PLS locais em cada região, sendo a região com melhor desempenho escolhida. Similarmente, o método biPLS [23] divide o espectro em regiões equidistantes, porém utiliza um algoritmo de eliminação *backwards* para determinar o melhor subconjunto de regiões que produz o modelo mais preciso. Já o algoritmo siPLS [4] utiliza uma busca exaustiva de todas as possíveis combinações de intervalos equidistantes para encontrar o melhor modelo, sendo um problema NP-completo com alto custo computacional.

Diferentemente dos métodos anteriores, o algoritmo SPA [34] utiliza uma abordagem de seleção *forwards* iniciando com somente um comprimento de onda. O comprimento de onda com a maior projeção ortogonal em relação aos já selecionados é adicionado ao subconjunto a cada iteração.

Os quatro métodos de seleção foram otimizados utilizando LOOCV na partição de treino, além dos modelos PLS e SVR com todos os comprimentos de onda. Após a otimização, o desempenho dos modelos foi avaliada na partição de testes. A Tabela 4.3 compara os resultados do método proposto de seleção de comprimentos de onda com os quatro tradicionais métodos anteriormente mencionados, bem como os modelos utilizando todo o espectro. O melhor resultado para cada banco de dados é apresentado em destaque.

A proposição do uso do SVR-RFE para seleção de comprimentos de onda, bem como o SVR para regressão, apresentou os melhores resultados na maior parte dos bancos de dados utilizados quando comparada com outros métodos de seleção baseados em PLS. É importante notar que o algoritmo siPLS teve desempenho superior aos os outros métodos tradicionais de seleção na maioria dos bancos de dados, e no banco de dados Milho 1 superior ao SVR-RFE, porém sua estratégia de força bruta tem alto custo computacional. Também nota-se o pequeno número de comprimentos de onda selecionados pelo SVR-RFE em comparação com os demais métodos, fornecendo modelos com menor complexidade.

Ao analisar a Tabela 4.3 verifica-se que os modelos utilizando a regressão por vetores de suporte apresentaram os melhores resultados em 8 dos 12 bancos de dados estudados. O SVR com todo o espectro obteve a melhor colocação em 4 casos, sendo um deles com 1050 comprimentos de onda. Um dos possíveis motivos para tal desempenho está na natureza das máquinas de vetores de suporte, as quais buscam minimizar o risco estrutural [26] ao invés de somente minimizar o risco empírico. Na minimização do risco empírico, princípio presente em diversos algoritmos de aprendizado de máquina, o objetivo é minimizar o erro na partição de treino, podendo levar a situações de sobreajuste [35,36]. Já na minimização do risco estrutural, o erro na partição de treino guia o procedimento de treinamento juntamente com a minimização da complexidade do modelo, favorecendo a generalização.

Em relação ao custo computacional, a complexidade para treinamento de um modelo SVR é tido como $O(ND)$, onde N é o número de amostras, e D o número de variáveis. Como

usualmente D >> N, e N pode ser considerado como uma constante, a complexidade pode ser considerada *O*(D) para simplificação. Para o algoritmo de eliminação recursiva SVR-RFE, o custo computacional para obtenção do ranking é $O(D^2 \log_2 D)$ [37]. Para o caso dos modelos envolvendo PLS, a complexidade do popular algoritmo SIMPLS é de *O*(ND) [38], que pode ser simplificado para *O*(D) quando D >> N. Para os algoritmos de seleção baseados em intervalos, como iPLS e biPLS, a complexidade é proporcional ao número I de intervalos considerados. Já para o algoritmo siPLS, a complexidade é combinatória em relação ao número de intervalos considerados e o número K de combinações estudadas, tal que a complexidade pode ser dada por $O(\frac{I!}{K!(I-K)!}\ D)$. É facilmente visível que o custo computacional de dividir o espectro em muitos intervalos, e explorar suas combinações, é extremamente alto.

**Tabela 4.3 Comparação com outros métodos – raiz quadrada do erro quadrático médio (RMSE) e número de comprimentos de onda (CO)**

| | SVR-RFE | | SVR Completo | | iPLS | | biPLS | | siPLS | | SPA-PLS | | PLS Completo | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSEP | #CO | RMSEP | #CO | RMSEP | #CO | RMSEP | #CO | RMSEP | #CO | RMSEP | #CO | RMSEP | #CO |
| Solo 1 | 1,1375 | 16 | **0,9014** | 1050 | 1,7282 | 131 | 1,9277 | 261 | 1,7693 | 131 | 2,5199 | 5 | 2,5981 | 1050 |
| Solo 2 | 32,4142 | 11 | 33,1045 | 1050 | 38,6073 | 131 | 35,6714 | 459 | 36,8731 | 131 | **28,7575** | 4 | 29,4638 | 1050 |
| Milho 1 | 0,1192 | 23 | 0,1305 | 700 | 0,1220 | 175 | 0,1303 | 131 | **0,1083** | 87 | 0,1415 | 5 | 0,1241 | 700 |
| Milho 2 | 0,0961 | 26 | 0,0807 | 700 | 0,0854 | 175 | 0,0834 | 109 | 0,0982 | 88 | 0,0845 | 22 | **0,0773** | 700 |
| Milho 3 | 0,1754 | 18 | **0,1639** | 700 | 0,1863 | 175 | 0,1857 | 175 | 0,1995 | 176 | 0,1847 | 9 | 0,1901 | 700 |
| Milho 4 | 0,3565 | 17 | 0,4311 | 700 | **0,3416** | 175 | 0,3955 | 350 | 0,3593 | 88 | 0,4318 | 12 | 0,3454 | 700 |
| Viscosidade | **0,0691** | 23 | 0,0945 | 401 | 0,1393 | 13 | 0,0887 | 237 | 0,1302 | 100 | 0,1139 | 20 | 0,1362 | 401 |
| Cetano | **1,7906** | 4 | 2,0035 | 401 | 1,9184 | 50 | 1,9285 | 50 | 1,8434 | 49 | 2,2270 | 2 | 2,1105 | 401 |
| Aromáticos totais | 0,5120 | 15 | **0,4891** | 401 | 0,8075 | 50 | 1,0087 | 125 | 0,7223 | 50 | 1,1644 | 17 | 0,8287 | 401 |
| Solidificação | **1,4644** | 7 | 2,3983 | 401 | 1,7457 | 100 | 2,0093 | 99 | 2,1718 | 100 | 1,7795 | 14 | 2,1372 | 401 |
| Densidade | 0,0009 | 12 | **0,0006** | 401 | 0,0015 | 100 | 0,0014 | 188 | 0,0011 | 100 | 0,0016 | 17 | 0,0011 | 401 |
| Ebulição | **2,6462** | 15 | 4,2301 | 401 | 5,2528 | 100 | 3,5817 | 161 | 2,6691 | 50 | 5,3784 | 18 | 4,0293 | 401 |

### 4.3.3 Comparação estatística entre algoritmos de seleção

Muitas vezes há a necessidade de obterem-se modelos simplificados, usualmente visando a reduzir o tempo gasto na aquisição dos dados, ou por questões de interpretação. Ao analisar a Tabela 4.3 é possível observar um desempenho geralmente maior do algoritmo SVR-RFE, porém em alguns bancos de dados os desempenhos entre os algoritmos de seleção são similares. Com o objetivo de reduzir o viés subjetivo na comparação de resultados é indicada a utilização de testes estatísticos para identificar a diferença entre tratamentos (ou algoritmos de seleção, neste caso).

Demsar [39] e Garcia e Herrera [40] estudaram a aplicação de testes de hipótese para comparação entre diferentes algoritmos de classificação quando mais de um banco de dados é utilizado. Como evidenciado por Demsar [39], os resultados obtidos quando técnicas de aprendizado de máquina são utilizadas não costumam seguir uma distribuição de probabilidades definida, portanto há a necessidade de se utilizar testes não-paramétricos. A utilização do teste de Friedman para identificação da existência de diferença significativa entre algoritmos é indicada na literatura [39,41,42]. Quando verificada diferença entre os algoritmos, o teste *post-hoc* de Hommel [43] é indicado para comparação múltipla [39].

Os cinco algoritmos de seleção de comprimentos de onda foram comparados utilizando a metodologia acima, baseando-se no RMSEP dos 12 bancos de dados. Inicialmente foi executado o teste de Friedman, alternativa não-paramétrica para a análise de variância, com significância a 5%. Com *p*-valor de 0,002114 a hipótese nula foi rejeitada, indicando a presença de diferença entre os algoritmos. O passo seguinte foi a realização da comparação múltipla entre os algoritmos, utilizando o teste de Hommel. A Tabela 4.4 apresenta os *p*-valores corrigidos pelo método de Hommel para cada par de comparações.

O algoritmo SVR-RFE apresentou desempenho significativamente melhor do que os demais algoritmos de seleção de comprimentos de onda, conforme evidenciado na Tabela 4.4. Ao comparar o algoritmo siPLS com o SVR-RFE, é possível verificar que o *p*-valor ficou próximo ao limite de 0,05, porém ainda apresentando diferença significativa (tendo ocupado a segunda posição em diversos bancos de dados).

**Tabela 4.4 Comparação múltipla entre algoritmos de seleção – *p*-valores**

| | SVR-RFE | iPLS | biPLS | siPLS | SPA-PLS |
|---|---|---|---|---|---|
| SVR-RFE | - | 0,032 | 0,032 | 0,047 | 0,001 |
| iPLS | 0,032 | - | 1,000 | 1,000 | 1,000 |
| biPLS | 0,032 | | - | 1,000 | 1,000 |
| siPLS | 0,047 | 1,000 | 1,000 | - | 0,981 |
| SPA-PLS | 0,001 | 1,000 | 1,000 | 0,981 | - |

Embora o teste estatístico aponte a superioridade do SVR-RFE dentre os algoritmos de seleção de comprimentos de onda estudados, não se pode afirmar que o mesmo apresentará os melhores resultados em todas as ocasiões. Particularidades específicas de cada banco de dados podem favorecer determinado algoritmo de seleção ou regressão, sendo aconselhado, sempre que possível, o estudo de diferentes alternativas.

## 4.4 Conclusão

O presente artigo estudou o uso da regressão por vetores de suporte (SVR) como potencial alternativa à tradicional regressão por mínimos quadrados parciais (PLS) na calibração multivariada de dados espectroscópicos. Também foi proposta a utilização do algoritmo SVR-RFE para seleção de um reduzido número de comprimentos de onda a serem utilizados na regressão por vetores de suporte, o qual foi comparado com os algoritmos de seleção para PLS: iPLS, biPLS, siPLS e SPA-PLS.

Doze bancos de dados de domínio público de espectroscopia NIR foram utilizados para avaliar o desempenho dos métodos de regressão. Os dados foram particionados em treino e teste utilizando o algoritmo Kennard-Stone. Foram treinados modelos PLS e SVR com espectro completo, bem como modelos com comprimentos de onda selecionados. A partição de testes foi utilizada para avaliação do desempenho de cada modelo.

Os modelos utilizando SVR, com espectro completo ou selecionado por SVR-RFE, obtiveram resultados superiores em 8 dos 12 bancos de dados estudados. Na comparação entre os modelos com comprimentos de onda selecionados, o algoritmo SVR-RFE apresentou desempenho significativamente superior aos métodos baseados em PLS, além de selecionar um reduzido número de comprimentos de onda.

O presente estudo comprovou a possibilidade do uso da regressão por vetores de suporte como ferramenta robusta para calibração de dados espectroscópicos. O algoritmo SVR-RFE proposto para seleção de comprimentos de onda provou-se estatisticamente superior aos tradicionais algoritmos PLS, obtendo melhor desempenho e modelos com menor complexidade. Futuras pesquisas incluem a utilização de diferentes *kernels* na regressão SVR, a combinação de predições SVR e PLS, bem como o estudo da influência de diferentes métodos de pré-tratamento dos dados.

## 4.5 Referências


[1] J.C.L. Alves, R.J. Poppi, Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)., Talanta. 104 (2013) 155–61. doi:10.1016/j.talanta.2012.11.033.

[2] A.C. Silva, L.F.B. Lira Pontes, M.F. Pimentel, M.J.C. Pontes, Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods, Talanta. 93 (2012) 129–134. doi:10.1016/j.talanta.2012.01.060.

[3] F. Been, Y. Roggo, K. Degardin, P. Esseiva, P. Margot, Profiling of counterfeit medicines by vibrational spectroscopy., Forensic Sci. Int. 211 (2011) 83–100. doi:10.1016/j.forsciint.2011.04.023.

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

[5] D. Chen, W. Cai, X. Shao, Removing uncertain variables based on ensemble partial least squares, Anal. Chim. Acta. 598 (2007) 19–26. doi:10.1016/j.aca.2007.07.023.

[6] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: A basic tool of chemometrics, Chemom. Intell. Lab. Syst. 58 (2001) 109–130. doi:10.1016/S0169-7439(01)00155-1.

[7] M.G. Ana, A.M. Gómez-Caravaca, R.M. Maggio, L. Cerretani, Chemometric applications to assess quality and critical parameters of virgin and extra-virgin olive oil. A review, Anal. Chim. Acta. 913 (2016) 1–21. doi:10.1016/j.aca.2016.01.025.

[8] M. Muratore, Raman spectroscopy and partial least squares analysis in discrimination of peripheral cells affected by Huntington's disease, Anal. Chim. Acta. 793 (2013) 1–10. doi:10.1016/j.aca.2013.06.012.

[9] M. Dyrby, S.B. Engelsen, L. Nørgaard, M. Bruhn, L. Lundsberg-Nielsen, Chemometric Quantitation of the Active Substance (Containing C≡N) in a Pharmaceutical Tablet Using Near-Infrared (NIR) Transmittance and NIR FT-Raman Spectra, Appl. Spectrosc. 56 (2002) 579–585. doi:10.1366/0003702021955358.

[10] I.A. Naguib, E.A. Abdelaleem, M.E. Draz, H.E. Zaazaa, Linear support vector regression and partial least squares chemometric models for determination of Hydrochlorothiazide and Benazepril hydrochloride in presence of related impurities: A comparative study, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. 130 (2014) 350–356. doi:10.1016/j.saa.2014.04.024.

[11] J.C.L. Alves, C.B. Henriques, R.J. Poppi, J. Cesar, L. Alves, C.B. Henriques, R.J. Poppi, Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system, Fuel. 97 (2012) 710–717. doi:10.1016/j.fuel.2012.03.016.

[12] P.R. Filgueiras, L.A. Terra, E.V.R. Castro, L.M.S.L. Oliveira, J.C.M. Dias, R.J. Poppi, Prediction of the distillation temperatures of crude oils using $^{1}$H NMR and support vector regression with estimated confidence intervals, Talanta. 142 (2015) 197–205. doi:10.1016/j.talanta.2015.04.046.

[13] A. Demiriz, K.P. Bennett, C.M. Breneman, M.J. Embrechts, Support Vector Machine Regression in Chemometrics, in: Comput. Sci. Stat. Proc. ofInterface, Vol. 33, 2001.

[14] U. Thissen, M. Pepers, B. Üstün, W.J. Melssen, L.M.C. Buydens, Comparing support vector machines to PLS for spectral regression applications, Chemom. Intell. Lab. Syst. 73 (2004) 169–179. doi:10.1016/j.chemolab.2004.01.002.

[15] D. Zhu, B. Ji, C. Meng, B. Shi, Z. Tu, Z. Qing, The performance of nu-support vector regression on determination of soluble solids content of apple by acousto-optic tunable filter near-infrared spectroscopy., Anal. Chim. Acta. 598 (2007) 227–34. doi:10.1016/j.aca.2007.07.047.

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

[17] M. Goodarzi, W. Saeys, Selection of the most informative near infrared spectroscopy wavebands for continuous glucose monitoring in human serum, Talanta. 146 (2015) TALD1501704. doi:10.1016/j.talanta.2015.08.033.

[18] M. Khanmohammadi, A. Bagheri Garmarudi, M. De La Guardia, Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis, Talanta. 104 (2013) 128–134. doi:10.1016/j.talanta.2012.11.032.

[19] O.Y. Rodionova, a. L. Pomerantsev, NIR-based approach to counterfeit-drug detection, TrAC Trends Anal. Chem. 29 (2010) 795–803. doi:10.1016/j.trac.2010.05.004.

[20] M.J. Anzanello, R.S. Ortiz, R.P. Limbergerb, P. Mayorga, A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes., J. Pharm. Biomed. Anal. 83 (2013) 209–14. doi:10.1016/j.jpba.2013.05.004.

[21] Z. Dashtbozorgi, H. Golmohammadi, E. Konoz, Support vector regression based QSPR for the prediction of retention time of pesticide residues in gas chromatography-mass spectroscopy, Microchem. J. 106 (2013) 51–60. doi:10.1016/j.microc.2012.05.003.

[22] P.R. Filgueiras, J.C.L.J.C.L.J.C.L. Alves, R.J. Poppi, Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval support vector regression, Talanta. 119 (2014) 582–589. doi:10.1016/j.talanta.2013.11.056.

[23] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares egression (iPLS): A comparative chemometric study with an sxample from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[24] J. Lee, K. Chang, C.-H. Jun, R.-K. Cho, H. Chung, H. Lee, Kernel-based calibration methods combined with multivariate feature selection to improve accuracy of near-infrared spectroscopic analysis, Chemom. Intell. Lab. Syst. 147 (2015) 139–146. doi:10.1016/j.chemolab.2015.08.009.

[25] I. Guyon, J. Weston, S. Barnhill, Gene selection for cancer classification using Support Vector Machines, Mach. Learn. (2009) 389–422. doi:10.1108/03321640910919020.

[26] C. Cortes, V. Vapnik, Support-Vector Networks, Mach. Learn. 20 (1995) 273–297. doi:10.1023/A:1022627411411.

[27] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, V. Vapnik, Support vector regression machines, Adv. Neural Inf. Process. Dystems. 1 (1997) 155–161. doi:10.1.1.10.4845.

[28] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222. doi:10.1023/B:STCO.0000035301.49549.88.

[29] K. Ito, R. Nakano, Optimizing Support Vector regression hyperparameters based on cross-validation, Int. Jt. Conf. Neural Networks. (2003) 2077–2082. doi:10.1109/ijcnn.2003.1223728.

[30] R. Rinnan, A. Rinnan, Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil, Soil Biol. Biochem. 39 (2007) 1664–1673. doi:10.1016/j.soilbio.2007.01.022.

[31] Y.W. Lin, B.C. Deng, Q.S. Xu, Y.H. Yun, Y.Z. Liang, The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework, Chemom. Intell. Lab. Syst. 150 (2016) 58–64. doi:10.1016/j.chemolab.2015.11.003.

[32] Y.H. Yun, W.T. Wang, M.L. Tan, Y.Z. Liang, H.D. Li, D.S. Cao, H.M. Lu, Q.S. Xu, A

strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta. 807 (2014) 36–43. doi:10.1016/j.aca.2013.11.032.

[33] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics. 11 (1969) 137–148. doi:10.2307/1266770.

[34] 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, Chemom. Intell. Lab. Syst. 57 (2001) 65–73. doi:10.1016/S0169-7439(01)00119-8.

[35] D. Wu, Y. He, S. Feng, D.-W. Sun, Study on infrared spectroscopy technique for fast measurement of protein content in milk powder based on LS-SVM, J. Food Eng. 84 (2008) 124–131. doi:10.1016/j.jfoodeng.2007.04.031.

[36] R. Meir, Empirical risk minimization versus maximum-likelihood estimation: a case study, Neural Comput. 7 (1995) 144–157.

[37] Y. Tang, Y.-Q. Zhang, Z. Huang, Development of Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis, IEEE/ACM Trans. Comput. Biol. Bioinforma. 4 (2007) 365–381. doi:10.1109/TCBB.2007.70224.

[38] G. Ji, Z. Yang, W. You, PLS-based gene selection and identification of tumor-specific genes, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 41 (2011) 830–841. doi:10.1109/TSMCC.2010.2078503.

[39] J. Demšar, Statistical Comparisons of Classifiers over Multiple Data Sets, J. Mach. Learn. Res. 7 (2006) 1–30. doi:10.1016/j.jecp.2010.03.005.

[40] S. Garcia, F. Herrera, An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons, J. Mach. Learn. Res. 9 (2008) 2677–2694.

[41] A. Kibekbaev, E. Duman, Benchmarking regression algorithms for income prediction modeling, Inf. Syst. 61 (2016) 40–52. doi:10.1016/j.is.2016.05.001.

[42] G. Loterman, I. Brown, D. Martens, C. Mues, B. Baesens, Benchmarking regression algorithms for loss given default modeling, Int. J. Forecast. 28 (2012) 161–170. doi:10.1016/j.ijforecast.2011.01.006.

[43] G. Hommel, A stagewise rejective multiple test procedure based on a modified bonferroni test, Biometrika. 75 (1988) 383–386. doi:10.1093/biomet/75.2.383.

# 5. Considerações finais

Este capítulo apresenta as conclusões da dissertação, além de sugestões para trabalhos futuros.

## 5.1 Conclusões

Esta dissertação teve como principal objetivo o desenvolvimento de métodos de seleção de variáveis com vistas à classificação e regressão em aplicações de química analítica. Modernas técnicas analíticas permitem a obtenção de um elevado volume de informações, sendo necessário identificar subconjuntos mais informativos dos dados originais para garantir análises mais precisas, interpretáveis e baratas.

Através da análise da literatura científica, objetivos específicos foram definidos. São eles: (*i*) Selecionar intervalos de comprimentos de onda não equidistantes provenientes de espectroscopia para classificar amostras de diesel e misturas de diesel/biodiesel; (*ii*) Aplicar o teste de Kruskal-Wallis como ferramenta de seleção preliminar, possibilitando o emprego de métodos mais sofisticados de seleção em etapa subsequente; (*iii*) Criar um Índice de Importância de Variáveis baseado na LDA que possibilite a posterior inserção ordenada em algoritmos de classificação; (*iv*) Empregar a regressão por vetores de suporte como ferramenta para calibração multivariada de dados espectroscópicos; e (*v*) Adaptar o algoritmo SVM-RFE para seleção de comprimentos de onda em modelos de regressão.

O objetivo (*i*) foi atingido no primeiro artigo, o qual apresentou um método baseado na dissimilaridade entre os espectros médios das classes de diesel e biodiesel/diesel analisadas. A distância entre os espectros médios foi utilizada para gerar intervalos não equidistantes, os quais foram posteriormente inseridos nos classificadores KNN, PNN e LDA. Foi observado que o método proposto reduziu o número de comprimentos de onda, ao passo que aumentou a acurácia de classificação. Além disso, os intervalos selecionados foram interpretados qualitativamente em relação às funções orgânicas responsáveis pela discriminação.

Os objetivos (*ii*) e (*iii*) foram alcançados no segundo artigo, que propôs um método envolvendo uma fase de filtragem e outra de *wrapping* para seleção das variáveis (elementos químicos) mais relevantes para classificação de vinhos de acordo com a origem geográfica. Na filtragem foram utilizados o teste Kruskal-Wallis e a LDA para ordenar as variáveis, sendo

a etapa de *wrapping* responsável pela identificação do melhor subconjunto para cada um dos quatro (NN, LDA, KNN e SVM) classificadores estudado. O melhor classificador, SVM com kernel linear, foi capaz de classificar 99,9% das amostras corretamente na validação cruzada de 10 porções com repetição, tendo retido em média 6,82 elementos químicos dos 45 originais.

Os objetivos (*iv*) e (*v*) foram atingidos no terceiro artigo, onde foi estudada a utilização da regressão por vetores de suporte (SVR) em 12 bancos de dados de espectroscopia, bem como foi proposta a utilização do método SVR-RFE para seleção dos comprimentos de onda. Os modelos envolvendo SVR, com todo o espectro e com comprimentos de onda selecionados, foram comparados com PLS com espectro completo e outras quatro técnicas tradicionais de seleção com PLS. Em 8 dos 12 bancos de dados os modelos envolvendo SVR obtiveram desempenho superior, sendo que na comparação entre os algoritmos de seleção o SVR-RFE obteve o melhor desempenho.

Os três métodos para seleção de variáveis em técnicas analíticas apresentados nesta dissertação abordaram diferentes problemas recorrentes na literatura científica e em aplicações industriais. Enquanto o primeiro artigo buscou encontrar regiões do espectro responsáveis pela maior discriminação entre duas classes de produtos, favorecendo a interpretabilidade do modelo, o terceiro artigo procurou obter um reduzido número de comprimentos de onda que, combinados com uma ferramenta de regressão, fornecessem modelos de predição com menor erro. Já o segundo artigo focou na classificação de produtos na situação onde há mais de duas classes, exigindo uma abordagem diferenciada para seleção de variáveis.

## 5.2 Sugestões para trabalhos futuros

Como extensões das proposições apresentadas nessa dissertação, sugerem-se as seguintes pesquisas futuras:

a) Expandir o conceito da seleção de intervalos de comprimentos de onda através da distância entre espectros médios para classificação em mais de duas classes;

b) Estudo da influência de diferentes pré-tratamentos nos algoritmos de seleção de variáveis;

c) Combinação de predições PLS e SVR para produção de estimativas mais precisas;

d) Construção de intervalos de confiança para índices de importância de variáveis através da técnica de *bootstrapping*;

e) Utilização de algoritmos estocásticos, como colônia de formigas ou enxame de partículas, combinados com índices de importância, para seleção de variáveis;

f) Estudo de diferentes funções de *kernel* para SVM, tanto em regressão como classificação; e

g) Estudar a aplicação de tais técnicas em dados MIR para análise de algas.