UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

TESE DE DOUTORADO

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Porto Alegre, 2017

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Tese submetida ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal do Rio Grande do Sul como requisito final à obtenção do título de Doutor em Engenharia, na área de concentração em Sistemas de Produção.

Orientador: Professor Michel J. Anzanello, *Ph.D.*

Porto Alegre, 2017

Alessandro Kahmann

SELEÇÃO DE VARIÁVEIS EM DADOS DE ESPECTROSCOPIA NO INFRAVERMELHO PARA CONTROLE DE QUALIDADE

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia 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.

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

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

Banca Examinadora:

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

Professor Marcelo Farenzena, Dr. (PPGEQ/UFRGS)

Rafael Scorsatto Ortiz, Dr. (Superintendência Regional do DPF no Rio Grande do Sul)

RESUMO

Nos últimos anos, a espectroscopia no infravermelho (IR) ganhou grande aceitação em diversas áreas de pesquisa por ser uma técnica rápida, simples e não destrutiva que permite a quantificação de diversos componentes químicos em amostras. Apesar de a IR resultar em valores de absorbância que auxiliam na caracterização da amostra, tal técnica acaba por gerar bancos de dados compostos por centenas, ou até milhares, de variáveis altamente correlacionadas e ruidosas, comprometendo o resultado de diversas técnicas de análise multivariada. Dentro deste cenário, esta Tese apresenta novas metodologias para seleção de variáveis, também chamada de seleção de comprimentos de onda quando aplicados em dados de IR, com o intuito de auxiliar o reconhecimento de padrões para o controle de qualidade em diversas áreas. Tais metodologias são apresentadas em três artigos onde as proposições visam à solução de problemas específicos: no primeiro artigo, amostras de erva mate são categorizadas de acordo com seu país de origem através de uma nova metodologia para seleção de variáveis. Para tanto, um problema de Programação Quadrática, combinado com a Informação Mútua entre as variáveis, é utilizado para reduzir a redundância entre as variáveis retidas e maximizar sua relação com o local de origem da amostra; por sua vez, o segundo artigo adequa as proposições do primeiro artigo para um problema de predição, onde o objetivo é determinar a concentração de cocaína e adulterantes em amostras de cocaína laboratoriais e apreendidas; por fim, o terceiro artigo utiliza a estatística do teste de Kolmogorov-Smirnov para duas amostras em uma abordagem de seleção de intervalos de comprimentos de onda com o intuito de identificar falsificações em medicamentos para disfunção erétil. A aplicação dos métodos em bancos de dados com distintas características e a validação dos resultados corrobora a adequabilidade das proposições desta tese.

Palavras-chave: Seleção de Comprimentos de Onda; NIR; FTIR; Classificação; Predição.

ABSTRACT

Over the last few years infrared (IR) spectroscopy gained wide acceptance in many research fields as a quick, simple and non-destructive technique allowing the quantification of many chemical compounds. Although IR provide many absorbance values that helps the sample characterization, this technique also generate databases comprised by hundreds, or even thousands, of highly noisy and correlated wavenumbers, jeopardizing the results of many multivariate analysis techniques. Under such scenario, this thesis presents new variables selection methodologies (also called wavenumber selection when applied in IR data) aimed to recognize patterns for quality control in many areas. Such methodologies are presented in three papers where the propositions are tailored for the solution of specific problems: on the first paper, yerba mate samples are categorized according to their country of origin through a novel variable selection methodology. Thereunto a quadratic programming problem, combined with the Mutual Information among variables, is utilized to reduce the redundancy among variables and increase their relationship with the samples' place of origin; the second paper adequate the first paper propositions for a prediction method which aims to determine cocaine and adulterants concentration in laboratorial and seized cocaine samples; lastly, the third paper uses the two-samples Kolmogorov-Smirnov statistic in an wavenumber interval selection method aimed for the identification of counterfeit erectile dysfunction medicines. The application of the methods in databases with distinct characteristics and the results validation corroborates the suitability of this thesis propositions.

Keywords: Wavenumber selection; NIR; FTIR; Classification; Prediction.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 Introdução

O rápido avanço de tecnologias para análise e monitoramento de processos e produtos tem gerado volumes crescentes de dados, os quais oferecem oportunidades para a identificação de padrões que expliquem eventos das mais diversas naturezas. Tais dados, no entanto, são tipicamente caracterizados por elevado número de variáveis, o que inviabiliza uma análise minuciosa das mesmas. Além disso, parcela significativa das ferramentas multivariadas de análise perde eficiência frente a dados impregnados por ruído ou multicolinearidade, o que é usualmente percebido em bancos com elevada dimensionalidade (LIU; YU, 2005).

Para quantificar a composição química de produtos, de forma a encontrar padrões que permitam verificar determinadas caraterísticas desejáveis, percebeu-se nos últimos anos um aumento substancial no número de estudos que se apoiam na espectroscopia no infravermelho (IR); tal técnica é tida como de simples execução, rápida e não destrutiva, permitindo estimar a composição química de observações com baixa preparação prévia (CRAIG et al., 2014; LIU; YANG; DENG, 2015; ZHANG; ZHANG; IQBAL, 2013). Apesar de dados do tipo NIR fornecerem diversas informações relevantes para a caracterização de amostras, tipicamente são compostos por diversas características indesejáveis a análises multivariadas. Tal cenário justifica a necessidade da utilização de técnicas de mineração de dados para identificação apropriada de padrões (MAIONE et al., 2016).

A mineração de dados consiste no processo computacional de identificação de padrões em grandes bancos de dados, tendo como principal objetivo extrair informações relevantes e implícitas destes bancos. Dentre as técnicas de mineração de dados, destaca-se a seleção de variáveis (também chamada de seleção de comprimentos de onda quando aplicada a dados do tipo NIR), a qual objetiva identificar as variáveis mais importantes através da remoção de variáveis irrelevantes ou que prejudiquem a interpretação dos dados. Os benefícios desta redução incluem melhor interpretação dos resultados, maior rapidez computacional na geração de modelos e aumento de acurácia de técnicas de predição e classificação. Tais benefícios estão alinhados com as justificativas trazidas pela literatura para seleção de variáveis: (*i*) evitar o *overfitting* de modelos; (*ii*) produzir modelos com menor necessidade de processamento e melhor custo-efetividade; e (*iii*) permitir um conhecimento aprofundado do processo, uma vez que a identificação de variáveis com base no conhecimento empírico de especialistas é frequentemente sujeita a equívocos (BLUM; LANGLEY, 1997; GUYON;

ELISSEEFF, 2003; HASTIE; TIBSHIRANI; FRIEDMAN, 2009; KETTANEH; BERGLUND; WOLD, 2005; SAEYS; INZA; LARRAÑAGA, 2007).

Dentro do escopo desta tese, a seleção de variáveis (ou comprimentos de onda) tem por objetivo criar um modelo de análise selecionando regiões do espectro que sejam significativas, reduzindo a quantidade de variáveis e, consequentemente, removendo dados ruidosos, redundantes, ou irrelevantes. A seleção de regiões relevantes do espectro também contribui na criação de modelos mais simples e, consequentemente, mais fáceis de interpretar, uma vez que tais modelos explicitam não apenas a relação dos comprimentos de onda entre si, como também sua relação com a variável resposta (XIE; YING; YING, 2009; ZHANG; ZHANG; IQBAL, 2013). Por fim, a remoção de comprimentos de onda que não possuem informações relevantes reduz a complexidade do modelo, resultando em ganhos computacionais e de precisão (CHEN et al., 2013).

Existem dois propósitos principais alinhados com a seleção de comprimentos de onda: (i) predição, onde o objetivo é encontrar um conjunto de variáveis independentes que viabilizam melhor predição de variáveis dependente quantitativa (GAUCHI; CHAGNON, 2001; PEREIRA et al., 2011); e (ii) classificação, a qual objetiva encontrar o conjunto de variáveis independentes que melhor insira novas observações em categorias (ANZANELLO et al., 2015; DINIZ et al., 2014). Para atingir tais objetivos, os métodos de seleção de comprimentos de onda se dividem em duas frentes: (i) seleção de comprimentos de onda individuais, como em Anzanello et al. (2015), e (ii) seleção de intervalos de comprimentos de onda, como em Soares et al. (2017) e Marcelo et al. (2014). Os artigos apresentados nesta tese abordam metodologias para classificação e predição voltadas à seleção individual e de intervalos de comprimentos de onda. O primeiro artigo apresenta um método de seleção de comprimentos de onda que visa à identificação do país de origem de amostras de erva mate; por sua vez, o segundo artigo apresenta um método com o intuito de predizer a concentração de cocaína e adulterantes em amostras de cocaína; por fim, um método de seleção de intervalos de comprimentos de onda é proposto no terceiro artigo com o objetivo de identificar falsificações de remédios para disfunção erétil.

1.1 TEMA E OBJETIVOS

O tema da presente tese é a proposição de novas abordagens para seleção de comprimentos de onda com vistas à classificação de amostras e predição de suas propriedades. Os objetivos específicos são:

(i) Criar novos índices de Importância de Comprimentos de onda com vistas a mensurar a relevância das variáveis analisadas;

(ii) Comparar métodos de seleção de intervalos de comprimentos de onda e de seleção individual de comprimentos de onda;

(iii) Comparar os resultados dos métodos propostos a outras metodologias de seleção de variáveis mais difundidas, aplicando-os em bancos de dados reais; e

(iv) Verificar a adequabilidade em dados oriundos de NIR com diferentes características em relação à dimensionalidade e características da variável resposta;

1.2 JUSTIFICATIVA DO TEMA E DOS OBJETIVOS

Nos últimos anos, a espectroscopia no infravermelho (IR) ganhou grande aceitação em diversas áreas de pesquisa por ser uma técnica rápida, simples e não destrutiva que permite a quantificação de diversos componentes químicos em amostras. A IR, combinada com diferentes tipos de técnicas de análise multivariada, tem sido utilizada nas mais diversas áreas de pesquisa, as quais incluem análise forense (BORILLE et al., 2017; MARCELO et al., 2016), engenharia de combustíveis (CRAMER; MORRIS; ROSE-PEHRSSON, 2010; SUN et al., 2011) e engenharia de alimentos (MARQUETTI et al., 2016; ZHANG et al., 2015).

Apesar da IR resultar em valores de absorbância que auxiliam na quantificação de diversos componentes químicos, a técnica acaba por gerar bancos de dados compostos por centenas, ou até milhares, de variáveis altamente correlacionadas e ruidosas, comprometendo o resultado de diversas técnicas de análise multivariada. Dentro deste cenário, a mineração de dados voltada à seleção de regiões relevantes do espectro se mostra necessária tanto para aumentar a qualidade da análise multivariada como para reduzir a influência de dados mal condicionados, gerando assim modelos mais simples e eficientes em termos de interpretação (HE et al., 2014; MARCELO et al., 2014). Tais benefícios justificam as abordagens aqui propostas em termos práticos.

Percebe-se ainda que diversas abordagens clássicas da literatura acabam por não mais produzir modelos satisfatórios quando aplicadas a dados espectrais mais detalhados, os quais são decorrentes da modernização das técnicas experimentais utilizadas na obtenção do NIR. Desta forma, é possível perceber no âmbito acadêmico um grande esforço devotado ao desenvolvimento de abordagens mais robustas e aptas à aplicação em bancos com tendência crescente de dimensionalidade, o que contribui na justificativa acadêmica do tema desta tese.

1.3 Estrutura da Pesquisa

A pesquisa é realizada em três etapas, onde cada etapa corresponde a um artigo que visa a atender os objetivos específicos supracitados. Com relação à estrutura da tese, cada artigo corresponde a um dos capítulos subsequentes a presente introdução. A Tabela 1-1 apresenta os artigos, ferramentas utilizadas e contribuição científica de cada artigo.

Artigo	Título	Ferramentas utilizadas	Contribuição científica
1	Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin	Programação Quadrática, Informação Mútua, Máquina de Suporte Vetorial, Análise Discriminante, K-vizinhos próximos	Proposição de um novo método de seleção de comprimentos de onda para categorização de amostras de erva mate de acordo com seu país de origem
2	Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples	Programação Quadrática, Informação Mútua, Regressão Linear Múlipla, Regressão por Componentes Principais, Regressão por Mínimos Quadráticos Parciais	Proposição de um novo método de seleção de comprimentos de onda para predição da concentração de cocaína e adulterantes em amostras de cocaína
3	Spectra interval selection to identify counterfeit medicines	Teste de Kolmogorov-Smirnov para duas amostras, Máquina de Suporte Vetorial, Análise Discriminante, K-vizinhos próximos	Proposição de um novo método de seleção de intervalos de comprimentos de onda para categorização de medicamentos falsificados e originais

Tabela 1-1 – Descrição dos artigos da tese

Dentre as principais contribuições desta pesquisa destacam-se: a integração da Programação Quadrática à Informação Mútua voltada à geração de um índice de importância de comprimentos de onda aplicável à seleção de variáveis em problemas de classificação e predição; a proposição de um índice de importância de intervalos de comprimentos de onda através da estatística do teste de Kolmogorov-Smirnov para duas amostras; e a comparação entre a seleção de comprimentos de onda individuais e a seleção de intervalos de comprimentos de onda, duas abordagens utilizadas na literatura.

1.4 Delimitações da Pesquisa

A pesquisa considera em seu escopo somente ferramentas clássicas de análise multivariada, bancos de dados de NIR voltados a problemas específicos e a validação dos resultados através da comparação com os resultados de técnicas difundidas de seleção de comprimentos de onda ou por especialistas. Desta forma, não foram considerados nesta pesquisa:

- Técnicas de análise multivariada alternativas às existentes na literatura;
- Dados públicos de NIR;
- Modelos alternativos ao *wrapper* com a inclusão de variáveis de forma *forward* ordenada;
- Avaliações de modelos baseados em métricas outras que acurácia e dimensionalidade; e
- A interpretação detalhada dos modelos gerados, analisando apenas os comprimentos de onda selecionados e não suas implicações.

1.5 Referências

ANZANELLO, M. J., KAHMANN, A., MARCELO, M. C. A., MARIOTTI, K. C., FERRÃO, M. F., ORTIZ, R. S. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, 2015. v. 115, p. 562–569.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, 1997. v. 97, n. 1–2, p. 245–271.

BORILLE, B. T., MARCELO, M. C. A., ORTIZ, R. S., MARIOTTI, K. de C., FERRÃO, M. F., LIMBERGER, R. P. Near infrared spectroscopy combined with chemometrics for growth stage classification of cannabis cultivated in a greenhouse from seized seeds. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, 2017. v. 173, p. 318–323.

CHEN, M., KHARE, S., HUANG, B., ZHANG, H., LAU, E., FENG, E. Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. **Industrial and Engineering Chemistry Research**, 2013. v. 52, n. 23, p. 7886–7895.

CRAIG, A. P., FRANCA, A. S., OLIVEIRA, L. S., IRUDAYARAJ, J., ILELEJI, K. Application of elastic net and infrared spectroscopy in the discrimination between defective and non-defective roasted coffees. **Talanta**, 2014. v. 128, p. 393–400.

CRAMER, J. A.; MORRIS, R. E.; ROSE-PEHRSSON, S. L. Use of Genetic Algorithms To Improve Partial Least Squares Fuel Property and Synthetic Fuel Modeling from Near-Infrared Spectra. **ENERGY & FUELS**, 2010. v. 24, p. 5560–5572.

DINIZ, P. H. G. D., GOMES, A. A., PISTONESI, M. F., BAND, B. S. F., de ARAÚJO, M. C. U. Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA. **Food Analytical Methods**, 2014. v. 7, n. 8, p. 1712–1718.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, 2001. v. 58, n. 2, p. 171–193.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, 2003. v. 3, n. 3, p. 1157–1182.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **Elements**, 2009. v. 1, p. 337–387.

HE, K., CHENG, H., DU, W., QIAN, F. Online updating of NIR model and its industrial application via adaptive wavelength selection and local regression strategy. **Chemometrics and Intelligent Laboratory Systems**, 2014. v. 134, p. 79–88.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS with very large data sets. **Computational Statistics and Data Analysis**, 2005. v. 48, n. 1, p. 69–85.

LIU, C.; YANG, S. X.; DENG, L. Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning. **Journal of Food Engineering**, 2015. v. 161, p. 16–23.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **Ieee Transactions on Knowledge and Data Engineering**, 2005. v. 17, n. 4, p. 491–502.

MAIONE, C., LEMOS, B., DOBAL, A., BARBOSA, F., MELGAÇO, R. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. **Computers and Electronics in Agriculture**, 2016. v. 121, p. 101–107.

MARCELO, M. C. A., MARTINS, C. A., POZEBON, D., FERRÃO, M. F. Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate. **Anal. Methods**, 2014. v. 6, n. 19, p. 7621–7627.

MARCELO, M. C. A., FIORENTIN, T. R., MARIOTTI, K. C., ORTIZ, R. S., LIMBERGER, R.P. Analytical Methods Determination of cocaine and its main adulterants in seized drugs from Rio Grande do Sul , Brazil , by a Doehlert optimized LC-DAD method. 2016. p. 5212–5217.

MARQUETTI, I., LINK, J. V., LEMES, A. L. G., SCHOLZ, M. B. dos S., VALDERRAMA, P., BONA, E. Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. **Computers and Electronics in Agriculture**, 2016. v. 121, p. 313–319.

PEREIRA, A. C., REIS, M. S., SARAIVA, P. M., MARQUES, J. C. Madeira wine ageing prediction based on different analytical techniques: UV-vis, GC-MS, HPLC-DAD. **Chemometrics and Intelligent Laboratory Systems**, 2011. v. 105, n. 1, p. 43–55.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. **Bioinformatics**, 2007. v. 23, n. 19, p. 2507–2517.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, 2017, v.167, n.15, p. 171-178.

SUN, X., ZIMMERMANN, C. M., JACKSON, G. P., BUNKER, C. E., HARRINGTON, P. B. 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**, jan. 2011. v. 83, n. 4, SI, p. 1260–1268.

XIE, L.; YING, Y.; YING, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. **Journal of Food Engineering**, 2009. v. 94, n. 1, p. 34–39.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 128, p. 17–24.

ZHANG, Y., ZHENG, L., LI, M., DENG, X., JI, R. Predicting apple sugar content based on spectral characteristics of apple tree leaf in different phenological phases. **Computers and Electronics in Agriculture**, 2015. v. 112, p. 20–27.

2 ARTIGO 1 - Near infrared spectroscopy and element concentration analysis for assessing yerba mate (*Ilex Paraguariensis)* samples according to the country of origin

Artigo publicado no Computers and Electronics in Agriculture (Elsevier, JCR 5yIF: 2.502, Qualis A2)

Abstract

Yerba mate (*Ilex Paraguariensis*) is used to produce a beverage typically consumed in South America countries, and presents peculiar land-based characteristics due to geographical origin. Such characteristics have recently become a matter of interest for many producers as specific features of yerba mate tend to influence product acceptance in new markets, prices and commercial advantages. This scenario justifies the developing of frameworks tailored to correctly classify products according to their authenticity. This paper uses Near Infrared (NIR) spectroscopy and data describing concentration of chemical elements to classify commercial yerba mate samples according to their place of origin. Aimed at enhancing data interpretability, we propose a novel variable selection method that applies quadratic programming to reduce redundant information among the retained variables and maximize their relationship regarding the sample place of origin; sample categorization is then performed using alternative classification techniques. When applied to the NIR dataset, the proposed method retained average 8.79% of the original wavenumbers, while leading to 1.9% more accurate classifications when compared to categorization using the full spectra. As for the elements dataset, we increased average classification accuracy by 3.5% and retained 47.22% of the original elements. The proposed method also outperformed two other approaches for variable selection from the literature. Our findings suggest that variable selection frameworks help to correctly identify the origin and authenticity of yerba mate samples, making model construction and interpretation easier.

Keywords: Yerba Mate, Near infrared (NIR), ICP-MS, ICP-OS, Variable selection

2.1 INTRODUCTION

Food and beverage producers typically associate their brands with the place of origin in order to increase product acceptance in new markets, and to obtain better prices and commercial advantages (DINIZ et al., 2015; KAROUI; BAERDEMAEKER, 2007; LUYKX; RUTH, VAN, 2008). The conditions of planting, harvesting and product processing, as well as intentional adulteration, may alter the final product quality and specifications. Thus, aligned with regulatory authorities, producers have presented an increasing interest in ensuring the precise categorization of products into proper classes according to place of origin, as well as improve methodologies for recognizing product authenticity (BORRÀS et al., 2015; MARCELO et al., 2014). Additionally, with the growth of international trades and potential markets, many countries have relied on several regulations or laws to ensure food traceability (ZHAO et al., 2013).

The *Ilex Paraguariensis*, also known as yerba mate, is a plant typically cultivated in the subtropical regions of South America, and its infusion is one of the most consumed beverages in countries from that continent. Despite being consumed throughout the continent, each region has different tastes and ways of consuming the beverage, forcing the producers to adapt the product to each region (FILIP et al., 2000; LINARES et al., 2010; NUNES et al., 2015). The commercialization of yerba mate has increased in recent years due to the benefits that its consumption brings to health as the substance contains bioactive components such as polyphenols, flavonoids, amino acids, xanthines and alkaloids (GAO et al., 2013; LÓPEZ-CÓRDOBA et al., 2015; MARCELO et al., 2014); these compounds are associated with antioxidant, anticancer, antiallergic, diuretic and hypocholesterolemic properties (BRACESCO et al., 2011; FILIP et al., 2000; LINARES et al., 2010). Although seminal researches on yerba mate were carried out in South America (where its use is mostly widespread), recent findings have also been extended to Japan, China and the USA (BRACESCO et al., 2011; HECK; DE MEJIA, 2007).

In the last few years, near infrared (NIR) spectroscopy has gained wide acceptance in many research fields as a simple, quick and non-destructive technique that allows the identification of chemical compounds from samples without previous preparation (CRAIG et al., 2014; LIU; YANG; DENG, 2015; ZHANG; ZHANG; IQBAL, 2013). NIR has been coupled with multivariate techniques aimed to identify patterns emerging from different products with categorization purposes in several fields, including food (KAROUI; DE BAERDEMAEKER, 2007; MARCELO et al., 2014), pharmaceuticals (ANZANELLO et al., 2013; GENDRIN; ROGGO; COLLET, 2007) and fuels (FERRÃO et al., 2011; VASCONCELOS et al., 2012). More aligned with the propositions of this paper, NIR and multivariate techniques have been used to confirm the authenticity and origin of food and beverage products, including tea (DINIZ et al., 2014), wine (CYNKAR et al., 2010; LIU et al., 2008), cheese (OTTAVIAN et al., 2012; PILLONEL et al., 2003), olive oil (CASALE et al., 2010; GALTIER et al., 2007), honey (WOODCOCK et al., 2007) and persimmon fruits (KHANMOHAMMADI et al., 2014).

Although the NIR technique efficiently provides absorbance values that help to identify and quantify several chemical components, it also generates databases comprised of hundreds (or even thousands) of highly correlated and noisy variables that jeopardize the prediction of a response variable. In this scenario, wavenumber selection techniques become fundamental to enhance the analysis and reduce the influence of such ill-conditioned data upon the multivariate techniques (BALABIN; SMIRNOV, 2011; COZZOLINO; RESTAINO; FASSIO, 2010; DONG et al., 2013; XIAOBO et al., 2010). Selecting the most relevant regions of the NIR spectra also makes it easier to interpret the generated models, once it highlights not only the relationship among wavenumbers but also the relationship of these wavenumbers with the investigated property (XIE; YING; YING, 2009; ZHANG; ZHANG; IQBAL, 2013). Additionally, the removal of uninformative wavenumbers reduces the model complexity and provides better results (CHEN et al., 2013).

Another approach to trace the origin of food products consists of analyzing their elemental composition and chemical concentration (DRIVELOS; GEORGIOU, 2012; Luykx et al., 2008). Multivariate analysis of elements concentration determined by ICP-OES and/or

ICP-MS has been widely used to determine the quality of products such as eggs (BARBOSA et al., 2014a), rice (MAIONE et al., 2016), organic coffee (BARBOSA et al., 2014b), and tea (DINIZ et al., 2015; MOREDA-PIÑEIRO; FISHER; HILL, 2003). Specifically in yerba mate, elements can provide useful insights on different forms of cultivation, type of soil and climatic conditions in each country, since macronutrient and micronutrient availability depends on several features such as rainfall, temperature, pH and type of soil (MARCELO et al., 2014a). The concentration of certain elements may also be altered by the plant age, use of fertilizers, pesticides, fungicides, and soil acidity (HÄNSCH; MENDEL, 2009; LAURSEN et al., 2011; MAATHUIS, 2009). In addition, elements concentration may change as yerba mate is industrially processed, especially in the drying and blanching steps (GIULIAN et al., 2009). In light of that, focusing on the chemical elements with higher discriminant ability becomes a crucial step to identify sample patterns tailored to classification purposes.

Some previous studies also aimed to classify yerba mate samples according to their place of origin: Cozzolino et al. (2010) used NIR full spectra coupled with Principal Components Analysis to perform such classification, while Marcelo et al. (2014a) identified the origin of yerba mate samples based on inductively coupled plasma mass spectrometry (ICP-MS) and inductively coupled plasma optical emission spectrometry (ICP-OES). Finally, Marcelo et al. (2014b) selected regions of the yerba mate NIR spectra using an interval-based approach to classify samples.

This paper proposes a new framework for variable (wavenumbers or elements) selection aimed at categorizing yerba mate samples into classes according to place of origin. For that matter, we use quadratic programming to simultaneously minimize the probability of retaining redundant variables, and to maximize the relationship between such variables and the geographical origin (which is the response variable). Aimed at verifying the quality of the retained subset, different classification techniques are used. We applied our propositions to two datasets emerging from the same yerba mate samples: one consisting of NIR and the other describing elements concentration. Such datasets were obtained from 54 yerba mate brands proceeding from four different countries (Brazil, Paraguay, Argentina and Uruguay). The method using the recommended technique achieved 100% in the training and 94.29% in the testing set in the NIR dataset, while retaining only 3.7% of the original wavenumbers. As for the dataset consisting of elements concentration, 50% of the original variables were retained to perform a perfect classification in both training set and testing set.

2.2 MATERIALS AND METHOD

2.2.1 Sample preparation and Instrumentation

Fifty-four (54) packages of different brands of commercial yerba mate were purchased in local markets of four South America countries: 19 from Brazil, 14 from Paraguay, 14 from Argentina and 7 from Uruguay. The different number of brands available in each country justifies the different number of samples derived from each country. Geographic origin and additional information were available on package labels. In addition, we have carefully screened all brands of yerba mate available in each assessed market, so the number of samples is near the population. It is noteworthy that we restricted our study to the traditional version of yerba mate (also called "native"), excluding commercial variations that present sugar, teas or other substances in their composition. Although such samples would increase the number of assessed samples, the added substances certainly would alter the spectral data, misleading the findings of our propositions. As for Brazilian samples, we focused on brands produced in Rio Grande do Sul and Santa Catarina states due to the vast territory of that country and potential heterogeneity of samples; such information was available on package labels.

The NIR spectra were obtained using a PerkinElmer 400 IR spectrometer equipped with integrating sphere and indium gallium-arsenic (In-Ga-As) detector. Background registration was taken using a Spectralon disc. Reflectance was measured in the range 10000–4200 cm^{-1} with a 4 cm^{-1} resolution. Due to excessive noise, in some NIR regions, the 8000-4200 cm^{-1} spectral region was selected, resulting in 3801 wavenumbers. For each sample, reflectance was normalized (maximum values to 1 and minimum values to 0) in order to correct multiplicative effects in the spectrum and to remove light effects. We performed a total of 32 scans for each sample; all scans were run in random order and in triplicate. One triplicate for a Brazilian sample was discarded; therefore the total number of NIR observations is 161. An aliquot (50 g) of each yerba mate brand was grounded in a cryogenic mill (Spex Certiprep, 6750 Freezer Mill, USA). The result was a homogenous green powder with particle size under 300 µm. This powder was transferred to a glass recipient recommended by the equipment manual (PerkinElmer 400 IR spectrometer). The glass recipient was put above the equipment sampler and covered with a black plastic cape to protect from external lights. Samples were not attached to the integrating sphere. Preliminary tests revealed that particle size had huge importance upon the analysis, so the cryogenic grinding was carried out in argon atmosphere with the sample frozen for 2 min and then ground for 2 min at 20 beats per second.

To determine the elements concentration a Varian/Vista MPX (USA) ICP-OES and ELAN DRC II (Perkin Elmer/SCIEX, Canada) ICP-MS were employed for quantification of the investigated elements. The ICP-MS instrument was operated in standard mode. Information about the elements, main instrumental parameters and settings are summarized in Table 2-1.

Additionally, we used nitric acid (Merck) purified by sub-boiling distillation (Duopor/Milestone, Italy) for sample and solution preparation. In order to obtain a resistivity of 18.2 MΩ cm the water used throughout the study was purified in a Milli-Q system (Millipore). The calibration solutions were prepared in 5% (v/v) HNO_3 by serial dilution of stock solutions as follows: a) 10 mgL^{-1} (Plasma Cal SCP33MS Science, Canada) for Li, Be, Ti, V, Cr, Ni, Co, As, Se, Rb, Mo, Ag, Cd, Sb, La, Ce, Pb, Bi and U; b) 1000 mgL^{-1} (SpecSol, Brazil) for Al, Ba, Ca, Cu, Fe, K, Mg, Mn, Sr and Zn; and c) 1000 mgL^{-1} (Titrisol, Merck) for P. The concentration of calibration solutions ranged from 0.05 to 10 μgL^{-1} for the elements determined by ICP-MS. As for the elements determined through ICP-OES, the concentrations of calibration solutions were 10 to 100 μgL^{-1} for Sr, Zn, Ba and Cu; 50 to 1500 μgL^{-1} for Fe, Al and Mn; and 1.0 to 8.0 mgL^{-1} for K, Ca, Mg and P. Sb, Se, Ag, Bi, Li and Be were not detected in some samples, yielding a dataset comprised of 24 variables.

Parameter	Setting for each technique	
	ICP-OES	ICP-MS
RF power	1300W	1300W
Plasma gas flow rate	15 L min^{-1}	15 L min^{-1}
Auxiliary gas	2.25 L min^{-1}	1.2 L min^{-1}
Nebulizer gas flow rate	230 kPa	1.0 L min^{-1}
Sample flow rate	1.5-2.5 mL min^{-1}	1.2 mL min^{-1}
Nebulizer	Ultrasonic (CETAC, 5000) and concentric	MicroMist MCN-600
Spray Chamber	Sturman-Masters (VARIAN)	Cyclonic
Wavenumber (nm)	Al (396.153), Ba (233.527), Ca (422.673), Cu (324.754), Fe (238.204), K (769.897), Mg (279.553), Mn (257.610), P (213.617), Sr (407.771), Zn (213.857)	-
Isotope	-	7Li, 9Be, 47Ti, 51V, 53Cr, 58Ni, 59Co, 75As, 82Se, 85Rb, 98Mo, 107Ag, 112Cd, 121Sb, 139La, 140Ce, 208Pb, 209Bi, 238U
Plasma view	Radial	-
Calibration type	External	External

Table 2-1 – Parameters and settings for elemental analysis using ICP-OES and ICP-MS

2.2.2 Multivariate Techniques

We now present the fundamentals of the multivariate techniques used in this paper (Mutual Information, Quadratic Programming, k-nearest neighbor, Support Vector Machine and Discriminant Analysis). All these techniques are widely available in computational packages for data analysis, justifying their use in our propositions.

The Mutual Information of two random variables (e.g. wavenumbers or elements concentration) is a non-negative symmetric dependency measure between these two variables (i.e., the Mutual Information quantifies the amount of information shared between such variables). Unlike other dependency measures, Mutual Information has the advantage of modeling nonlinear dependences (ROSSI et al., 2006). The Mutual Information between variables X and Y is defined as

$$\mathrm{I}\,(X;Y) = \sum_{y\epsilon Y}\sum_{x\epsilon\, X}\mathrm{p}(x,y)\log(\frac{\mathrm{p}(x,y)}{\mathrm{p}(x)\mathrm{p}(y)})(1)$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively (LONG, et al., 2013; RACHOW et al., 2011; RODRÍGUEZ-ROSARIO et al., 2008). Such distributions can be either discretized (also known as “Histogram Method”, approach used in this paper) or estimated by density function methods (DUDA; HART; STORK, 2001).

Quadratic programming (QP) is a type of optimization applied to a quadratic function consisting of several variables subjected to linear constraints (CHEN; CHEN; LIN, 2005; GILL; WONG, 2015). An optimization problem seeks to determine the function domain that reaches the extreme values of a function, i.e. the largest or the smallest value that a function can achieve. The generic vector formulation of a QP is depicted in equation (2)

$$\underset{x\,\in\,R^N}{minimize}\; f(\mathbf{x}) = \{\frac{1}{2}\mathbf{x}^{\mathrm{t}}\mathbf{H}\mathbf{x} - \mathbf{f}^{\mathrm{t}}\mathbf{x}\,\} \qquad (2)$$

where **x** is a *N*-dimensional variable vector, **H** is a *N*x*N* symmetric matrix that represents the second order elements of the polynomial, and **f** is a *N*-dimensional vector that represents the first order elements (RODRIGUEZ-LUJAN et al., 2010). In the propositions of this paper, the QP aims at identifying the most relevant variables for sample classification.

We now present the fundamental of the classification techniques we test. The first, k-Nearest Neighbor (KNN), is a non-parametric technique that categorizes a new sample according to the class that appears the most among the *k* nearest samples; the nearest neighbors are typically identified by means of the Euclidian or the Mahalanobis Distances. The KNN stands out for its simplicity and for requiring a single parameter, *k*, which can be defined by cross-validation (BARBON et al., 2016; DUDA; HART; STORK, 2001; REBOLO et al., 2000). Discriminant Analysis (DA), the second used technique, determines the hyperplane that maximizes the ratio of variances between classes and within each class. Similarly to several other statistical techniques, such hyperplane can be achieved through the calculation of eigenvalues and correspondent eigenvectors. The resulting hyperplane is then used to classify new samples (DUDA; HART; STORK, 2001; HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The third technique is the Support Vector Machine (SVM), which uses a training sample to create a linear hyperplane maximizing the separation between two classes (CORTES; VAPNIK, 1995); the location of a sample with regards to that hyperplane defines its allocation to a class. In addition, kernel functions can be applied to transform nonlinear data and create a proper hyperplane (COLMAN et al., 2015; HUANG, C. L.; WANG, C. J., 2006; RAKOTOMAMONJY, 2003). SVM was originally created to classify samples into two classes, but different approaches (e.g. one versus the remaining classes, used in this paper) allows one to classify new samples into multiple classes (BURGES, 1998; LUTS et al., 2010).

2.2.3 Proposed framework for variable selection

There are four methodological steps in the proposed framework for variable selection aimed to identify the origin of yerba mate samples: (i) divide the original data into training and testing sets using the Kennard Stone algorithm; (ii) compute the mutual information among variables and between variables and the response variable using the training set; (iii) create a Variable Importance Index (VII) to assess variable relevance using the QP optimization; and (iv) iteratively classify the training set samples using a forward procedure according to the order suggested by the VII, and retain the subset that yields the best result. Finally, classify the testing set using the retained variables. We now detail such steps.

Consider samples represented by *C* classes, *c*=1,…,*C*. In step (i), for each class *c*, we split the M_c samples consisting of *N* variables into two sets: a training set from class *c*, containing TR_c samples, and a testing set from class *c*, containing TS_c samples, where TR_c+TS_c=M_c. The Kennard-Stone algorithm was used to split the dataset into training and testing set due to its robustness and wide acceptance in chemometrics (DONG et al., 2013; KHANMOHAMMADI; GARMARUDI; LA GUARDIA, 2013; PONTES et al., 2005). The proportion between TR_c and TS_c is defined as 80%-20%. The union of all *c* training sets is used to select the most important variables, while the union of all *c* testing sets denotes new samples to be classified. Such unions will hereafter be called just as training set and testing set, respectively.

When modeling a dataset, one typically intends to retain variables that have minimum relationship among themselves and maximum information regarding the response variable (GUYON; ELISSEEFF, 2003; LIU; YU, 2005). In order to tackle such points, in step (ii) we use the training set to generate a matrix **R** and a vector **s**. Since each class may be differently described by sets of variables, a matrix $\mathbf{R}_c$ with dimensionality *N*x*N*, where r_{cij}=*I*(V_{ci}, V_{cj}), i.e., the mutual information of variables *i* and *j* within class c, is calculated for each class *c*. In that notation, in the R matrix of class *c*, *i*th line and *j*th column is the mutual information between variable *i* and variable *j* within class *c*. Matrix **R** is then defined as the average Mutual Information for each variable among classes, as in equation (3), while **s** is a *N*-dimensional vector with elements s_i=*I*(V_i, $origin$); i.e., the *i*th element of vector **s** is the mutual information between variable *i* and the response variable (place of origin).

$$\mathbf{R} = \sum_{\mathrm{c}=1}^{\text{number of classes}} \frac{\mathbf{R}_{\mathrm{c}}}{\text{number of classes}} \qquad (3)$$

In step (iii), a QP optimization problem is formulated using **R** and **s**. The objective function aims to simultaneously reduce the shared information and increase the categorization ability among variables, as in equation (4); in that equation, α is a scalar parameter proposed in Rodriguez-Lujan et al. (2010) that prevents the predominance of **R** or **s** in the optimization (i.e., α avoids one term of the optimization to dominate the other term, which may lead to loss of important information, as claimed by Rodriguez-Lujan et al. (2010). Constraints presented in equations (5) and (6) restrict the domain function to the interval [0,1].

$$\underset{x \in R^N}{minimize}\, f(\mathbf{x}) = \{\tfrac{1}{2}(1-\alpha)\mathbf{x}^{\mathrm{t}}\mathbf{R}\mathbf{x} - \alpha\mathbf{s}^{\mathrm{t}}\mathbf{x}\,\} \qquad (4)$$

Subject to

$$\mathrm{x_i} \geq 0 \qquad (5)$$

$$\textstyle\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{x_i} = 1 \qquad (6)$$

In the propositions of this paper, the weight vector **x** (consisting of variables x_i) in equation (4) represents the weight of variable *i* in a scenario where the association among independent variables is minimized and the association between independent variables and classes is maximized by means of the QP (RODRIGUEZ-LUJAN et al., 2010). Such weight enables assessing the importance of each variable for the classification procedure carried out in step (iv). Therefore, **x** is used as a VII; the greater the VII, the more important such variable is deemed for classification.

In step (iv), the training samples are inserted into proper classes applying the KNN on the variable with the highest VII; the classification accuracy (i.e., ratio of correct classifications) is calculated. Next, the variable with the second highest VII is added to the dataset and a new classification using the two most important variables is carried out. Such iterative procedure is repeated until all variables have been inserted into the dataset used for classification. The variable subset yielding the highest accuracy is retained; in case multiple variable subsets yield the maximum accuracy, the subset with fewer variables is retained. That course of action allows one to interrupt step (iv) whether perfect classifications are obtained before all variables are inserted into the dataset (avoiding unnecessary computational processing). The selected variables are then used to classify the testing set samples, denoting observations not used during the development of the classifier. Finally, we repeat Step (iv) replacing KNN by DA and SVM classification techniques and compared their categorization performance.

2.3 RESULTS AND DISCUSSION

We now apply the proposed method to the datasets (yerba mate NIR and chemical elements) described in section 2.1; the classification performance obtained after variable selection is then compared with the categorization using all original variables (i.e., full databases) and with two frameworks for variable selection. All computational experiments were performed in Matlab® R2014a, using Statistics and Machine Learning Toolbox and Optimization Toolbox.

Parameters for the classification techniques were defined using cross validation (BARBON et al., 2016; ZHANG et al., 2015). For the KNN classification technique, we defined the parameters "number of neighbors" (assessed from k=3 to k=5), and the "distance metric" (Euclidean or Mahalanobis). As for the DA, we tested the parameter "discriminant type" (linear or quadratic); we also set the prior probability for each class as "uniform", since there were no evidences to believe that a sample had different probabilities to belong to a specific class. Finally, SVM parameters were assessed in terms of "kernel function" (linear, polynomial with order 2 or 3, and Radial Basis Function with sigma equal to 1, 3, or 5), and "box constraint" representing the misclassification cost (1, 5 or 10). If more than one combination of parameters depicted the same result, the one closest to the default values was used.

2.3.1 NIR data

2.3.1.1 Overview of the NIR spectra for yerba mate

The 8000–4200 cm^{-1} spectral region was selected due to excessive noise observed in other NIR regions. The infrared spectra were normalized using sup norm (maximum absorbance equal to 1) to remove systematic variation associated to the particle size, and to equalize the magnitude of each sample in the model. Figure 2-1 displays the normalized spectra for the 161 samples.

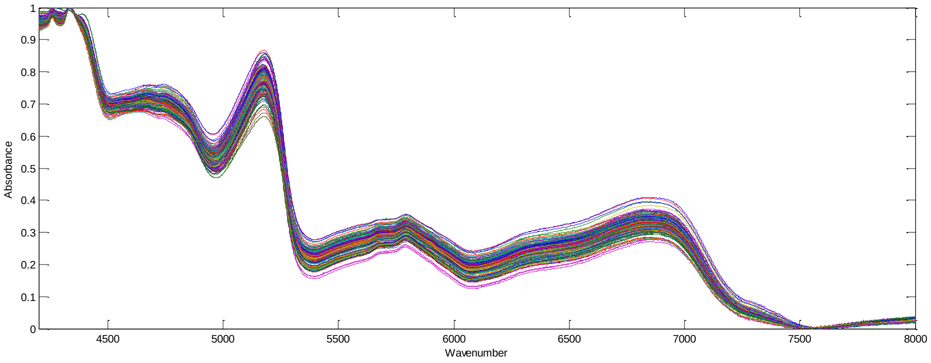


Figure 2-1 Raw spectra of NIR absorbance for 161 yerba mate samples

2.3.1.2 NIR wavenumber selection

Table 2-2 depicts the selected parameters of the classification techniques for the NIR data. As for the KNN parameters, wavenumbers do not require the Mahalanobis Distance, since the Euclidean Distance (the default parameter) produces similar results; in addition, *k*=3 is set to avoid overfitting as the number of Uruguayan samples is substantially smaller than other countries' samples and may jeopardize the classification. DA and SVM depict different results, as DA indicates that the classes have linear boundaries while SVM rely on a nonlinear classifier.

KNN	DA	SVM
k=3 (default) *Euclidean Distance (default)*	*Linear Discriminant (default)*	*Radial Basis Function* *Sigma = 3* *Box constraint = 5*

Table 2-2 – Parameters of multivariate techniques on NIR dataset

Table 2-3 depicts the confusion matrix for each classification technique applied to the yerba mate training set using the recommended subset of wavenumbers. The training set is comprised of 126 samples: 32 from Argentina, 45 from Brazil, 15 from Uruguay and 33 from Paraguay. DA and SVM led to perfect classifications, while KNN yielded several misclassifications. Table 2-4 depicts the number of retained wavenumbers for each classification technique after the selection procedure suggested in section 2. Regardless the classification accuracy in the training set, it is noteworthy that SVM and DA needed substantially fewer wavenumbers than the KNN. The joint analysis of Table 2-3 and Table 2-4 suggests that the variables are highly overlapped and jeopardize the KNN performance as this last classification such technique relies on the Euclidean distance for categorization.

Training Set													
		Real											
		KNN				DA				SVM			
		ARG	BRA	URU	PAR	ARG	BRA	URU	PAR	ARG	BRA	URU	PAR
Predicted	ARG	31	4	3	1	33	0	0	0	33	0	0	0
	BRA	1	40	3	2	0	45	0	0	0	45	0	0
	URU	0	0	9	0	0	0	15	0	0	0	15	0
	PAR	1	1	0	30	0	0	0	33	0	0	0	33

Table 2-3 – Confusion matrix of classification results on NIR training set

Classification Technique	Number of Retained Wavenumbers	% of Retained Wavenumbers
KNN	759	19.97%
DA	102	2.68%
SVM	141	3.71%

Table 2-4 – Number and ID of retained wavenumbers

Figure 2-1 illustrates the initial stages of the classification accuracy profile as the first wavenumbers are inserted into the training set following the order suggested by the VII (SVM is used as classification technique); the dots in that figure represent the classification accuracy when an increasing number of wavenumbers is used. After the addition of the 141th wavenumber the accuracy stands on 100%, suggesting that the insertion of additional wavenumbers into the model does not provide relevant information for the classification.

In order to illustrate the better classification performance yielded by the proposed VII in the wavenumber selection process, we carried out an alternative selection procedure by randomly adding wavenumbers (one by one) to the subset of wavenumbers used for classification until 100% accurate classifications were obtained or all wavenumbers were added to the subset. We repeated that procedure 100 times. The results from such alternative course of action are represented by the bars in Figure 2-1, which depict the mean accuracy ± 1 standard deviation (numerical results of such alternative procedure are presented in section 3.3, which brings a comparison between different approaches). It can be noticed that the accuracy profile yielded by the random selection of wavenumbers leads to smaller accuracies (especially after the method achieved 100% of accuracy), suggesting the proposed method as a reliable way of selecting the most relevant wavenumbers for sample classification.

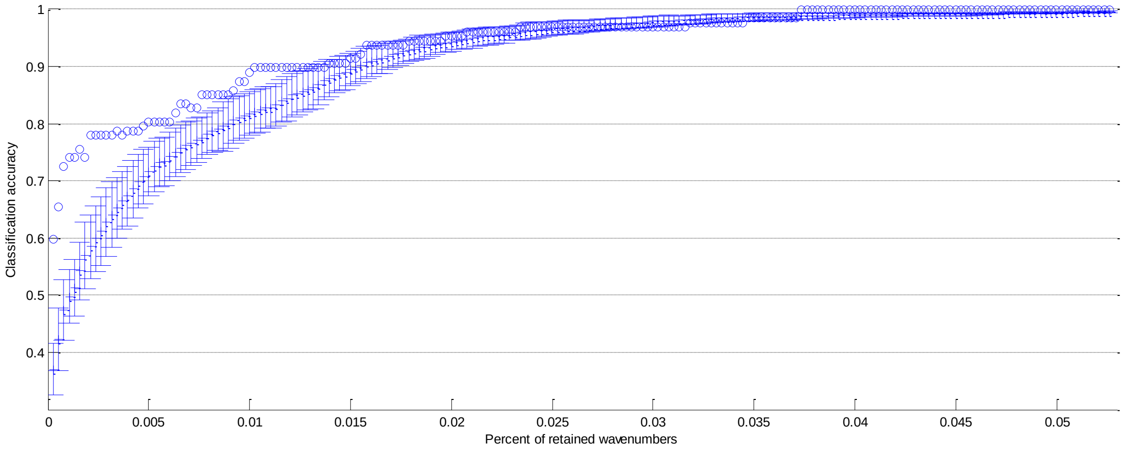


Figure 2-1 – NIR training set accuracy profile with SVM

Table 2-5 depicts the confusion matrix comparing the classification performance of both full spectra and selected wavenumbers in the testing set. The testing set is comprised of 35 samples: 9 Argentinean, 11 Brazilian, 6 Uruguayan and 9 Paraguayan. It can be noticed that only SVM provides satisfactory results, while both KNN and DA have their classification performance substantially reduced in the testing set. The poor results achieved by KNN were expected, once the training set results were poor; as for the DA low accuracy in the testing set, there are strong evidences of model overfitting yielded by this classification technique. Such results indicate that linear techniques are not adequate to create a good classification model, therefore a high dimensionality transformation (such as RBF) is needed.

Testing Set														
			Real											
			KNN				DA				SVM			
			ARG	BRA	URU	PAR	ARG	BRA	URU	PAR	ARG	BRA	URU	PAR
Predicted	Full data	ARG	8	1	0	2	8	0	0	2	6	0	0	0
		BRA	1	9	0	0	0	6	0	3	1	9	0	0
		URU	0	0	6	0	0	2	1	2	0	0	2	1
		PAR	0	1	0	7	1	3	5	2	2	2	4	8
	Selected Wavenumbers	ARG	8	1	0	2	2	0	1	2	9	1	0	1
		BRA	1	7	0	2	1	5	1	2	0	10	0	0
		URU	0	0	6	0	5	5	4	1	0	0	6	0
		PAR	0	3	0	5	1	1	0	4	0	0	0	8

Table 2-5 – Confusion matrix of classification results on NIR testing set

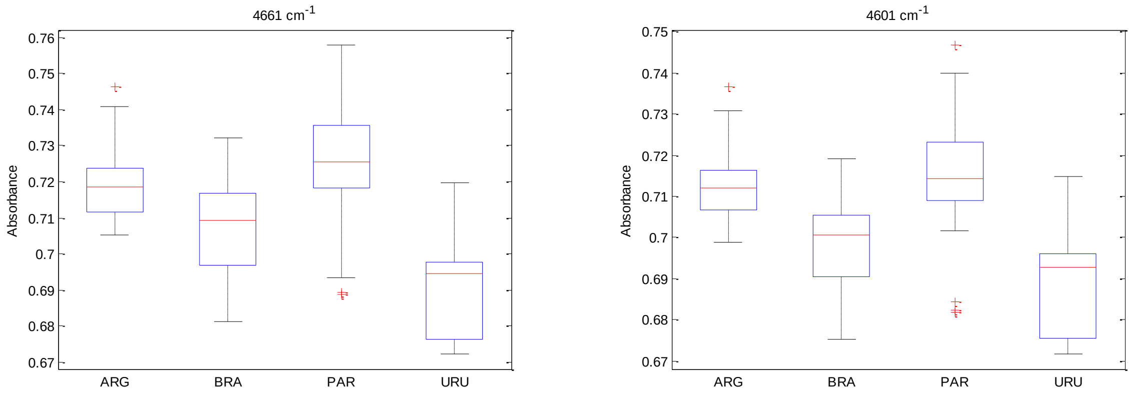
2.3.1.3 Analysis of the retained wavenumbers

We now discuss on the wavenumbers retained by the proposed method using the SVM; we recommend that classification technique as it yields the best overall results. As the retained subset contains very close (and presumably redundant) wavenumbers, it may suggest that the retained subset is not optimum. Although that may be the case of the proposed method, such condition does not testify against the categorization ability of the subset recommended by the method. According to Guyon and Elisseeff (2003), the addition of redundant variables in subsets used for classification may, in some scenarios, reduce the noise of the whole retained subset and conduct to better categorizations. Such authors also state that such scheme is likely to happen when approaches based on variable ranking are used; although producing a suboptimum subset, it should not be considered a harmful characteristic.

In order to better analyze the retained wavenumbers, they are divided into two groups: group (*i*), containing wavenumbers in the 4530 to 4684 cm^{-1} interval; and group (*ii*), containing wavenumbers in the 4264 and 4381 cm^{-1} interval. The wavenumbers inside the groups are homogeneous, and have the same chemical interpretation, while the groups are heterogeneous. To visually describe these groups a box plot of three wavelengths of each group is presented in Figure 2-2.

Wavenumbers inserted into Group (*i*) are closely related to combinations of carbon-carbon bonds, aldehydes and amine functional groups. Figure 2-2 suggests a large overlapping between Argentinean and Paraguayan samples and Brazilian and Uruguayan samples, but a reasonable separation between these two overlapped groups. Wavenumbers belonging to Group (*ii*) are related to combinations of the CH_3, CH_2 and CH bonds. Figure 2-2 indicates that the main contribution to this group to classification is the notable separation between the Uruguayan samples to other countries samples that wavenumbers inside this interval provides. Additionally, the notable overlap among different categories samples depicted in Figure 2-2 enhances the necessity of a nonlinear classifier.

The differences among samples highlighted by the retained variables may not be only related to soil and weather conditions, but also to processing stages. Different methods of drying yerba mate (or *sapecado* stage) can lead to different moisture migrations along the heat process, as well as changes in the organic composition of the final product. The drying step can also lead to different organoleptic characteristics, which can be used as a commercial advantage for trading (GIULIAN et al., 2009; SCHMALKO; LOVERA; KOLOMIEJEC, 2011).



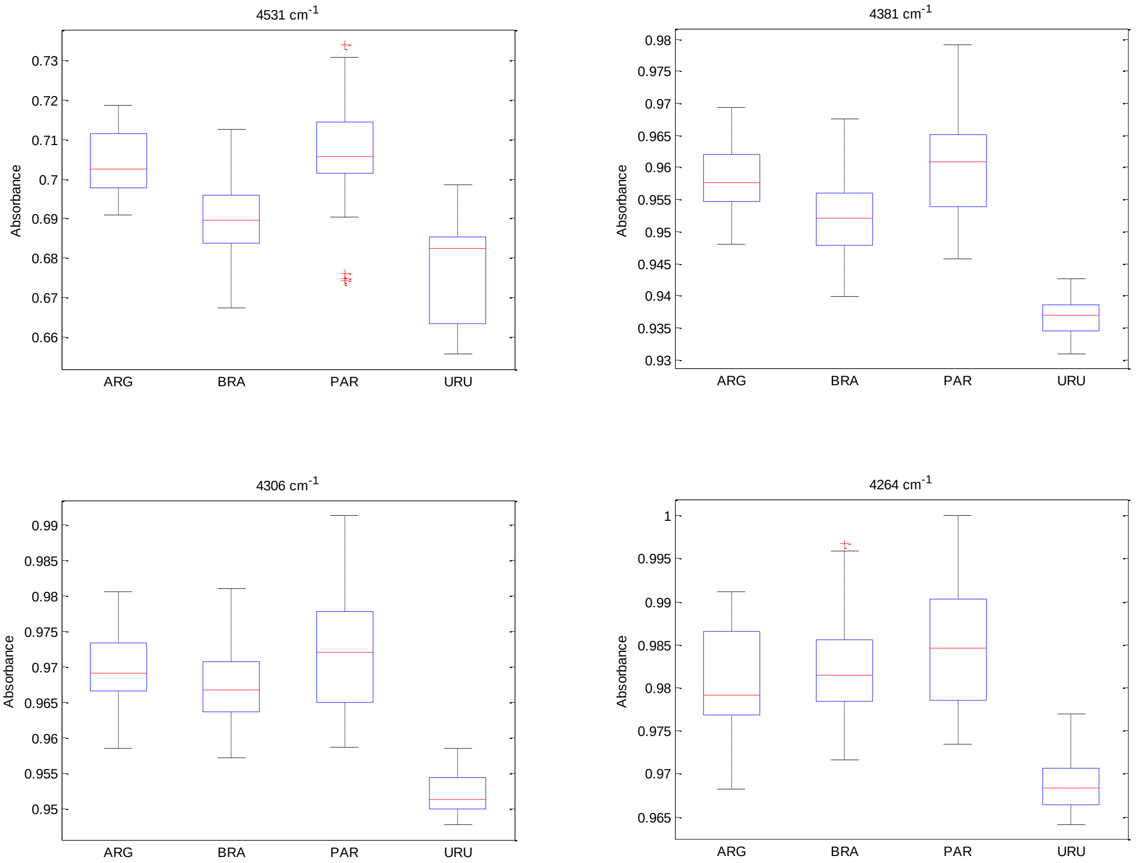


Figure 2-2 – Boxplot of NIR retained wavenumbers

2.3.2 Elements data

2.3.2.1 Elements selection

Table 2-6 depicts the parameters of the classification techniques for the elements dataset. Most parameters are the same used in the NIR dataset, with the exception of the kernel function RBF with sigma=3 and “box constraint”=10 required by the SVM. Such modification suggests the existence of a nonlinear boundary among the classes.

KNN	DA	SVM
k=3 (default) Euclidean Distance (default)	Linear Discriminant (default)	RBF kernel function Sigma = 3 Box constraint = 10

Table 2-6 – Parameters of multivariate techniques on elements dataset

Table 2-7 depicts the confusion matrix of the classification on the training set using the recommend subset of variables. KNN was unable to correctly classify all samples, misclassifying two samples (one Brazilian and one Uruguayan sample into the Argentinean class). On the other hand, both DA and SVM correctly classified all training set samples.

		Training Set											
		Real											
		KNN				DA				SVM			
		ARG	BRA	URU	PAR	ARG	BRA	URU	PAR	ARG	BRA	URU	PAR
Predicted	ARG	11	1	1	0	11	0	0	0	11	0	0	0
	BRA	0	14	0	0	0	15	0	0	0	15	0	0
	URU	0	0	4	0	0	0	5	0	0	0	5	0
	PAR	0	0	0	11	0	0	0	11	0	0	0	11

Table 2-7 – Confusion matrix of classification results on training set of the elements dataset

Table 2-8 depicts the variables (i.e., elements) retained by each classification technique. SVM and DA retained practically the same elements, with exception of Al for the DA; such elements yielded perfect classifications using both techniques. KNN relied on fewer variables, but it did not correctly classify all training set samples; such results suggested that elements V, Ni and La were relevant for sample classification and should have remained in the classifier.

Classification Technique	Number of Retained Variables	% of Retained Variables	Retained Elements
KNN	9	37.50%	Cu, K, As, Zn, P, Cd, Co, Mn, Mg
DA	13	54.16%	Cu, K, As, Zn, P, Cd, Co, Mn, Mg, V, Ni, La, Al
SVM	12	50.00%	Cu, K, As, Zn, P, Cd, Co, Mn, Mg, V, Ni, La

Table 2-8 – Retained elements

Figure 2-3 illustrates the classification accuracy profile with SVM as variables were added into the training set. We decided to recommend SVM as classification technique since it led to perfect classifications with the smallest number of retained variables. Figure 2-3 displays that the accuracy stood on 100% after inserting the 12th variable (50% of full dataset), implying that the addition of extra variables did not improve classification performance. Results depicted in Table 2-9 suggest that the full dataset presents lack of generalization, making of variable selection an important step to correctly determine the origin of new samples.

As similarly performed for the NIR, we also added variables in a random way to the subset of elements used for sample classification of the training set aimed at comparing results with the proposed VII-based method. Differently from the previous dataset, a visual assessment of the accuracy profile does not suggest the proposed method as superior when compared to a random insertion of variables (see Figure 2-3). This can be justified by the fact that the Elements dataset is comprised of a substantially smaller number of variables than the NIR, so some of the replications relying on random insertions of variables may have led to optimum (or close to optimum) results. However, despite the similar results on the training set (and graph below), the performance of the suggested method led to 100% precise classification in the testing set, against average 84.16% yielded by the random procedure (results discussed in section 3.3).

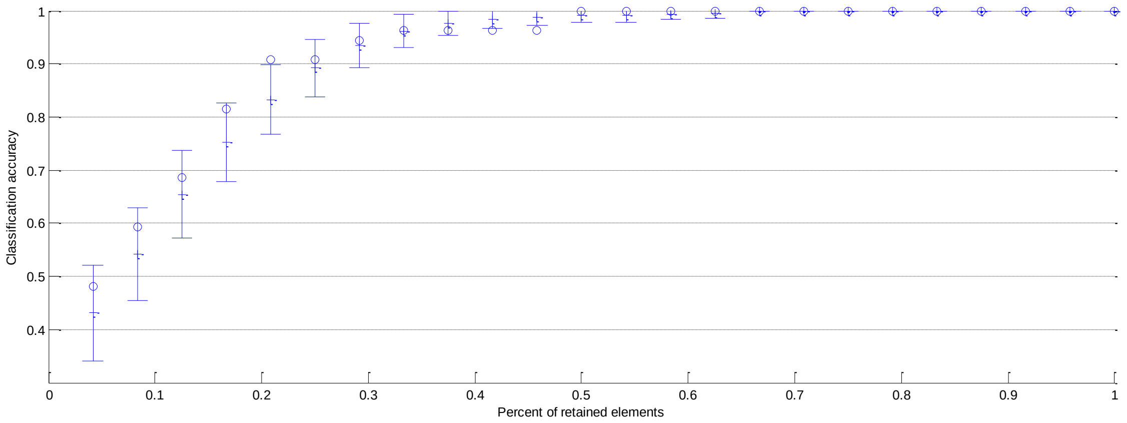


Figure 2-3 – Training set accuracy profile with SVM for the elements dataset

Table 2-9 depicts the confusion matrix for the testing set considering full data (all variables) and data consisting of the selected elements for the testing set. DA correctly classified all samples in both scenarios (i.e., full data and with selected elements). As for the SVM, it improved its categorization performance after variable selection as it led to a perfect classification for the Argentinean samples. Regardless of variable selection, KNN misclassified a Uruguayan sample as Argentinean. The poorer performance of KNN can be justified by the small number of Uruguayan samples available for analysis, and by similarities on the elements found in Argentinean and Uruguayan samples.

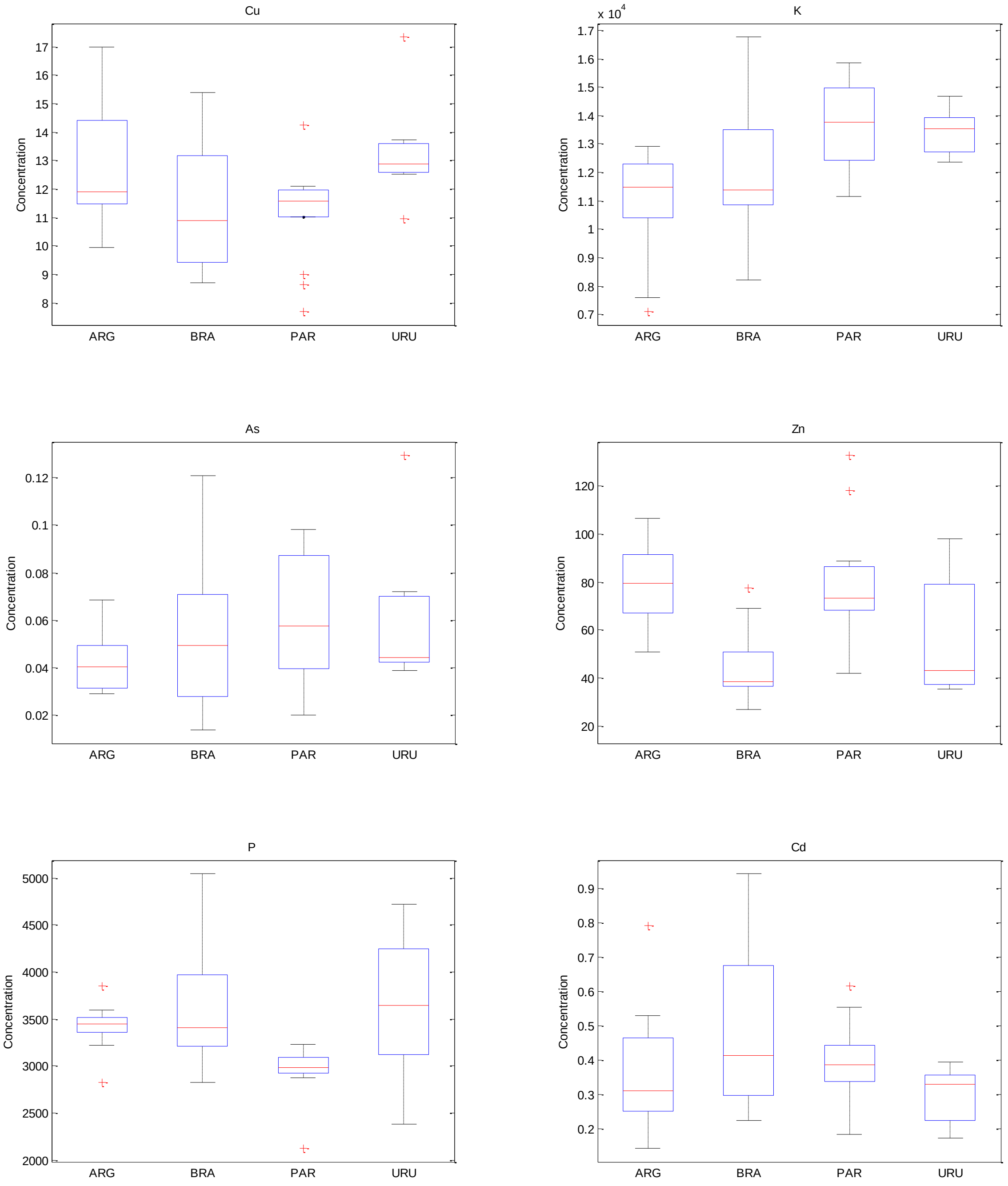
Testing Set														
			Real											
			KNN				DA				SVM			
			ARG	BRA	URU	PAR	ARG	BRA	URU	PAR	ARG	BRA	URU	PAR
Predicted	Full data	ARG	3	0	1	0	3	0	0	0	2	0	0	0
		BRA	0	4	0	0	0	4	0	0	0	4	0	0
		URU	0	0	1	0	0	0	2	0	1	0	2	0
		PAR	0	0	0	3	0	0	0	3	0	0	0	3
	Selected Elements	ARG	3	0	1	0	3	0	0	0	3	0	0	0
		BRA	0	4	0	0	0	4	0	0	0	4	0	0
		URU	0	0	1	0	0	0	2	0	0	0	2	0
		PAR	0	0	0	3	0	0	0	3	0	0	0	3

Table 2-9 – Confusion matrix of classification results on Elements testing dataset

2.3.2.2 Analysis of the retained elements

In light that the SVM classification technique leads to the best results for the elements data, we assess the boxplots yielded by the retained elements in Figure 2-4 (concentrations of such elements are given in μg L^{-1}). Differently from the selected NIR wavenumbers, the elements concentration present substantial overlapping regions, justifying the necessity of kernel transformation for the SVM, and the incapacity of KNN to correctly classify all samples. It is also noteworthy the large number of potential outliers; e.g., 4 samples from Paraguay (PAR) regarding Cu, and three for Argentina (ARG) in element V. Such outliers also contribute to justify the poor results emerging from the KNN, since this technique is more sensitive to outliers than DA and SVM.

Among the selected elements, there are macronutrients (K and P), micronutrients (Cu, Zn, Mn and Ni) and beneficial elements (Co, V). The concentration of these elements is related to soil contents and farming practices (e.g., fertilizers and limestones used for adjusting soil pH). In addition, elements contents in the yerba mate can be modified by industrial processing stages, as previously mentioned for the NIR data (GIULIAN et al., 2009). Among the retained elements, we understand that Lanthanum is a relevant discriminating element that can be related to the soil composition (and consequent place of origin): Uruguay and Brazil have similarly low La levels, whereas Argentina and Paraguay present high concentrations regarding that element. In addition, Cd and As (toxic elements for human and plants) were selected as variables for discriminating some Brazilian samples; such samples should certainly require further assessment given that dangerous aspect.



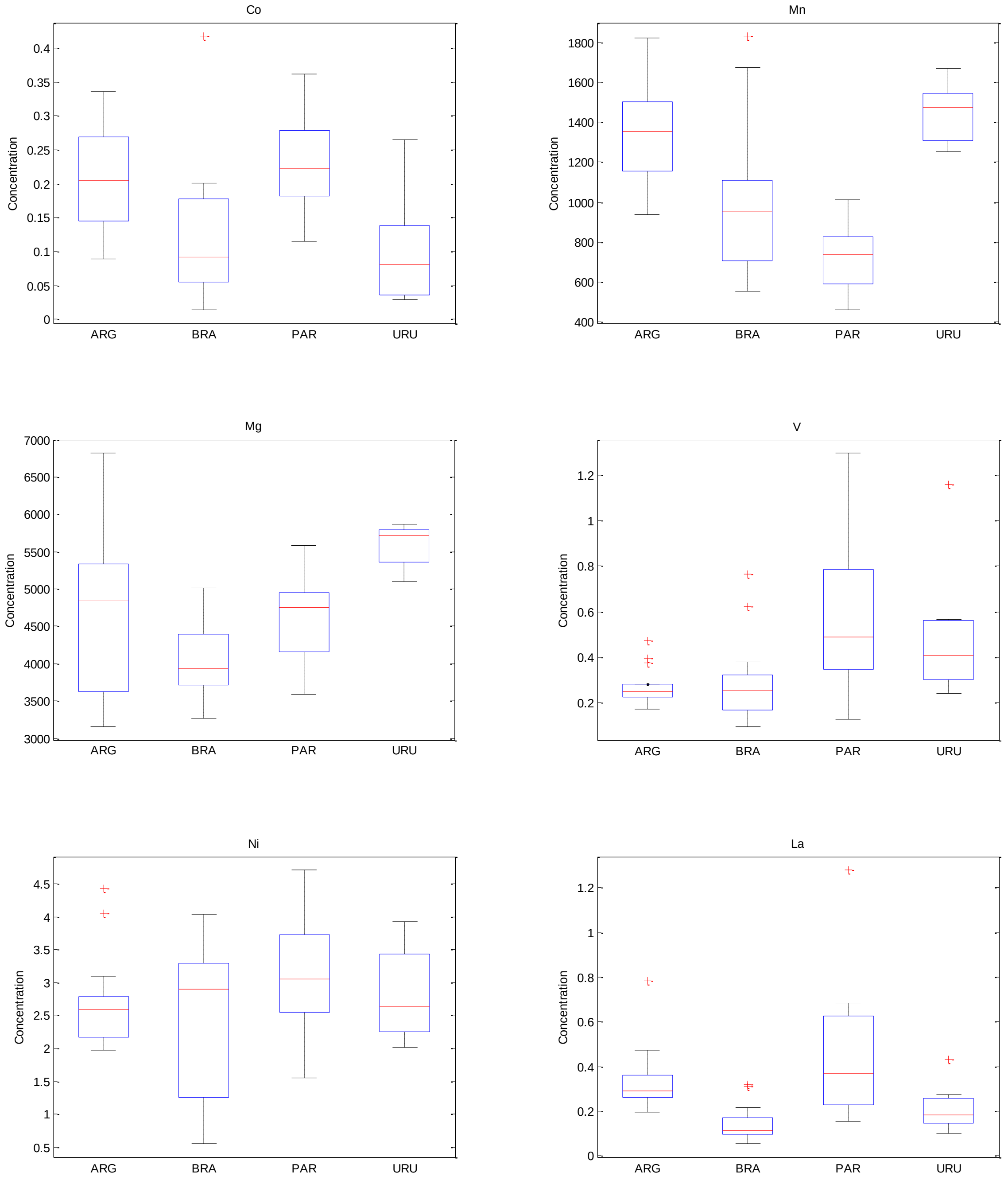


Figure 2-4 – Boxplot of retained elements

2.3.3 Comparison with other variable selection methods

We now compare the proposed method using SVM as classification technique with two other approaches for variable selection: (i) a forward selection that relies on a variable importance index derived from Principal Component Analysis (PCA) parameters; and (ii) the performance of a random forward selection, as described in section 3.1.2. The fundamentals of such methods are based on a wrapper approach and forward selection (i.e., principles aligned with the ones from our propositions), and cited by the literature as well suited courses of action to find the most relevant variables for classification techniques (JIANG; LI, 2015; LIU; YU, 2005). The SVM parameters are the same presented on Table 2-2 and Table 2-6. We now briefly describe the fundamentals of each method.

The first method we compare our propositions relies on the propositions of (ANZANELLO et al., 2013), and will be referred as PCAVII. Such framework builds a VII based on the eigenvalues and eigenvectors from PCA assuming that variables with higher variability are more relevant for sample classification. In the propositions of this paper, we used the Scree Graph approach to define the number of principal components to be retained (RENCHER, 2002); 3 principal components were retained to generate the VII. That VII guides a forward-based procedure similar to the one employed in the proposed method, in which variables are added one by one to the recommended subset and classification accuracy assessed after each addition. In the other tested method variables are randomly included one by one in the subset used for classification and accuracy is evaluated; this method also does not require any parameter to be trained. All aforementioned methods were carried out in Matlab® R2014a using the same Toolboxes employed in the proposed method.

Table 2-10 depicts the average classification accuracy and percent of retained variables of the proposed method compared to the aforementioned variable selection methods. The proposed method outperforms the PCAVII and Random Selection methods as it provides better classification accuracy in both NIR and Elements datasets. Regarding the PCAVII, it clearly overfits the model once the retained variables can correctly classify all training samples, but it performs poorly on samples from the testing set. Those results can be justified by the fact that the PCAVII only considers the variance of the variables, regardless of their discrimination capacity. On the other hand, the proposed method combines these two features.

As for the Random Selection, it also achieves perfect classifications in the training set for all repetitions, but it suffers from an expected lack of generalization. It is also noteworthy that the Random Selection approach retains more variables than the proposed method for the NIR dataset, while in the Elements dataset it retains fewer variables (although yielding unsatisfactory accuracies in the testing set).

Dataset	Classification Performance	Proposed Method	PCAVII	Random Selection
NIR	Accuracy on Training set	1	1	1
	Accuracy on Testing set	0.9428	0.4000	0.8827 (0.048)
	% of retained wavenumbers	3.71%	6.60%	5.58%
Elements	Accuracy on Training set	1	1	1
	Accuracy on Testing set	1	0.9167	0.8416 (0.105)
	% of retained Variables	50%	41.66%	39.84%

Table 2-10 – Performance of the proposed method compared to other variable selection approaches (standard deviation in parenthesis)

2.4 Conclusions

The novelty of this paper relied on the proposition of a novel wavenumber selection method to provide a reduced, easier to interpret model to classify yerba mate samples according to their place of origin. For that matter, we employed both NIR spectroscopy and elements concentration data describing 54 samples from four South America countries. Three classification techniques were tested: k-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Discriminant Analysis (DA). Aimed at better assessing the performance and limitations of our propositions, we used triplicates for the NIR spectra. Although such course of action has been consistently employed as a valid way to increase the dataset size and enable subsequent analyses (ANZANELLO et al., 2013), it represents a limitation on the scope range of this manuscript. By retaining 3.7% (141 wavenumbers) and 50% (12 elements) of the original wavenumbers and elements variables, respectively, SVM was able to correctly classify all samples from the training set in both datasets; as for the testing set, it correctly classified 94.29% in the NIR data and 100% in Elements data. When compared to other variable selection techniques, our propositions proved to be more robust by providing better classification in both datasets. Future research includes the development of different similarity measures to integrate to quadratic programming for identifying the most relevant variables for yerba mate sample categorization. We also intend to apply the proposed method to NIR and data describing concentration of chemical elements of substances other than yerba mate.

2.5 REFERENCES

ANZANELLO, M.J., ORTIZ, R.S., LIMBERGERB, R.P., MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, 2013. v. 83, p. 209–214.

BALABIN, R. M.; SMIRNOV, S. V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. **Analytica Chimica Acta**, 2011. v. 692, n. 1, p. 63–72.

BARBON, A.P.A.C., BARBON, S., MANTOVANI, R.G., FUZYI, E.M., PERES, L.M., BRIDI, A.M. Storage time prediction of pork by Computational Intelligence. **Computers and Electronics in Agriculture**, 2016. v. 127, p. 368–375.

BARBOSA, R.M., BATISTA, B.L., VARRIQUE, R.M., COELHO, V.A., CAMPIGLIA, A.D., BARBOSA, F. The Use of Decision Trees and Naive Bayes Algorithms and Trace Element Patterns for Controlling the Authenticity of Free-Range-Pastured Hens' Eggs. **Journal of Food Science**, 2014a. v. 79, n. 9, p. C1672–C1677.

BARBOSA, R.M., NACANO, L.R., FREITAS, R., BATISTA, B.L., BARBOSA, F. The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee. **Food Research International**, 2014b. v. 61, p. 246–251.

BORRÀS, E., FERRÉ, J., BOQUÉ, R., MESTRES, M., ACEÑA, L., BUSTO, O. **Data fusion methodologies for food and beverage authentication and quality assessment - A review**. **Analytica Chimica Acta**. Elsevier.

BRACESCO, N., SANCHEZ, A.G., CONTRERAS, V., MENINI, T., GUGLIUCCI, A. Recent advances on Ilex paraguariensis research: Minireview. **Journal of Ethnopharmacology**, 2011. v. 136, n. 3, p. 378–384.

BURGES, C. C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. **Data Mining and Knowledge Discovery**, 1998. v. 2, n. 2, p. 121–167.

CASALE, M., CASOLINO, C., OLIVERI, P., FORINA, M. The potential of coupling information using three analytical techniques for identifying the geographical origin of

Liguria extra virgin olive oil. **Food Chemistry**, 2010. v. 118, n. 1, p. 163–170.

CHEN, M., KHARE, S., HUANG, B., ZHANG, H., LAU, E., FENG, E. Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. **Industrial and Engineering Chemistry Research**, 2013. v. 52, n. 23, p. 7886–7895.

CHEN, M. J.; CHEN, K. N.; LIN, C. W. Optimization on response surface models for the optimal manufacturing conditions of dairy tofu. **Journal of Food Engineering**, 2005. v. 68, n. 4, p. 471–480.

COLMAN, E., WAEGEMAN, W., DE BAETS, B., FIEVEZ, V. Prediction of subacute ruminal acidosis based on milk fatty acids: A comparison of linear discriminant and support vector machine approaches for model development. **Computers and Electronics in Agriculture**, 2015. v. 111, p. 179–185.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, 1995. v. 20, n. 3, p. 273–297.

COZZOLINO, D.; RESTAINO, E.; FASSIO, A. Discrimination of yerba mate (Ilex paraguayensis St. Hil.) samples according to their geographical origin by means of near infrared spectroscopy and multivariate analysis. **Sensing and Instrumentation for Food Quality and Safety**, 2010. v. 4, n. 2, p. 67–72.

CRAIG, A.P., FRANCA, A.S., OLIVEIRA, L.S., IRUDAYARAJ, J., ILELEJI, K. Application of elastic net and infrared spectroscopy in the discrimination between defective and non-defective roasted coffees. **Talanta**, 2014. v. 128, p. 393–400.

CYNKAR, W., DAMBERGS, R., SMITH, P., COZZOLINO, D. Classification of Tempranillo wines according to geographic origin: Combination of mass spectrometry based electronic nose and chemometrics. **Analytica Chimica Acta**, 2010. v. 660, n. 1–2, p. 227–231.

DINIZ, P.H.G.D., GOMES, A.A., PISTONESI, M.F., BAND, B.S.F., DE ARAÚJO, M.C.U. Simultaneous Classification of Teas According to Their Varieties and Geographical Origins by Using NIR Spectroscopy and SPA-LDA. **Food Analytical Methods**, 2014. v. 7, n. 8, p. 1712–1718.

DINIZ, P.H.G.D., PISTONESI, M.F., ALVAREZ, M.B., BAND, B.S.F., DE ARAÚJO, M.C.U. Simplified tea classification based on a reduced chemical composition profile via successive projections algorithm linear discriminant analysis (SPA-LDA). **Journal of Food Composition and Analysis**, 2015. v. 39, p. 103–110.

DONG, Y., XIANG, B., GENG, Y., YUAN, W. Rough set based wavelength selection in near-infrared spectral analysis. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 126, p. 21–29.

DRIVELOS, S. A.; GEORGIOU, C. A. Multi-element and multi-isotope- ratio analysis to determine the geographical origin of foods in the European Union. **Trends in Analytical Chemistry**, 2012. v. 40, p. 38–51.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. **New York: John Wiley, Section**.

FERRÃO, M.F., VIERA, M.D.S., PAZOS, R.E.P., FACHINI, D., GERBASE, A.E., MARDER, L. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. **Fuel**, 2011. v. 90, n. 2, p. 701–706.

FILIP, R., LOTITO, S.B., FERRARO, G., FRAGA, C.G. ANTIOXIDANT ACTIVITY OF ILEX PARAGUARIENSIS AND RELATED SPECIES Rosana Filip, M.S., Silvina. B. Lotito, M.S. t, Graciela Ferraro, Ph.D., Cesar G. Fraga, Ph.D. l Pharmacognosy, and l Physical Chemistry-PRALIB, School of Pharmacy and Biochemistry, Universit. **Nutrition Research**, 2000. v. 20, n. 10, p. 1437–1446.

GALTIER, O., DUPUY, N., LE DRÉAU, Y., OLLIVIER, D., PINATEL, C., KISTER, J., ARTAUD, J. Geographic origins and compositions of virgin olive oils determinated by chemometric analysis of NIR spectra. **Analytica Chimica Acta**, 2007. v. 595, n. 1, p. 136–144.

GAO, H., LONG, Y., JIANG, X., LIU, Z., WANG, D., ZHAO, Y., LI, D., SUN, B. Beneficial effects of Yerba Mate tea (Ilex paraguariensis) on hyperlipidemia in high-fat-fed hamsters. **Experimental Gerontology**, 2013. v. 48, n. 6, p. 572–578.

GENDRIN, C.; ROGGO, Y.; COLLET, C. Content uniformity of pharmaceutical solid dosage forms by near infrared hyperspectral imaging: A feasibility study. **Talanta**, 2007. v. 73, n. 4, p. 733–741.

GILL, P. E.; WONG, E. Methods for convex and general quadratic programming. **Mathematical Programming Computation**, 2015. v. 7, n. 1, p. 71–112.

GIULIAN, R., DOS SANTOS, C.E.I., SHUBEITA, S. DE M., DA SILVA, L.M., YONEAMA, M.L., DIAS, J.F. The study of the influence of industrial processing on the elemental composition of mate tealeaves (Ilex paraguariensis) using the PIXE technique. **LWT - Food Science and Technology**, 2009. v. 42, n. 1, p. 74–80.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, 2003. v. 3, n. 3, p. 1157–1182.

HÄNSCH, R.; MENDEL, R. R. Physiological functions of mineral micronutrients (Cu, Zn, Mn, Fe, Ni, Mo, B, Cl). **Current Opinion in Plant Biology**, 2009. v. 12, n. 3, p. 259–266.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **Elements**, 2009. v. 1, p. 337–387.

HECK, C. I.; MEJIA, E. G. DE. Yerba mate tea (Ilex paraguariensis): A comprehensive review on chemistry, health implications, and technological considerations. **Journal of Food Science**, 2007. v. 72, n. 9.

HUANG, C. L.; WANG, C. J. A GA-based feature selection and parameters optimizationfor support vector machines. **Expert Systems with Applications**, 2006. v. 31, n. 2, p. 231–240.

JIANG, Y.; LI, C. MRMR-based feature selection for classification of cotton foreign matter using hyperspectral imaging. **Computers and Electronics in Agriculture**, 2015. v. 119, p. 191–200.

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

KHANMOHAMMADI, M., KARAMI, F., MIR-MARQUÉS, A., BAGHERI GARMARUDI, A., GARRIGUES, S., DE LA GUARDIA, M. Classification of persimmon fruit origin by near infrared spectrometry and least squares-support vector machines. **Journal of Food Engineering**, 2014. v. 142, p. 17–22.

KHANMOHAMMADI, M.; GARMARUDI, B. A.; LA GUARDIA, M. DE. Feature selection strategies for quality screening of diesel samples by infrared spectrometry and linear discriminant analysis. **Talanta**, 2013. v. 104, p. 128–134.

LAURSEN, K.H., SCHJOERRING, J.K., OLESEN, J.E., ASKEGAARD, M., HALEKOH, U., HUSTED, S. Multielemental fingerprinting as a tool for authentication of organic wheat, barley, faba bean, and potato. **Journal of Agricultural and Food Chemistry**, 11 maio. 2011. v. 59, n. 9, p. 4385–4396.

LINARES, A.R., HASE, S.L., VERGARA, M.L., RESNIK, S.L. Modeling yerba mate aqueous extraction kinetics: Influence of temperature. **Journal of Food Engineering**, 2010. v. 97, n. 4, p. 471–477.

LIU, C.; YANG, S. X.; DENG, L. Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning. **Journal of Food Engineering**, 2015. v. 161, p. 16–23.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **Ieee Transactions on Knowledge and Data Engineering**, 2005. v. 17, n. 4, p. 491–502.

LIU, L., COZZOLINO, D., CYNKAR, W.U., DAMBERGS, R.G., JANIK, L., O'NEILL, B.K., COLBY, C.B., GISHEN, M. Preliminary study on the application of visible-near infrared spectroscopy and chemometrics to classify Riesling wines from different countries. **Food Chemistry**, 2008. v. 106, n. 2, p. 781–786.

LONG, X.-X., LI, H.-D., FAN, W., XU, Q.-S., LIANG, Y.-Z. A model population analysis method for variable selection based on mutual information. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 121, p. 75–81.

LÓPEZ-CÓRDOBA, A., MATERA, S., DELADINO, L., HOYA, A., NAVARRO, A., MARTINO, M. Compressed tablets based on mineral-functionalized starch and co-crystallized sucrose with natural antioxidants. **Journal of Food Engineering**, 2015. v. 146, p. 234–242.

LUTS, J., OJEDA, F., VAN DE PLAS RAF, R., DE MOOR, B., VAN HUFFEL, S., SUYKENS, J.A.K. A tutorial on support vector machine-based methods for classification problems in chemometrics. **Analytica Chimica Acta**, 2010. v. 665, n. 2, p. 129–145.

LUYKX, D. M. A. M.; RUTH, S. M. VAN. An overview of analytical methods for determining the geographical origin of food products. **Food Chemistry**, 2008. v. 107, n. 2, p. 897–911.

MAATHUIS, F. J. Physiological functions of mineral macronutrients. **Current Opinion in Plant Biology**, 2009. v. 12, n. 3, p. 250-258

MAIONE, C., LEMOS, B., DOBAL, A., BARBOSA, F., MELGAÇO, R. Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. **Computers and Electronics in Agriculture**, 2016. v. 121, p. 101–107.

MARCELO, M.C.A., MARTINS, C.A., POZEBON, D., DRESSLER, V.L., FERRÃO, M.F. Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate. **Anal. Methods**, 2014. v. 6, n. 19, p. 7621–7627.

MARCELO, M.C.A., MARTINS, C.A., POZEBON, D., FERRÃO, M.F. Classification of yerba mate (Ilex paraguariensis) according to the country of origin based on element concentrations. **Microchemical Journal**, 2014. v. 117, p. 164–171.

MOREDA-PIÑEIRO, A.; FISHER, A.; HILL, S. J. The classification of tea according to region of origin using pattern recognition techniques and trace metal data. **Journal of Food Composition and Analysis**, 2003. v. 16, n. 2, p. 195–211.

NUNES, G.L., BOAVENTURA, B.C.B., PINTO, S.S., VERRUCK, S., MURAKAMI, F.S., PRUDÊNCIO, E.S., DE MELLO CASTANHO AMBONI, R.D. Microencapsulation of freeze concentrated Ilex paraguariensis extract by spray drying. **Journal of Food Engineering**, 2015. v. 151, p. 60–68.

OTTAVIAN, M., FACCO, P., BAROLO, M., BERZAGHI, P., SEGATO, S., NOVELLI, E., BALZAN, S. Near-infrared spectroscopy to assist authentication and labeling of Asiago d'allevo cheese. **Journal of Food Engineering**, 2012. v. 113, n. 2, p. 289–298.

PILLONEL, L., LUGINBÜHL, W., PICQUE, D., SCHALLER, E., TABACCHI, R., BOSSET, J.O. Analytical methods for the determination of the geographic origin of Emmental cheese: Mid- and near-infrared spectroscopy. **European Food Research and Technology**, 2003. v. 216, n. 2, p. 174–178.

PONTES, M.J.C., GALVAO, R.K.H., ARAUJO, M.C.U., NOGUEIRA, P., MOREIRA, T., NETO, O.D.P., JOSE, G.E., SALDANHA, T.C.B. The successive projections algorithm for spectral variable selection in classification problems. **Chemometrics and Intelligent Laboratory Systems**, jul. 2005. v. 78, n. 1–2, p. 11–18.

RACHOW, T., BERGER, S., BOETTGER, M.K., SCHULZ, S., GUINJOAN, S., YERAGANI, V.K., VOSS, A., BÄR, K.J. Nonlinear relationship between electrodermal activity and heart rate variability in patients with acute schizophrenia. **Psychophysiology**, 2011. v. 48, n. 10, p. 1323–1332.

RAKOTOMAMONJY, A. Variable Selection Using SVM-based Criteria. **Journal of Machine Learning Research**, 2003. v. 3, p. 1357–1370.

REBOLO, S., PEÑA, R.M., LATORRE, M.J., GARCÍA, S., BOTANA, A.M., HERRERO, C. Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. **Analytica Chimica Acta**, 2000. v. 417, n. 2, p. 211–220.

RENCHER, A.C., 2002. Methods of Multivariate Analysis, Second Edition, Wiley, New York.

RODRIGUEZ-LUJAN, I., HUERTA, R., ELKAN, C., SANTA CRUZ, C. Quadratic Programming Feature Selection. **Journal of Machine Learning Research**, 2010. v. 11, p. 1491–1516.

RODRÍGUEZ-ROSARIO, C.A., MODI, K., KUAH, A., SHAJI, A., SUDARSHAN, E.C.G. Completely positive maps and classical correlations. **Journal of Physics A: Mathematical and Theoretical**, 2008. v. 41, n. 20, p. 205-301.

ROSSI, F., LENDASSE, A., FRANÇOIS, D., WERTZ, V., VERLEYSEN, M. Mutual information for the selection of relevant variables in spectrometric nonlinear modelling.

Chemometrics and Intelligent Laboratory Systems, 2006. v. 80, n. 2, p. 215–226.

SCHMALKO, M. E.; LOVERA, N. N.; KOLOMIEJEC, G. C. Moisture migration during a tempering time after the heat treatment step in yerba mate processing. **Latin American Applied Research**, 2011. v. 41, n. 2, p. 153–156.

DE VASCONCELOS, F.V.C., DE SOUZA, P.F.B., PIMENTEL, M.F., PONTES, M.J.C., PEREIRA, C.F. Using near-infrared overtone regions to determine biodiesel content and adulteration of diesel/biodiesel blends with vegetable oils. **Analytica Chimica Acta**, 2012. v. 716, p. 101–107.

WOODCOCK, T., DOWNEY, G., KELLY, J.D., O’DONNELL, C. Geographical classification of honey samples by near-infrared spectroscopy: A feasibility study. **Journal of Agricultural and Food Chemistry**, 2007. v. 55, n. 22, p. 9128–9134.

XIAOBO, Z., JIEWEN, Z., POVEY, M.J.W., HOLMES, M., HANPIN, M. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, 2010. v. 667, n. 1–2, p. 14–32.

XIE, L.; YING, Y.; YING, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. **Journal of Food Engineering**, 2009. v. 94, n. 1, p. 34–39.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 128, p. 17–24.

ZHANG, Y., ZHENG, L., LI, M., DENG, X., JI, R. Predicting apple sugar content based on spectral characteristics of apple tree leaf in different phenological phases. **Computers and Electronics in Agriculture**, 2015. v. 112, p. 20–27.

ZHAO, H., GUO, B., WEI, Y., ZHANG, B. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. **Food Chemistry**, 2013. v. 138, n. 2–3, p. 1902–1907.

3 Artigo 2 - Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples

Artigo a ser submetido no Journal of Pharmaceutical and Biomedical Analysis (Elsevier, JCR 5yIF: 2.953, Qualis A2)

Abstract

Street cocaine is typically altered with several compounds that increase its harmful health-related side effects, most notably depression, convulsions, and severe damages to the cardiovascular system, lungs, and brain. Thus, determining the concentration of cocaine and adulterants in seized drug samples is important from both health and forensic perspectives. Although FTIR has been widely used to identify the fingerprint and concentration of chemical compounds, spectroscopy datasets are usually comprised of thousands of highly correlated wavenumbers which, when used as predictors in regression models, tend to undermine the predictive performance of multivariate techniques. In this paper, we propose an FTIR wavenumber selection method aimed at identifying FTIR spectra intervals that best predict the concentration of cocaine and adulterants (e.g. caffeine, phenacetin, levamisole, and lidocaine) in cocaine samples. For that matter, the Mutual Information measure is integrated into a Quadratic Programming problem with the objective of minimizing the probability of retaining redundant wavenumbers, while maximizing the relationship between retained wavenumbers and compounds' concentrations. Optimization outputs guide the order of inclusion of wavenumbers in a predictive model, using a forward-based wavenumber selection method. After the inclusion of each wavenumber, parameters of three alternative regression models are estimated, and each model's prediction error is assessed through the Mean Average Error (MAE) measure; the recommended subset of retained wavenumbers is the one that minimizes the prediction error with maximum parsimony. Using our propositions in a dataset of 115 cocaine samples we obtained a best prediction model with average MAE of

0.0502 while retaining only 2.29% of the original wavenumbers, increasing the predictive precision by 0.0359 when compared to a model using the complete set of wavenumbers as predictors.

Keywords: Wavenumber selection, prediction, FTIR, Cocaine, Adulterants

3.1 INTRODUCTION

The estimated global production of cocaine is around 900 tons per year with a number of users surpassing 18 million people, making cocaine one of the most consumed drugs around the world. South America concentrates most of its production (approximately 60% of global cocaine seizures occur in the continent), and the drug is mostly trafficked to North America and Western/Central Europe. In Brazil, the increasing number of cocaine seizures is attributed to a combination of improved law enforcement, growing domestic demand for the drug, and increasing number of shipments to overseas markets departing from Brazilian ports (UNITED NATIONS, 2016). In 2006 the Brazilian Federal Police (BFP) created the PeQui project ("Perfil Químico de Drogas" in Portuguese), which aims at providing investigative forces with detailed chemical analyses of drugs trafficked in the country. Since BFP mainly deals with federal crimes, local law enforcement agents usually carry out drug analyses using samples from street drug seizures (ANZANELLO et al. 2015; BOTELHO et al. 2014; MARCELO et al. 2015).

Cocaine is extracted from the leaves of *Erytroxylum coca*, and is mainly consumed in salt, crack or freebase forms. While cocaine hydrochloride is obtained as a powder that can be administrated intravenously or via aspiration, crack is usually presented as a rock easily volatilized when heated due to its low melting point. In spite of the powerful anesthetic properties of cocaine, its abusive intake may lead to depression and a large variety of harmful effects to the cardiovascular system, lungs and brain (GOLDSTEIN; DESLAURIERS;

BURDA, 2009; PAWLIK et al., 2015; SOUZA et al., 2016). In addition, several adulterants are typically mixed with the drug, most notably caffeine (stimulant), phenacetin (analgesic), levamisole (anthelmintic), and lidocaine (local anesthetic) (BOTELHO et al., 2014; MAGALHÃES et al., 2013; PAWLIK et al., 2015). Such substances produce similar effects to those obtained by ingesting pure cocaine, but are normally less expensive (GROBÉRIO et al., 2015; INDORATO; ROMANO; BARBERA, 2016; LAPACHINSKE et al., 2015; MAGALHÃES et al., 2013). Thus, determining the concentration of cocaine and its adulterants in seized samples is not only valuable from a clinical perspective, but also provides relevant information to investigative forces towards interrupting drug trafficking (BERNARDO et al., 2003).

Over the last years, Fourier Transformed Infrared (FTIR) spectroscopy has gained wide acceptance in many research fields as a fast and non-destructive technique for identifying the fingerprint of several chemical compounds (CRAIG et al., 2014; LIU; YANG; DENG, 2015; ZHANG; ZHANG; IQBAL, 2013). In addition, such technique does not require previous preparation of samples. FTIR datasets are usually treated using multivariate techniques with applications in several fields and products, including fuels (FERRÃO et al., 2011; SILVA et al., 2012), food (CRAIG et al., 2014; MARCELO; POZEBON; FERRÃO, 2015), and pharmaceuticals (ANZANELLO et al., 2013). Better aligned with the propositions of this paper, FTIR and multivariate techniques have been used to analyze datasets obtained through spectroscopy on drug samples, such as amphetamines (PRAISLER et al., 2000), cocaine (MARCELO et al., 2015) and heroin (YUSOFF et al., 2017).

FTIR analyses produce information on absorbance values that enable identification and quantification of several chemical compounds. FTIR datasets are comprised of a large number of highly correlated and noisy variables known as wavenumbers. When used as input data in multivariate prediction techniques, such type of data tend to compromise the performance of predictive models. To overcome such drawback wavenumber selection techniques have become key in reducing the influence of ill-conditioned FTIR data on multivariate techniques (BALABIN; SMIRNOV, 2011; COZZOLINO; RESTAINO; FASSIO, 2010; DONG et al., 2013; XIAOBO et al., 2010). The analysis and selection of the most relevant regions of the FTIR spectra also generates models that are simpler and easier to interpret by highlighting not only relationships between wavenumbers, but of wavenumbers and the investigated property (XIE; YING; YING, 2009; ZHANG; ZHANG; IQBAL, 2013). Finally, model complexity and data collection costs are reduced through wavenumber selection (CHEN et al., 2013).

This paper proposes a new method for wavenumber selection aimed at predicting the concentration of cocaine and adulterants in cocaine samples. We propose the use of Quadratic Programming (QP) to simultaneously minimize the probability of retaining redundant wavenumbers, and to maximize the relationship between retained wavenumbers and the response variable (compound's concentration). Optimization gives rise to an importance index that quantifies the predictive potential of each wavenumber. Using the index, wavenumbers are inserted into three regression techniques for predicting the concentration of cocaine and adulterants; namely: Multiple Linear Regression, Principal Components Regression, and Partial Least Squares Regression.

We applied our propositions to an FTIR dataset consisting of 115 cocaine samples (58 seized and 57 synthetic); each sample is described by 662 wavenumbers. Using the recommended regression technique, we obtained (*i*) a Mean Average Error (MAE) of 0.0879 for cocaine concentration, while retaining only 2.03% of the original wavenumbers in average, and (*ii*) an MAE of 0.0408 for adulterants concentration while retaining only 2.35% of the original wavenumbers in average.

3.2 MATERIALS AND METHOD

3.2.1 Samples, Sample Preparation, and Instrumentation

Fifty-eight samples of cocaine (crack, freebase, and salt cocaine) seized by the BFP in the state of Rio Grande do Sul between 2013 and 2015 were used in this study. In addition, 57 solid samples were prepared mixing cocaine and its adulterants. All samples were homogenized using an agate mortar prior to analysis; the cocaine standard was provided by the BFP. Lidocaine (Delaware, Brazil), levamisole (Sigma-Aldrich), caffeine (Acrosorganics 98.5%, NJ, USA) and phenacetin (Delaware, Brazil) were used as adulterants in the prepared mixtures, which had two to five components each.

FTIR spectra were acquired through a Nicolet 380 FTIR Spectrometer (Nicolet Instrument Co., Madison, USA) equipped with DTGS (deuterated triglycine sulphate) detector and a smart orbit single reflection diamond ATR sampling accessory. Thirty-two scans were performed with resolution of 4 cm-1. The fingerprint spectral region ranging from 550 to 1800 cm-1 was selected for the multivariate analysis; spectra data was not preprocessed.

A liquid chromatography system with diode array detector (Agilent Technologies, USA) was used as reference method for determining the concentration of cocaine and its adulterants in the seized samples. The separation was carried out in isocratic mode, C18 column (Zorbax EclipsePlus, 4.6° 250 mm; Agilent, USA) at 30 °C with the full UV-Vis monitored, a bandwidth of 4 nm and 0.5 nm resolution. For the mobile phase a 1:1 solution of acetonitrile (Panreac, Spain) and water purified in a Milli-Q system (Millipore) was used, with amonium acetate (F. Maia, Brazil) as buffer (pH 8.3) and a flow of 1 mL min-1. Compound concentrations were assessed in mg⁄g through external calibration. For more details on methodology and figures of merit, see (MARCELO et al., 2016).

3.2.2 Multivariate Techniques

We now present the fundamentals of the data analysis and optimization techniques used in the proposed method; namely: Mutual Information, Quadratic Programming, Multiple Linear Regression, Principal Component Regression, and Partial Least Squares Regression. Such techniques are used in our propositions due to its availability in many statistical packages for data analysis and suitability for FTIR analysis.

Mutual Information quantifies the amount of information shared between two wavenumbers using a symmetric and non-negative dependency measure. The Mutual Information between wavenumbers W and Y is given by:

$$\mathrm{I}\,(W;Y) = \sum_{y\epsilon Y}\sum_{w\epsilon\, W} \mathrm{p}(w,y)\log(\frac{\mathrm{p}(w,y)}{\mathrm{p}(w)\mathrm{p}(y)}) \qquad (1)$$

where $p(w,y)$ represents the joint probability distribution function of W and Y, and $p(w)$ and $p(y)$ are the marginal probability distribution functions of W and Y, respectively (LONG et al., 2013; RACHOW et al., 2011; RODRÍGUEZ-ROSARIO et al., 2008). These distributions can be either estimated by fitting tests to known density function methods or through the discretization of the density function (DUDA; HART; STORK, 2001).

A Multiple Linear Regression (MLR) equation provides a mathematical description of the relationship between a dependent variable (e.g. compound concentration) and one or more independent variables (e.g. wavenumbers). By assumption, independent variables should not display multicollinearity, which is usually the case when wavenumbers from an FTIR analysis are used as inputs. To overcome that the wavenumber dataset may be treated using dimensionality reduction techniques (ARAÚJO et al., 2001; YU; JIANG; LAND, 2015; ZHAO et al., 2013).

Principal Component Analysis (PCA) is a data reduction technique that replaces the original correlated variables (such as FTIR wavenumbers) by new uncorrelated variables known as Principal Components (PCs). Dimensionality reduction occurs when only a few components that account for most of the variability in the data replace the original set of variables; that enables, for example, the handling of high-dimensional datasets as the one resulting from spectroscopy analyses (ANZANELLO et al., 2013; DHARMARAJ et al., 2006; INSAUSTI et al., 2012).

Merging concepts from MLR and PCA, Principal Components Regression (PCR) builds a regression equation using PC scores instead of observations from the original variables used to obtain the PCs. PCR is an alternative to MLR when independent variables present multicollinearity (ANZANELLO et al., 2015a; ZHANG; ZHANG; IQBAL, 2013); drawbacks include the potential information loss resulting from selecting a limited number of PCs, and interpretability issues arising from using transformed rather than original variables in the regression (ARAÚJO et al., 2001; LIN et al., 2016).

Finally, Partial Least Squares Regression (PLSR) seeks a regression model describing the relationship between two sets of PCA-reduced variables. The first set corresponds to dependent variables; in that sense, PLSR may be viewed as a multiresponse expansion of PCR. The second set corresponds to independent variables. In PLSR regression coefficients are determined by maximizing the covariance between the two sets of reduced variables, which is not the same criterion used in PCR. Thus, modeling a single response problem through PLSR and PCR, which is our case in this paper, should lead to different results. In a dataset containing *N* wavenumbers, PLSR regression generates *A* orthogonal linear combinations (where $A<N$) of the original wavenumbers (ANZANELLO et al., 2015b; WOLD; SJÖSTRÖM; ERIKSSON, 2001; YEH et al., 2016). Data projection enables PLSR to handle data with strong collinearity, high levels of noise and substantially fewer samples than wavenumber. Although being more popular than PCR there is no obvious advantage of PLSR over PCR (LIN et al., 2016), justifying the comparison of methods proposed here.

3.2.3 Method

The proposed method for wavenumber selection aimed at predicting the concentration of cocaine and adulterants in cocaine samples is implemented in five steps: (i) split the original data into training and test sets; (ii) generate a redundancy matrix and a similarity vector via Mutual Information using the training set; (iii) create a QP problem to evaluate the relevance of wavenumbers and generate a Wavenumber Importance Index (WII); (iv) predict the concentration of cocaine and adulterants in samples belonging to the training set using the model building techniques in section 3.2.2; use a forward procedure entering wavenumbers in models according to the order given by WII, and retain the most parsimonious subset of wavenumbers that yields the smallest prediction error; and (v) predict cocaine and adulterants' concentration in samples from the test set using the retained subset of wavenumbers. Steps are detailed next.

In step (i) split the M samples described by N wavenumbers into two sets: a training set containing TR samples, and a test set containing TS samples, such that $TR + TS = M$. The recommended proportion between TR and TS is 80%-20% (GARCÍA NIETO et al., 2016). The training set is used to select the most important wavenumbers. The test set is comprised of new samples whose concentration will be predicted to verify the performance of the final model.

When selecting wavenumbers for prediction the typical goal is to retain a minimum redundancy set for parsimony, preserving as much information as possible regarding the relationship between wavenumbers and response variable (GUYON; ELISSEEFF, 2003; LIU; YU, 2005). For that, in step (ii) we propose the determination of a redundancy matrix **R** describing wavenumbers' shared information, and a similarity vector **s** describing shared information between wavenumbers and response variable (i.e. compound concentration). Using equation (1) to calculate the Mutual Information between pairs of wavenumbers, **R** is constructed as a (*N*×*N*) matrix with element r_{ij} giving the Mutual Information between wavenumbers *i* and *j*. Similarity vector **s** is a *N*–dimensional vector with *i–th* element giving the Mutual Information between wavenumber *i* and the response variable.

In step (iii) we propose a Quadratic Programming formulation that uses the information in **R** and **s**. The objective function in eqn. (2) minimizes the probability of retaining redundant wavenumbers while maximizing the similarity between wavenumbers and the response variable. In that equation, parameter α is a scalar in the interval [0,1] that ensures the balance between quadratic (written as a function of **R**) and linear (written as a function of **s**) terms in the optimization (RODRIGUEZ-LUJAN et al., 2010). Constraints in equations (3) and (4) restrict the domain of **x** to the [0,1] interval. Therefore, vector **x** in equation (2) gives percentage weights associated to wavenumbers that guarantees the dual goal in the objective function (RODRIGUEZ-LUJAN et al., 2010). They are realizations of the Wavenumber Importance Index (WWI) used later in the prediction step; the larger the value of WII_i, the more important wavenumber *i* is in predicting the response.

$$\underset{x \in R^N}{minimize}\ f(\mathbf{x}) = \{\tfrac{1}{2}(1-\alpha)\mathbf{x}^{\mathrm{t}}\mathbf{R}\mathbf{x} - \alpha\mathbf{s}^{\mathrm{t}}\mathbf{x}\,\} \qquad (2)$$

Subject to

$$x_i \geq 0 \qquad (3)$$

$$\textstyle\sum_{i=1}^{N} x_i = 1 \qquad (4)$$

In the final steps of the proposed method, we first use the training set to build regression models relating each compound’s concentration with wavenumbers [step (iv)], and then use the test set to evaluate the performance of resulting models [step (v)]. These two steps are repeated for each response variable (cocaine, lidocaine, levamisole, caffeine, and phenacetin) and regression strategy (MLR, PCR, and PLSR).

In step (iv) we initially obtain a regression model using the wavenumber with largest WII as independent variable, compute the MAE, and then calculate d_n [equation (5)], which gives the distance of the iteration results in terms of accuracy and percentage of retained wavenumbers to a hypothetical ideal point (corresponding to a model with a single wavenumber as predictor and zero prediction error), as illustrated in Figure 3-1.

$$d_n = \sqrt{(0 - MAE_n)^2 + (\frac{1-n}{N})^2} \qquad (5)$$

In eqn. (5), MAE_n is the mean average error of prediction of a model consisting of n wavenumbers from a total of N. Next, the wavenumber with second largest WII is added to the set of predictors in the model, and iteration results are calculated. The procedure is repeated until all wavenumbers have been inserted in the predictors' dataset.

Some papers favor the use of the Root Mean Squared Error as prediction error measure. Here, we use the Mean Average Error to maintain both terms in a linear scale. Terms related to error and percentage of retained wavenumbers in equation (5) are in the [0,1] domain establishing a balanced tradeoff between these two goals. Alternatively, importance weights may be used to enhance the influence of accuracy or wavenumber retention in the selection process.

The predictors' subset yielding the smallest distance to the optimum point is retained. In case multiple subsets yield the minimum distance, the subset with fewer wavenumbers is chosen.

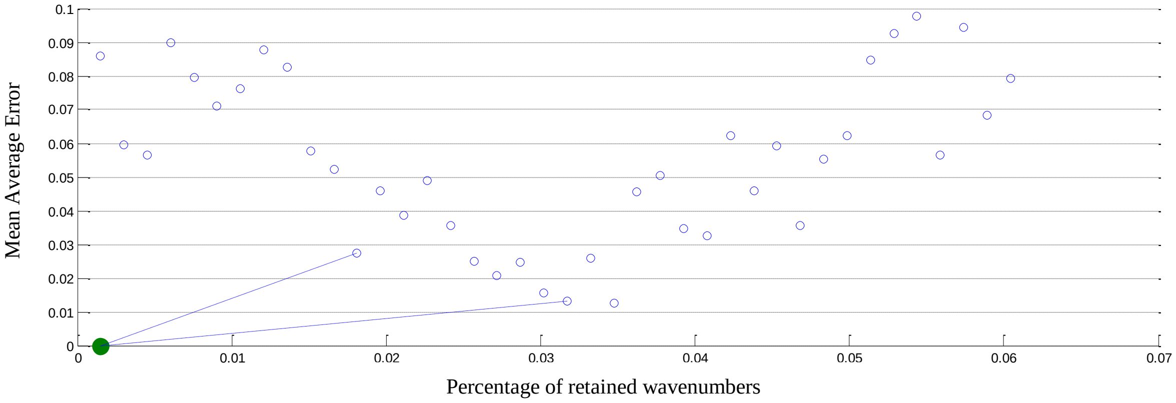


Figure 3-1 – Illustration of distance between model results and hypothetical optimum point

In step (v), a model based on the subset of selected wavenumbers is used to predict the compound’s concentration in samples from the test set, validating the model’s prediction performance in observations not used in step (iv).

3.3 RESULTS AND DISCUSSION

We now present the results of applying the proposed method to an FTIR dataset comprised of seized cocaine and laboratory mixed samples; the scope is to predict the concentration of cocaine and adulterants (caffeine, phenacetin, levamisole, and lidocaine) in samples. There are 115 samples in the dataset. Each sample is described by 662 wavenumbers in the 526-1801 cm^{-1} interval; the raw spectra are displayed in Figure 3-2.

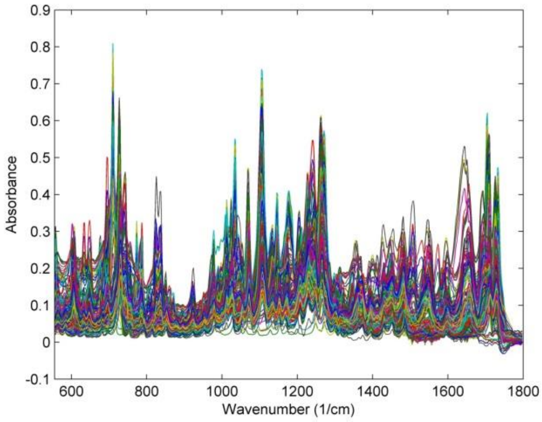


Figure 3-2 – Raw spectra of FTIR absorbance for 115 cocaine samples

To avoid bias by sampling, we run 500 replications of the method in section 2.3 by shuffling the original dataset following an 80%-20% training-test proportion (i.e. 92 samples in the training set, and 23 samples in the test set). At each replication, models obtained through MLR, PCR and PLSR are used to predict the concentration of each assessed compound. The number of retained components in PCR was determined by cross-validation (DUDA; HART; STORK, 2001; REBOLO et al., 2000), parameter *A* of PLSR was defined as recommended in Wold et al. (2001).

3.3.1 Cocaine

The first dependent variable tested was cocaine concentration. Figure 3-3 illustrates the MAE profile for one of the 500 replications as wavenumbers are inserted into a PLSR model according to the WII. MAE values decrease substantially as relevant wavenumbers are inserted in model, reaching a local minimum at 6.73% when approximately 12.39% of the full spectra is retained. After that MAE values decrease slightly as remaining wavenumbers are inserted in the model. The global minimum MAE value (6.22%) is attained when around 94.41% of wavenumbers are kept in the model. However, the small increase in prediction accuracy is not compensated by the large number of predictors required in the model.

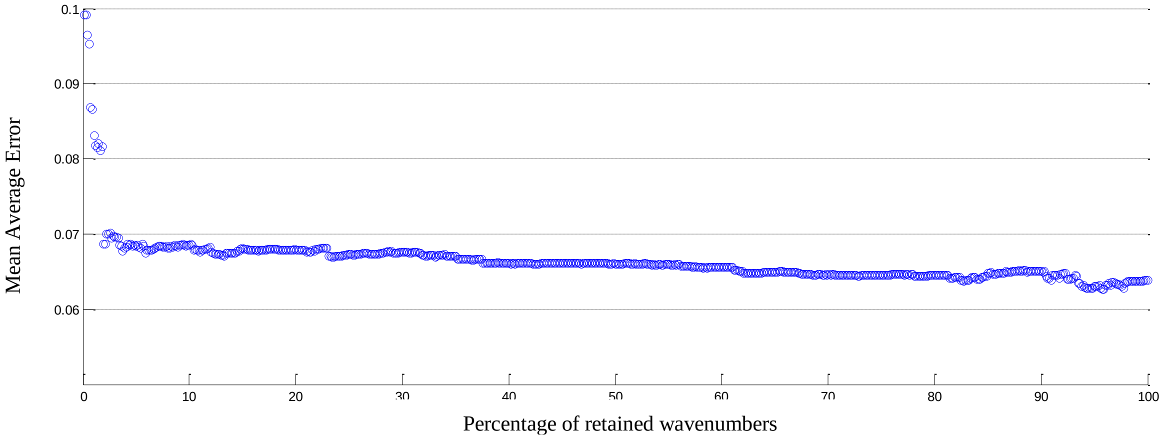


Figure 3-3 – Realization of cocaine MAE profile as wavenumbers are inserted in a PSLR model

Table 3-1 depicts the average MAE in training and test sets, the average percentage of retained wavenumbers, and the standard deviation of such metrics (in parentheses) for 500 replications of each regression method; best results are highlighted in bold. Note the substantial reduction in the percentage of wavenumbers required in the models: from 1.42% to 3.76% of the original 662 wavenumbers. MLR and PLSR models with the optimized set of predictors displayed substantial improvements in accuracy if compared to models using all wavenumbers: MAE dropped from average 26.44% to 10.25% for MLR, and from average 9.13% to 8.79% for PLSR. That was not observed when PCR was the modeling strategy: MAE increased from average 8.19% using all wavenumbers as predictors to 9.93% using the reduced subset. That suggests some information was lost when replacing original wavenumbers by PCA scores in the regression model.

	MLR	PCR	PLSR
MAE training set	0.0618 (0.0072)	0.0839 (0.0051)	0.0781 (0.0051)
MAE test set	0.1025 (0.0202)	0.0993 (0.0167)	0.0879 (0.0161)
% of retained wavenumbers	3.76% (1.08%)	1.42% (0.67%)	2.03% (0.73%)
MAE for the test set with all wavenumbers	0.2644 (0.0592)	0.0819 (0.0152)	0.0913 (0.0172)

Table 3-1– Performance of regression models for predicting cocaine concentration using all wavenumbers and the reduced subset of predictors (standard deviations in parentheses)

Although yielding the lowest average MAE in the training set MLR presents the highest variation compared to other regression strategies. That leads to lack of generalization, with good results in the training set that are not replicable in the test set. PLSR provides the lowest and more consistent MAE in the testing set when using a reduced subset of wavenumbers (2.03%), being recommended to predict cocaine concentration.

Figure 3-4 illustrates the distribution of the most frequently retained wavenumbers when using PLSR to predict cocaine concentration, based on 500 replications. Peaks from largest to smallest correspond to the following wavenumbers: 1724-1728 cm^{-1}, 1525-1533 cm^{-1}, 1068 cm^{-1}, and 741 cm^{-1}. The first peak (1724-1728 cm^{-1}) is associated to the stretching vibration of the carbonyl group; the second peak (1525-1533 cm^{-1}) is related to C-H bending vibrations; third and fourth peaks (1068 cm^{-1} and 741 cm^{-1}) correspond to out-of-plane bending and mono substituted benzene stretching, respectively.

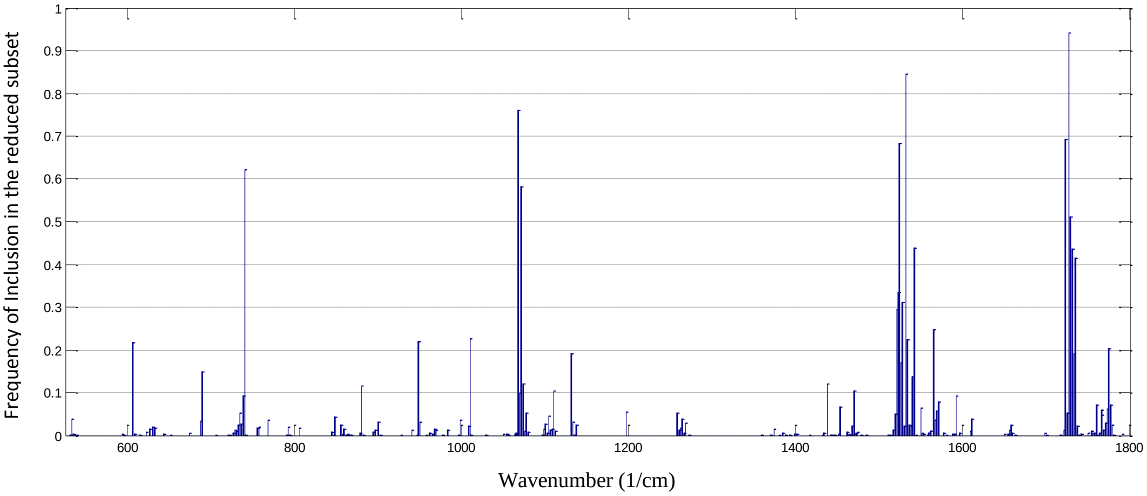


Figure 3-4 – Frequency of retained wavenumbers when using PLSR to predict cocaine concentration

3.3.2 Adulterants

We now present our method’s results when 4 cocaine adulterants (caffeine, phenacetin, levamisole, and lidocaine) are used as dependent variables in the regression models. Figure 3-5 illustrates a realization of the MAE profile for caffeine prediction as wavenumbers are inserted into the reduced dataset using MLR as regression strategy; other adulterants yielded similar MAE profiles and are not presented. MAE is greatly reduced when nearly 2% of the original wavenumbers are retained in the model, and approaches 0% when approximately 14% of the original wavenumbers are used in the model; i.e. 92 wavenumbers, which is same number of samples in the training set. Such behavior suggests that MLR tends to overfit when there are more variables than samples, requiring some wavenumber selection approach.

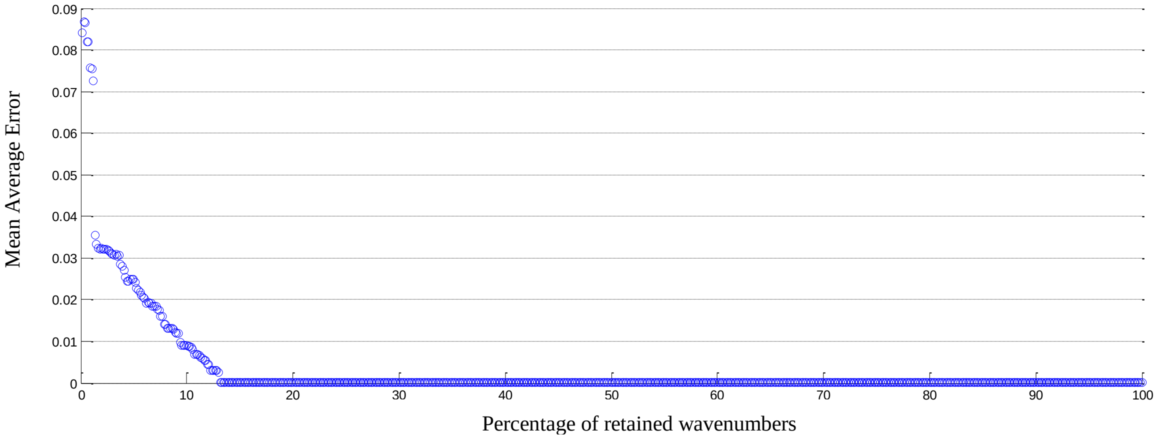


Figure 3-5 – Realization of caffeine MAE profile as wavenumbers are inserted in an MLR model

Table 3-2 – presents average MAE values for both training and test sets, and the percentage of retained wavenumbers obtained over 500 repetitions of each regression strategy when adulterants were used as dependent values (best results are highlighted in bold). Differently from Table 3-1– Performance of regression models for predicting cocaine concentration using all wavenumbers and the reduced subset of predictors (standard deviations in parentheses), MLR yields more homogenous and slightly better predictions of adulterants' concentrations than PLSR. PCR clearly leads to overfitted models, since differences in MAE in training and test sets are substantially higher when compared to PLSR and PCR.

	Caffeine			Phenacetin		
	MLR	PCR	PLS	MLR	PCR	PLS
MAE Training	0.0298 (0.0038)	0.0361 (0.0055)	0.0316 (0.0039)	0.0295 (0.0032)	0.0361 (0.0049)	0.0319 (0.0031)
MAE Test	0.0405 (0.0119)	0.0805 (0.0181)	0.0449 (0.0137)	0.0402 (0.0113)	0.0702 (0.0178)	0.0486 (0.0140)
% of retained wavenumbers	2.19% (0.53%)	1.94% (0.56%)	2.11% (0.52%)	2.15% (0.37%)	1.84% (0.35%)	2.03% (0.45%)
MAE Test with all variables	0.0727 (0.0317)	0.0334 (0.0088)	0.0351 (0.0093)	0.0813 (0.0356)	0.0406 (0.0110)	0.0383 (0.0109)
	Levamisole			Lidocaine		
	MLR	PCR	PLS	MLR	PCR	PLS
MAE Training	0.0332 (0.0052)	0.0427 (0.0049)	0.0374 (0.0051)	0.0277 (0.0034)	0.0359 (0.0055)	0.0304 (0.0037)
MAE Test	0.0459 (0.0125)	0.0689 (0.0176)	0.0492 (0.0132)	0.0366 (0.0115)	0.0950 (0.0185)	0.0438 (0.0152)
% of retained wavenumbers	2.72% (0.57%)	2,04% (0.51%)	2.48% (0.57%)	2.34% (0.41%)	2,12% (0.63%)	2.27% (0.46%)
MAE Test with all variables	0.1209 (0.0513)	0.0449 (0.0135)	0.0331 (0.0074)	0.0643 (0.0312)	0.0328 (0.0083)	0.0267 (0.0077)

Table 3-2 – Performance of regression models for predicting adulterants' concentrations using all wavenumbers and the reduced subset of predictors (standard deviations in parentheses)

On average MLR requires 15.5 wavenumbers to predict adulterants' concentrations (i.e. 2.35% of the original 662 wavenumbers), while PLSR requires 17.7 wavenumbers (or 2.22% of the total). Despite the similar number of retained wavenumbers in each regression strategy, MLR leads to less variable predictions. Additionally, when compared to models that predict cocaine concentration the standard deviation of the percentage of retained wavenumbers for adulterants' prediction is noticeably smaller.

Given the results in Table 3-2, we recommend MLR as regression strategy for predicting adulterants' concentrations in cocaine samples. In addition to its better performance in terms of prediction and wavenumber retention, MLR relies on simple mathematical foundations, enables direct model interpretation, and is widely available in statistical packages. Most important peaks appear in the 1637-1686 cm^{-1} region, suggesting that bands associated to the carbonyl group are relevant to quantify caffeine concentration.

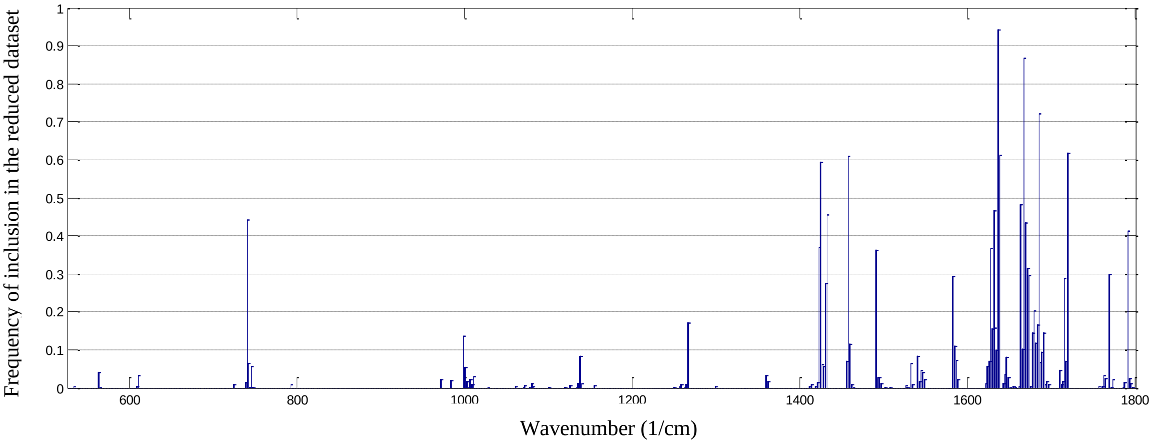


Figure 3-6 – Frequency of retained wavenumbers for caffeine prediction

The analysis of most frequently retained wavenumbers for phenacetin prediction in Figure 3-7 indicates peaks concentrated in 3 regions: 1506-1512 cm^{-1}, which is related to C-N stretch, and 825-837 cm^{-1} and 924-926 cm^{-1}, which are associated to C-H aromatic.

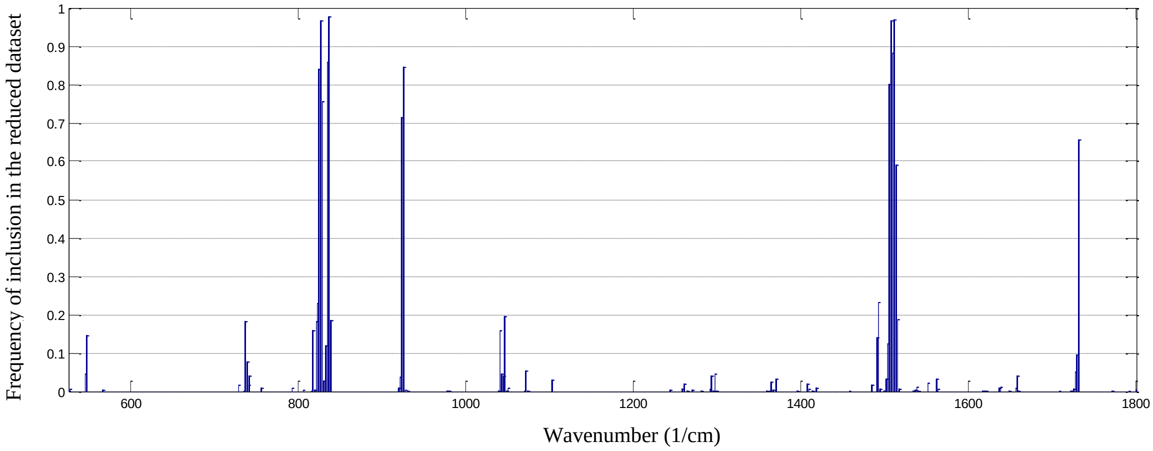


Figure 3-7 – Frequency of retained wavenumbers when using MLR to predict phenacetin concentration

Differently from Figure 3-6 and Figure 3-7, the most frequently retained wavenumbers to predict levamisole concentration are dispersed in several intervals of the spectra, with some of them selected only a few times, as illustrated in Figure 3-8. Such results suggest that shuffling of the original dataset to create training and test samples has influence on the selected wavenumbers. Wavenumbers retained in more than 60% of the 500 replications belong to the 607-696 cm^{-1} interval, related to aromatic stretching, and to the 1763-1772 cm^{-1} interval, which it is not related to levamisole but may be associated to carbonyl groups from cocaine and other adulterants.

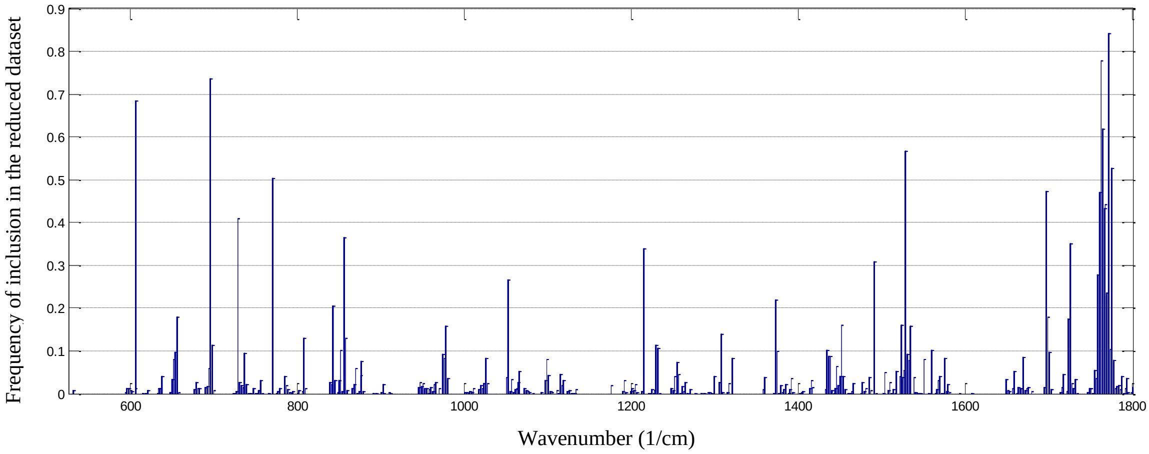


Figure 3-8 – Frequency of retained wavenumbers when using MLR to predict levamisole concentration

As for the most frequently retained wavenumbers to predict lidocaine, there is a predominance of wavenumbers in the 1209-1225 cm^{-1} and 1535-1547 cm^{-1} regions (see Figure 3-9). The regions correspond to C-O streching and an aromatic C-H, respectively.

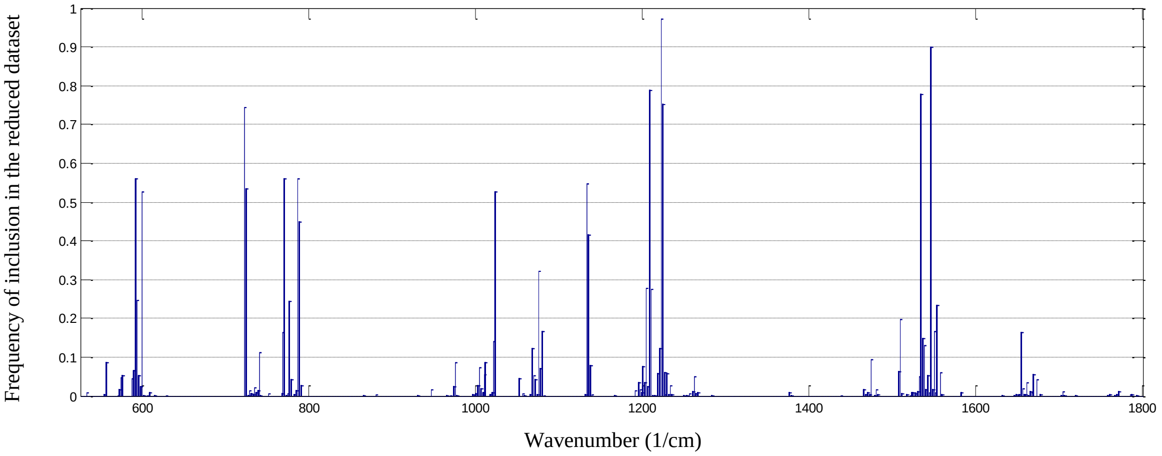


Figure 3-9 – Frequency of retained wavenumbers when using MLR to predict lidocaine concentration

3.4 Conclusions

FTIR spectra typically yield datasets comprised of a large number of noisy and correlated wavenumbers that tend to undermine the performance of several regression models. Thus, improving predictability of such models by selecting the most relevant FTIR regions is a relevant matter to properly determine chemical characteristics of cocaine samples and its adulterants.

A method for selecting the most relevant wavenumbers to be included in regression models is proposed in this paper. Its main contributions are: (i) a WII tailored to identify the wavenumbers that better identify the variation of compounds in cocaine samples; (ii) a multicriteria method to select a reduced part of the spectra considering the precision of the regression model to predict the compounds concentration and its dimensionality; and (iii) the analysis of the wavenumbers used in the regression model of each compound.

When applied to an FTIR dataset of 115 cocaine samples, PLSR yielded best results for the prediction of cocaine concentration, while MLR was recommended for predicting the concentration of cocaine adulterants. In all cases, a significant reduction in the percentage of wavenumbers required for prediction was observed. Analyzing the most frequently retained wavenumbers to predict compounds' concentrations it was possible to identify chemical functions related to them.

Future research includes the analysis of alternative predictive techniques (e.g. Neural Networks) in association with wavenumber importance indices derived from PLSR parameters. We will also explore extensions for classification models to simply determine the existence of adulterants in seized cocaine samples.

3.5 References

ANZANELLO, M.J., ORTIZ, R.S., LIMBERGERB, R.P., MAYORGA, P. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. **Journal of Pharmaceutical and Biomedical Analysis**, 2013. v. 83, p. 209–214.

ANZANELLO, M.J., FU, K., FOGLIATTO, F.F., FLÔRES, M. Chemometrics and Intelligent Laboratory Systems HATR – FTIR wavenumber selection for predicting biodiesel / diesel blends flash point. **Chemometrics and Intelligent Laboratory Systems**, 2015. v. 145, p. 1–6.

ANZANELLO, M.J., KAHMANN, A., MARCELO, M.C.A., MARIOTTI, K.C., FERRÃO, M.F., ORTIZ, R.S. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, 2015. v. 115, p. 562–569.

ARAÚJO, M.C.U., SALDANHA, T.C.B., GALVÃO, R.K.H., YONEYAMA, T., CHAME, H.C., VISANI, V. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. **Chemometrics and Intelligent Laboratory Systems**, 2001. v. 57, n. 2, p. 65–73.

BALABIN, R. M.; SMIRNOV, S. V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. **Analytica Chimica Acta**, 2011. v. 692, n. 1–2, p. 63–72.

BERNARDO, N.P., SIQUEIRA, M.E.P.B., DE PAIVA, M.J.N., MAIA, P.P. Caffeine and other adulterants in seizures of street cocaine in Brazil. **International Journal of Drug Policy**, 2003. v. 14, n. 4, p. 331–334.

BOTELHO, É.D., CUNHA, R.B., CAMPOS, A.F.C., MALDANER, A.O. Chemical profiling of cocaine seized by Brazilian federal police in 2009-2012: Major components. **Journal of the Brazilian Chemical Society**, 2014. v. 25, n. 4, p. 611–618.

CHEN, M., KHARE, S., HUANG, B., ZHANG, H., LAU, E., FENG, E. Recursive wavelength-selection strategy to update near-infrared spectroscopy model with an industrial application. **Industrial and Engineering Chemistry Research**, 2013. v. 52, n. 23, p. 7886–7895.

COZZOLINO, D.; RESTAINO, E.; FASSIO, A. Discrimination of yerba mate (Ilex paraguayensis St. Hil.) samples according to their geographical origin by means of near infrared spectroscopy and multivariate analysis. **Sensing and Instrumentation for Food Quality and Safety**, 2010. v. 4, n. 2, p. 67–72.

CRAIG, A.P., FRANCA, A.S., OLIVEIRA, L.S., IRUDAYARAJ, J., ILELEJI, K. Application of elastic net and infrared spectroscopy in the discrimination between defective and non-defective roasted coffees. **Talanta**, 2014. v. 128, p. 393–400.

DHARMARAJ, S., HOSSAIN, M.A., ZHARI, S., HARN, G.L., ISMAIL, Z. The use of principal component analysis and self-organizing map to monitor inhibition of calcium oxalate crystal growth by Orthosiphon stamineus extract. **Chemometrics and Intelligent Laboratory Systems**, 2006. v. 81, n. 1, p. 21–28.

DONG, Y., XIANG, B., GENG, Y., YUAN, W. Rough set based wavelength selection in near-infrared spectral analysis. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 126, p. 21–29.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. **New York: John Wiley**, 2001.

FERRÃO, M.F., VIERA, M.D.S., PAZOS, R.E.P., FACHINI, D., GERBASE, A.E., MARDER, L. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. **Fuel**, 2011. v. 90, n. 2, p. 701–706.

GARCÍA NIETO, P.J., GARCÍA-GONZALO, E., ARBAT, G., DURAN-ROS, M., RAMÍREZ DE CARTAGENA, F., PUIG-BARGUÉS, J. A new predictive model for the filtered volume and outlet parameters in micro-irrigation sand filters fed with effluents using the hybrid PSO-SVM-based approach. **Computers and Electronics in Agriculture**, 2016. v. 125, p. 74–80.

GOLDSTEIN, R. A.; DESLAURIERS, C.; BURDA, A. M. Cocaine: History, Social Implications, and Toxicity-A Review. **Disease-a-Month**, 2009. v. 55, n. 1, p. 6–38.

GROBÉRIO, T.S., ZACCA, J.J., BOTELHO, É.D., TALHAVINI, M., BRAGA, J.W.B. Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR. **Forensic Science International**, 2015. v. 257, p. 297–306.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, 2003. v. 3, n. 3, p. 1157–1182.

INDORATO, F.; ROMANO, G.; BARBERA, N. Levamisole-adulterated cocaine: Two fatal case reports and evaluation of possible cocaine toxicity potentiation. **Forensic Science International**, 2016. v. 265, p. 103–106.

INSAUSTI, M., GOMES, A.A., CRUZ, F. V., PISTONESI, M.F., ARAUJO, M.C.U., GALVÃO, R.K.H., PEREIRA, C.F., BAND, B.S.F.Screening analysis of biodiesel feedstock using UV-vis, NIR and synchronous fluorescence spectrometries and the successive projections algorithm. **Talanta**, 2012. v. 97, p. 579–583.

LAPACHINSKE, S.F., OKAI, G.G., DOS SANTOS, A., DE BAIRROS, A.V., YONAMINE, M. Analysis of cocaine and its adulterants in drugs for international trafficking seized by the Brazilian Federal Police. **Forensic Science International**, 2015. v. 247, p. 48–53.

LIN, Y., DENG, B., XU, Q., YUN, Y., LIANG, Y. Chemometrics and Intelligent Laboratory Systems The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework. **Chemometrics and Intelligent Laboratory Systems**, 2016. v. 150, p. 58–64.

LIU, C.; YANG, S. X.; DENG, L. Determination of internal qualities of Newhall navel oranges based on NIR spectroscopy using machine learning. **Journal of Food Engineering**, 2015. v. 161, p. 16–23.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **Ieee Transactions on Knowledge and Data Engineering**, 2005. v. 17, n. 4, p. 491–502.

LONG, X.-X., LI, H.-D., FAN, W., XU, Q.-S., LIANG, Y.-Z. A model population analysis method for variable selection based on mutual information. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 121, p. 75–81.

MAGALHÃES, E.J., NASCENTES, C.C., PEREIRA, L.S.A., GUEDES, M.L.O., LORDEIRO, R.A., AULER, L.M.L.A., AUGUSTI, R., DE QUEIROZ, M.E.L.R. Evaluation of the composition of street cocaine seized in two regions of Brazil. **Science and Justice**, 2013. v. 53, n. 4, p. 425–432.

MARCELO, M.C.A., FIORENTIN, T.R., MARIOTTI, K.C., ORTIZ, R.S., LIMBERGER, R.P. Profiling cocaine by ATR-FTIR. **Forensic Science International**, 2015. v. 246, p. 65–71.

MARCELO, M.C.A., MARIOTTI, K.C., FERRÃO, M.F., ORTIZ, R.S. Analytical Methods Determination of cocaine and its main adulterants in seized drugs from Rio Grande do Sul , Brazil , by a Doehlert optimized LC-DAD method. 2016. p. 5212–5217.

MARCELO, M. C. A.; POZEBON, D.; FERRÃO, M. F. Authentication of yerba mate according to the country of origin by using Fourier transform infrared (FTIR) associated with chemometrics. **Food Additives & Contaminants: Part A**, 2015. v. 32, n. 8, p. 1215–1222.

PAWLIK, E., MAHLER, H., HARTUNG, B., PLÄSSER, G., DALDRUP, T. Drug-related death: Adulterants from cocaine preparations in lung tissue and blood. **Forensic Science International**, 2015. v. 249, p. 294–303.

PRAISLER, M., DIRINCK, I., BOCXLAER, J. VAN, LEENHEER, A. DE, MASSART, D.L. Exploratory analysis for the automated identification of amphetamines from vapour-phase FTIR spectra. 2000. v. 404, p. 303–317.

RACHOW, T., BERGER, S., BOETTGER, M.K., SCHULZ, S., GUINJOAN, S., YERAGANI, V.K., VOSS, A., BÄR, K.J. Nonlinear relationship between electrodermal activity and heart rate variability in patients with acute schizophrenia. **Psychophysiology**, 2011. v. 48, n. 10, p. 1323–1332.

REBOLO, S., PEÑA, R.M., LATORRE, M.J., GARCÍA, S., BOTANA, A.M., HERRERO, C. Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. **Analytica Chimica Acta**, 2000. v. 417, n. 2, p. 211–220.

RODRIGUEZ-LUJAN, I., HUERTA, R., ELKAN, C., SANTA CRUZ, C. Quadratic Programming Feature Selection. **Journal of Machine Learning Research**, 2010. v. 11, p. 1491–1516.

RODRÍGUEZ-ROSARIO, C.A., MODI, K., KUAH, A., SHAJI, A., SUDARSHAN, E.C.G. Completely positive maps and classical correlations. **Journal of Physics A: Mathematical and Theoretical**, 2008. v. 41, n. 20, p. 205-301.

SILVA, A.C., LIRA PONTES, L.F.B., PIMENTEL, M.F., PONTES, M.J.C. Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods. **Talanta**, 2012. v. 93, p. 129–134.

SOUZA, L.M., RODRIGUES, R.R.T., SANTOS, H., COSTA, H.B., MERLO, B.B., FILGUEIRAS, P.R., POPPI, R.J., VAZ, B.G., ROMÃO, W. A survey of adulterants used to cut cocaine in samples seized in the Espírito Santo State by GC-MS allied to chemometric tools. **Science and Justice**, 2016. v. 56, n. 2, p. 73–79.

United Nations Office on Drugs and Crime (UNODC), World Drug Report, UNO, New York, USA, 2014.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: A basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, 2001. v. 58, n. 2, p. 109–130.

XIAOBO, Z., JIEWEN, Z., POVEY, M.J.W., HOLMES, M., HANPIN, M. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, 2010. v. 667, n. 1–2, p. 14–32.

XIE, L.; YING, Y.; YING, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. **Journal of Food Engineering**, 2009. v. 94, n. 1, p. 34–39.

YEH, Y.H., CHUNG, W.C., LIAO, J.Y., CHUNG, C.L., KUO, Y.F., LIN, T. Strawberry foliar anthracnose assessment by hyperspectral imaging. **Computers and Electronics in Agriculture**, 2016. v. 122, p. 1–9.

YU, H.; JIANG, S.; LAND, K. C. Multicollinearity in hierarchical linear models. **Social Science Research**, 2015. v. 53, p. 118–136.

YUSOFF, M.Z., CHANG, K.H., FAHMI, A., ABDULLAH, L. Attenuated total reflectance – Fourier transform infra-red spectral profiling of illicit heroin for forensic intelligence. **Australian Journal of Forensic Sciences**, 2017. v. 618, n. January, p. 1–9.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, 2013. v. 128, p. 17–24.

ZHAO, H., GUO, B., WEI, Y., ZHANG, B. Near infrared reflectance spectroscopy for determination of the geographical origin of wheat. **Food Chemistry**, 2013. v. 138, n. 2–3, p. 1902–1907.

4 Artigo 3 - Interval-based wavenumber selection framework for drug classification

Artigo a ser submetido no Forensic Science International (Elsevier, JCR 5yIF: 2.307, Qualis A2)

Abstract

The commerce of counterfeit drugs has substantially grown in years due to the ease access to the necessary technology for copying original pharmaceutical products. Attenuated Total Reflectance coupled with Fourier Transform Infrared (ATR-FTIR) spectroscopy has been widely employed as an efficient analytical tool to identify fraudulent medicines and help investigative forces to interrupt illegal operations. Despite the useful information obtained from ATR-FTIR, such data typically relies on hundreds of highly correlated wavenumbers which may jeopardize the performance of classical multivariate techniques tailored to sample analysis. This paper proposes a new wavenumber interval selection method aimed to select the region of the spectra that better inserts samples of seized drugs into two classes, i.e., original or counterfeit. For that matter, the Two Sample Kolmogorov Smirnov test statistic is estimated for each wavenumber, and an interval importance index is built to guide an iterative forward approach for wavenumber selection. At each iteration, one interval is added to the subset following the order suggested by the proposed index, and samples are classified towards different data mining techniques. The wavenumber subset yielding the best accuracy is chosen. In 100 replications using the adequate classification technique and interval size, the proposed method yielded average 99.87% accurate classifications on a Cialis® dataset, while retaining 12.5% of the original wavenumbers without variability on the selected subset; as for the Viagra® dataset, the method led to average 99.43% accurate categorizations with 23.75% of the original wavenumbers retained in the model. When compared to an individual wavenumber selection, the interval selection retained more consistent and easier to interpret wavenumber subsets.

Keywords: Interval Selection, Classification, ATR-FTIR, Cialis®, Viagra®

4.1 INTRODUCTION

Over the last years, the commerce of counterfeit medicines has increased worldwide due to the ease access of criminal organizations to the resources and technologies required to falsify original pharmaceuticals. Such commerce has also been prompted by online sales, which tend to difficult the investigation of illegal operations by police forces (Fernandez et al., 2011; Rodomonte et al., 2010; Sacré et al., 2011). Since the production of counterfeit medicines is not subjected to any mechanism of quality control, there are no guaranties about their composition, representing a serious risk to public health.

The falsification of phosphodiesterase type 5 (PDE-5) inhibitors for the treatment of erectile dysfunction is a particularly concerning problem in Brazil. The Brazilian Federal Police (BFP) reported 371 events in which forged PDE-5 inhibitors were apprehended between January 2007 and September 2010 (Anzanello et al., 2013). In light of that, the Fourier Transform Infrared (FTIR) (Soares et al., 2009) has been successfully coupled with the Attenuated total reflectance (ATR) and chemometrics techniques tailored to assess physical and chemical properties of seized Viagra® and Cialis® samples (Jung et al., 2012; Ortiz et al., 2012, 2013). Attenuated total reflectance (ATR) is a less expensive FTIR sampling technique that eliminates the use of solvents, and reduces the need for sample preparation (Grobério et al., 2015; Ortiz et al., 2013). ATR-FTIR has been widely used in many segments as food (Marcelo et al., 2014); fuels (Ferrão et al., 2011) and forensic areas (Grobério et al., 2015). In spite of providing useful information towards sample characterization, ATR-FTIR typically results in high dimensional databases with intrinsic multicollinearity and possible noise, which may jeopardize the prediction of a response variable with both classification and prediction purposes. To address such issue, two classical courses of action have been employed: (i) projection techniques, which combine all wavenumbers into new variables, and (ii) region selection techniques, which select a smaller subset from the set of original wavenumbers (Xiaobo et al., 2010).

Projections techniques (e.g. Principal Components Analysis and Partial Least Squares) transform the original set of wavenumbers into a subset of uncorrelated, allegedly more relevant variables for sample prediction or classification, overcoming the multicollinearity problem. Such data compression, however, may suppress important information from the original data, add bias to the model, and may not remove noisy spectral regions that yield poor classification results (Anzanello et al., 2014; Xiaobo et al., 2010). On the other hand, an efficient variable selection, also called wavenumber selection when applied to infrared spectroscopy data, can remove uninformative, noisy and redundant spectra regions, resulting on a smaller, easier to interpret model (Xie et al., 2009; Zhang et al., 2013). The wavenumber selection can be tailored to select wavenumber intervals (Marcelo et al., 2014; Soares et al., 2017) or individual wavenumbers (Anzanello et al., 2013; Kahmann et al., 2017).

This paper proposes a novel framework for identifying the most relevant spectral intervals tailored to improve the categorization of erectile dysfunction medicines into counterfeit or authentic classes. An interval selection method typically relies on combining spectral intervals, as in Soares et al. (2017), which may be computationally prohibitive in spectral data comprised of thousands of wavenumbers. To overcome such limitation, an Interval Importance Index (III) is proposed to guide the inclusion of the most relevant and informative wavenumber intervals in the analyzed subset. The suggested III relies on the two-sample Kolmogorov-Smirnov statistical test, which assesses the discriminant ability presented by each wavenumber to separate samples into classes; once that is assessed for each wavenumber, the III is created based on the importance of the wavenumbers inserted in that interval. Wavenumber subsets are then inserted into the set used for classification; the subset yielding the higher accuracy is then selected. We applied our propositions to two datasets comprised of authentic and forged samples; the Cialis® dataset consisted comprised of 300 samples, and the Viagra® had 177 samples. Using the recommended interval size and classification technique, we obtained 100% correct classifications in the training set, and average 99.5% accurate categorization in the testing set by retaining average 18.12% of the original 661 wavenumbers. When compared to an individual wavenumber selection, the interval selection identifies with more consistency the spectra area to be analyzed.

4.2 MATERIALS AND METHOD

4.2.1 Sample preparation

Eight authentic Cialis® tablets containing 20 mg of TAD, and six authentic Viagra® tablets containing 50 mg of SLD were supplied by Pfizer Ltda Laboratories were supplied by Eli Lilly do Brasil Ltda Laboratories. Twenty authentic Cialis® tablets (TAD, 20 mg) from eight distinct batches and nineteen authentic Viagra® tablets (SLD, 50 mg) from six distinct batches were purchased in local pharmacies. As for the counterfeit samples, one hundred and four tablets were sent to BPF (Porto Alegre, Rio Grande do Sul State) for forensic analysis through ATR-FTIR.

A Nicolet 380 FTIR Spectometer (Nicolet Instrument Co., Madison/ WI, USA) equipped with Deuterated Triglycine Sulphate detector and smart orbit single reflection diamond ATR sampling accessory was employed for all experiments. Spectra deriving from a small amount of sample positioned on the ATR crystal were measured, and the transmittance values then converted to absorption. No sample treatment was necessary for measurement; genuine and counterfeit tablets were prepared without their coats and homogenized by milling.

Next, a sample portion was directly placed on the ATR element. Each mixture was sampled 3 times, i.e., in triplicate. Identical pressure was used for all measurements. Each spectrum consists of 16 co-added scans measured at a spectral resolution of 4 cm^{-1} in the 4000–525 cm^{-1} range, yielding 661 wavenumbers. Spectral data were acquired with EZ OMNIC software, version 7.2a (Nicolet Instrument Co.). After measurement, the crystal was cleaned with acetone and dried in air ambient. Using the same instrumental conditions as the samples, an hourly background spectrum was obtained against air with clean and dry ATR element. No spectra pretreatments were performed. All spectra were saved in SPA format for works in EZ OMNIC and TQ Analyst EZ edition (Nicolet Instrument Co.) software.

4.2.2 Statistical and multivariate techniques

The two-sample Kolmogorov Smirnov (TSKS) test is a non-parametric test of equality of one-dimensional, continuous probability distribution. It aims to identify whether two samples originate from the same distribution without assuming any underlying parametric model for the samples (Mora-López and Mora, 2015). The TSKS statistic quantifies the maximum difference between the empirical distribution of both samples, as in equation (1).

$$\mathrm{D}_{\mathrm{KS}} = \max\{ \max_{1\leq \mathrm{i} \leq \mathrm{n1}} \left|F_{n1}^{1}\,(X_i^1\,) - F_{n2}^{2}\,(X_i^1\,)\right|, \max_{1\leq \mathrm{j} \leq \mathrm{n2}} \left|F_{n1}^{1}\left(X_j^2\,\right) - F_{n2}^{2}\left(X_j^2\,\right)\right|\} \quad (1)$$

where D_{KS} is the TSKS statistic, F^1 and F^2 are the continuous probability distributions and $\{X^1\}$ and $\{X^2\}$ are samples from the respective distributions. The domain of D_{KS} lies within the [0,1] interval; values closer to 1 suggest high separability between classes (Xiao, 2017).

We now present the fundamentals of the classification techniques tested in the proposed framework: k-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM). Such techniques are used due to their wide availability in statistical packages and suitability for ATR-FTIR analysis.

The first classification technique tested is the k-Nearest Neighbor (KNN), a non-parametric tool that classifies a new sample according to the majority class among the *k* nearest samples. The KNN stands out for its theoretical simplicity and for requiring a simple parameter, *k*, which can be defined by cross validation (Barbon et al., 2016; Rebolo et al., 2000). The second classification technique is the Linear Discriminant Analysis (LDA), which finds the hyperplane that maximizes the variance ratio between classes and within each class. Such hyperplane, which is defined by one or more discriminant functions comprised of the eigenvectors from the original data, is used to insert new samples into proper categories (Duda et al., 2001; Hastie et al., 2009). Support Vector Machine, the third tested technique, creates a linear hyperplane that maximizes the distance between the frontier samples of two categories, called as support vector. Similarly to LDA, the resulting hyperplane is used to categorize new samples. When non-linear problems are under analysis, kernel transformations can be used to find the proper hyperplane (Colman et al., 2015; Huang and Wang, 2006; Rakotomamonjy, 2003).

4.2.3 Framework for interval selection

There are four methodological steps in the proposed framework for interval selection aimed to categorize samples of erectile dysfunction medicines in original or counterfeit: (*i*) divide the original dataset samples in training and testing sets, and split the wavenumbers of the training set into equally sized (i.e. equidistant) intervals; (*ii*) create an Interval Importance Index (III) for each interval from (*i*) using the TSKS statistic to access the intervals with most dissimilarity between classes; (*iii*) iteratively classify the training set samples using a forward procedure according to the order suggested by the III, and retain the subset responsible for the highest classification accuracy; and (*iv*) classify the testing set using the retained intervals to determine the method accuracy. We now detail such steps.

In step (*i*), randomly split the *M* samples into two sets: a training set containing *TR* samples, and a testing set, containing *TS* samples, where *TR* + *TS* = *M*. The training set is used to select the most important intervals, while the testing set represents new samples to be classified by the selected model. The assessed proportion between *TR* and *TS* is 80%-20% (García Nieto et al., 2016), although other proportions could be tested. Next, split the *J* wavenumbers of the training set in *I* equally sized intervals.

In step (ii), apply the TSKS statistic [see equation (1)] to each wavenumber. The D_{KS} is then used to derive an Interval Importance Index for interval *i*, as in equation (2).

$$\mathrm{III}_i = \left(\sum_{j=LB_i}^{UB_i} D_{KS(j)}\right) / \left(\frac{J}{I}\right) \qquad (2)$$

where III_i is the importance index for interval *i*, with lower bound LB_i and upper bound UB_i. Note that the proposed III_i results from the average of the TSKS values inside each interval *i*. Different values for *I* are tested to assess the robustness of the method when different interval sizes are under analysis. Although *I* can reach the total number of wavenumbers (i.e., *I*=661), we recommend testing *I*=[2,4,…,64], similarly as in Ferrão et al. (2011), as small intervals of wavenumbers are normally more intuitive to be interpreted than isolated wavenumbers.

In step (*iii*), training samples described by the interval with the highest III are categorized into authentic and unauthentic classes applying the SVM; next, the interval with the second highest III is added to the dataset, and a new classification using both intervals is performed. This iterative procedure is repeated until all intervals have been inserted into the dataset according to the order suggested by the III; the subset yielding the highest accuracy is retained. In case multiple subsets of intervals yield the maximum accuracy, the subset comprised of the smallest number of retained intervals is chosen.

The chosen subset is then used in step (*iv*) to classify the testing set samples aimed at assessing the generalization capacity of the model. To verify the suitability of other classification techniques, we repeat steps (*iii*) and (*iv*) replacing SVM by KNN and DA, and compare their categorization performance. To avoid sampling bias in the analysis, 100 replications with different training and testing sets were performed for each classification technique.

4.3 Results

We now apply our propositions to the datasets described in section 2.1. Parameters for the classification techniques were defined by cross validation (Barbon et al., 2016; Zhang et al., 2015). All computational experiments were performed in Matlab® R2014a, using Statistics and Machine Learning Toolboxes.

4.3.1 Cialis® dataset

The Cialis® dataset is comprised of 300 samples (84 authentic and 216 counterfeit) described by 661 wavenumbers; the raw spectra are displayed in Figure 4-1.

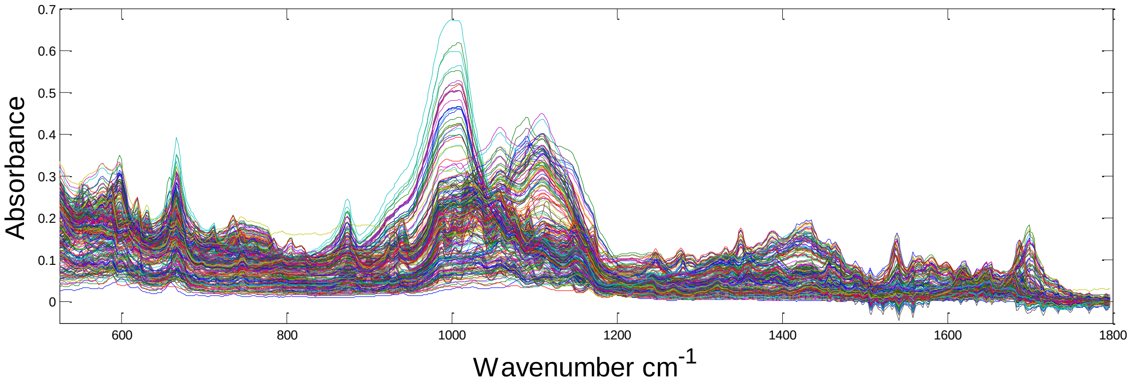


Figure 4-1 – Raw spectra of ATR-FTIR absorbance for 300 Cialis® samples

Table 4-1 depicts the classification accuracy on the training set and average percentage of retained wavenumbers after carrying out the selection procedure in section 4.2.3. The average percent of retained wavenumbers is calculated based on the number and size of retained intervals in relation to the 661 original wavenumbers for the 100 replications (e.g., for *I*=16, 12.5% of retained wavenumbers suggest that 2 intervals comprised of nearly 41 [=661/16] wavenumbers each were kept). SVM leads to the best results as it yields 100% accurate classifications from *I*=2 to *I*=64, and retains fewer wavenumbers when compared to the other classification techniques tested. LDA and KNN also lead to good classification accuracy when *I* increases, but the percentage of retained wavenumber is substantially higher than SVM. Figure 4-2 depicts the retained wavenumber intervals for SVM. The retention of two well-defined wavenumbers regions (i.e., 920-1000 cm^{-1} and 1560-1640 cm^{-1}) when *I*=16 is deemed the most consistent result as both regions were kept in all replications. When *I*>16, selected intervals tend to mostly spread around 1100, 1600 and 1800 cm^{-1}; there are also some intervals around 840 cm^{-1}. Such spreading can turn the interpretation of the retained regions less precise.

Training set average accuracy for Cialis® samples						
Classification Technique	Number of equidistant intervals (I)					
	2	4	8	16	32	64
KNN	0.9113 (100%)	0.9633 (98.5%)	0.9883 (92%)	0.9996 (74%)	1 (39.06%)	1 (19.53%)
LDA	0.8842 (100%)	0.9122 (96%)	0.9211 (63.75%)	1 (90.75%)	1 (45.37%)	1 (22.69%)
SVM	1 (50%)	1 (25%)	1 (12.75%)	1 (12.5%)	1 (8.5%)	1 (4.94%)

Table 4-1 – Training set average classification accuracy for Cialis® data set (% of retained wavenumbers in parenthesis)

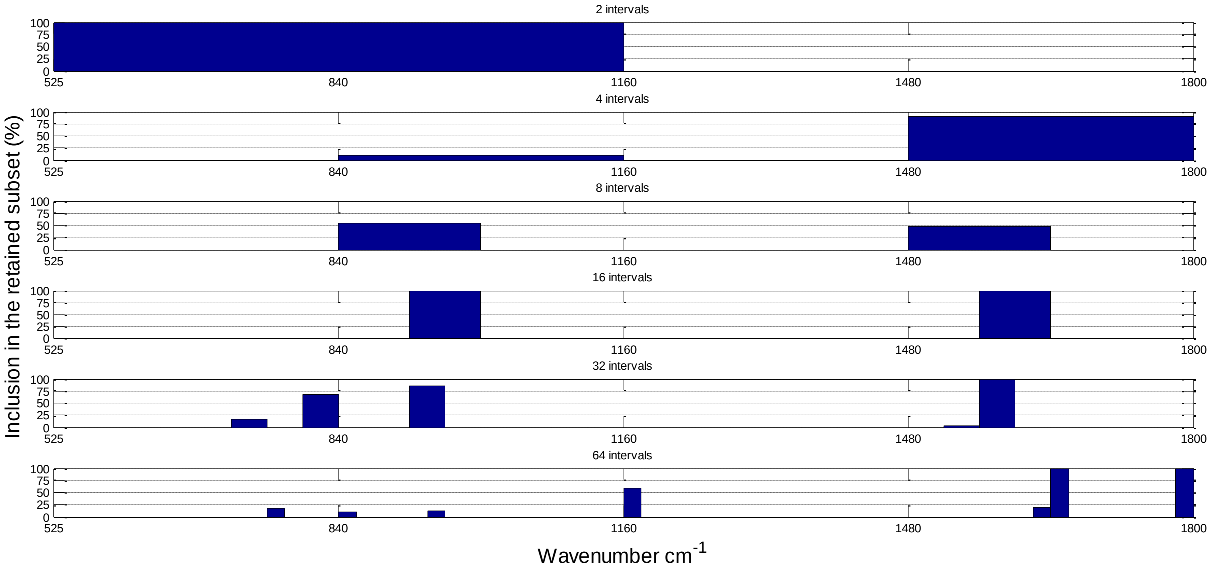


Figure 4-2– Retained intervals using SVM in Cialis® dataset

Table 4-2 depicts classification accuracy for Cialis® testing set. All classification techniques perform satisfactorily on the testing set independently of the number of intervals, especially for *I*=16. In light of that, we recommend applying SVM to intervals 920-1000 cm^{-1} and 1560-1640 cm^{-1} to classify Cialis® samples into original or counterfeit; such combination of classification technique and spectra regions yielded only 8 misclassifications out of 6000 testing samples.

Testing set accuracy for Cialis® samples						
Classification Technique	Number of equidistant intervals (I)					
	2	4	8	16	32	64
KNN	0.9992	0.9992	0.9992	0.9992	0.9992	0.9939
LDA	0.9979	0.9979	0.9984	0.9989	0.9971	0.9987
SVM	1	0.9936	0.9955	0.9987	0.9915	0.9904

Table 4-2 – Testing set classification accuracy for Cialis®

4.3.2 Viagra® dataset

The Viagra® dataset contains 102 counterfeit and 75 original samples, described by 661 wavenumbers. Figure 4-3 shows the raw spectra for the 177 Viagra® samples.

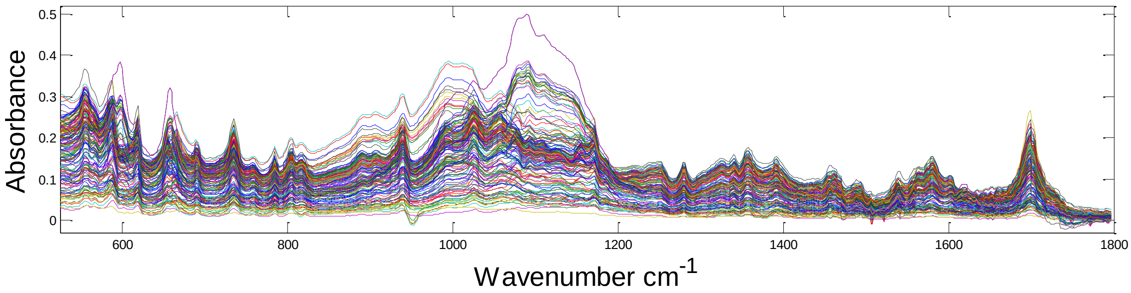


Figure 4-3 – Raw spectra of ATR-FTIR absorbance for 177 Viagra® samples

The average classification performance and percent of retained wavenumbers for different *I*‘ for the Viagra® training set are depicted in Figure 4-3. Once again, SVM correctly classifies all samples in the training set regardless of *I*. Despite yielding satisfactory results when *I* is increased, both KNN and LDA retain a substantially higher number of intervals when compared to SVM.

The retained intervals by SVM are depicted in Figure 4-4, which indicates 1240-1440 cm^{-1} as the most frequently selected interval. As the number of intervals increases, the retention of wavenumbers within the 680-840 cm^{-1} is replaced by wavenumbers near the 1780-1800 cm^{-1} interval.

Training set accuracy (% of retained variable in parenthesis)						
Classification Technique	Number of equidistant intervals (I)					
	2	4	8	16	32	64
KNN	0.8547 (61%)	0.8904 (95.5%)	0.9241 (89.5%)	0.9341 (64.12%)	0.9492 (78.13%)	0.9686 (65.5%)
LDA	0.7343 (68%)	0.8266 (89.5%)	0.9023 (83%)	0.9368 (94.63%)	1 (68.37%)	1 (34.19%)
SVM	1 (50%)	1 (25%)	1 (24.5%)	1 (23.75%)	1 (19.25%)	1 (14.94%)

Table 4-3 – Training set classification accuracy of Viagra® data set (% of retained variable in parenthesis)

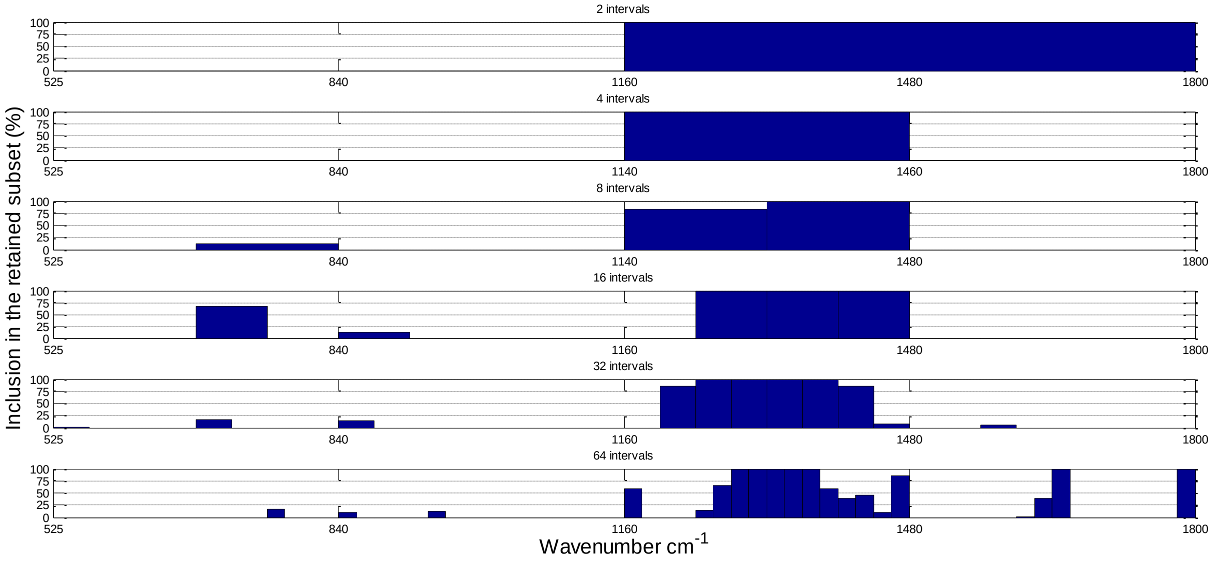


Figure 4-4 – Retained intervals using SVM in Viagra® dataset

As for the classification accuracy in the testing set (see Table 4-4 – Testing set classification accuracy of Viagra® data set), both SVM and LDA present higher accuracy than KNN. In terms of wavenumber retention, SVM needs fewer wavenumbers than LDA, becoming the recommended classification technique.

Testing set accuracy						
Classification Technique	Number of equidistant intervals (I)					
	2	4	8	16	32	64
KNN	0.9505	0.9645	0.9577	0.9341	0.9505	0.9468
LDA	0.9859	0.9923	0.9918	0.9950	0.9905	0.9714
SVM	0.9850	0.9891	0.9859	0.9943	0.9832	0.9745

Table 4-4 – Testing set classification accuracy of Viagra® data set

4.4 INDIVIDUAL SELECTION

When the number of intervals (*I*) approaches the number of wavenumbers (i.e. *I*→*J*), the interval selection becomes similar to an individual wavenumber selection. To compare such scenario with the interval selection, we performed an individual wavenumber selection (i.e. *I*=*J*) using the same training and testing sampling used on the interval approach.

4.4.1 Cialis®

The results of the individual wavenumber selection on Cialis® dataset is presented in Table 4-5, which shows that all classification techniques performed well in the training set when *I*=*J*. Classification performance in the testing set from both individual wavenumber selection and interval selection are very similar. Although LDA requires the smallest percent of wavenumbers (1.86%), the retained spectra presents some regions that were not consistently selected by the method (e.g., wavenumbers around 1500 and 1700 cm^{-1}) in Figure 4-5. Such sparse regions may offer additional complexity and non-straightforward interpretation of the relevance of those wavenumbers for sample discrimination; that situation is avoided when the proposed interval-based selection is carried out (TAN; LI, 2008; SONG et al., 2016).

Classification Technique	Training set accuracy	% of retained wavenumbers	Testing set accuracy
KNN	1	6.70%	0.9957
LDA	1	1.86%	0.9923
SVM	1	4.77%	0.9943

Table 4-5 – Classification results applying individual wavenumber selection to Cialis® dataset

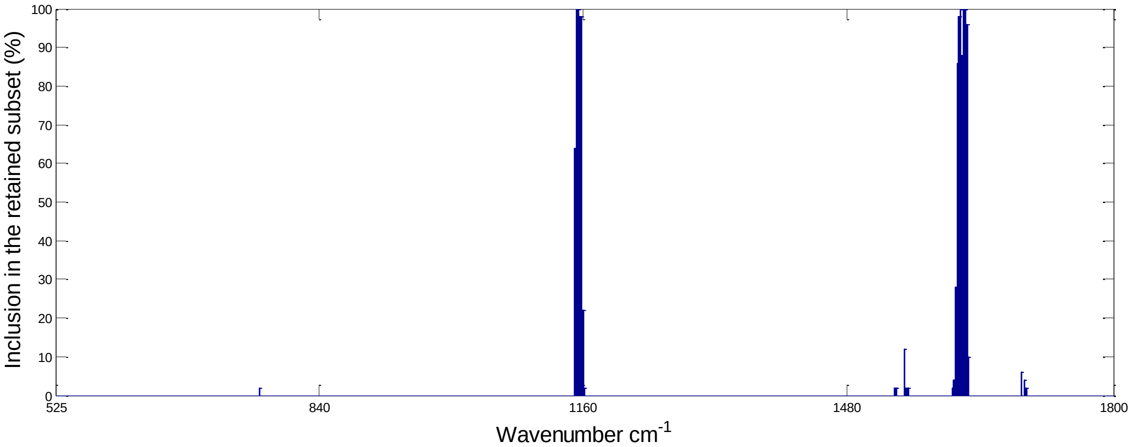


Figure 4-5 – Individual wavenumbers retained using LDA in Cialis® dataset

4.4.2 Viagra®

As for the individual wavenumber selection results for Viagra® dataset, LDA not only yields the better classification accuracy on the testing set, but also retains fewer wavenumbers when comparing to the other classification techniques; see Table 4-6. On the other hand, both SVM and KNN retain an elevated number of wavenumbers, indicating that the individual wavenumber selection is not an adequate approach when using such classification techniques coupled with the proposed III. Additionally, the interval selection approach seems to yield slightly higher accuracy than the individual wavenumber selection.

Figure 4-6 depicts the wavenumbers mostly retained by the LDA classification technique, which are located near 1280-1360 cm^{-1}. Once again, the individual-based selection approach yields some regions not consistently retained along the replications, which tend to compromise model interpretation (TAN; LI, 2008; SONG et al., 2016).

Classification Technique	Training set accuracy	% of retained wavenumbers	Testing set accuracy
KNN	0.9854	60.33%	0.9406
LDA	1	5.34%	0.9731
SVM	1	36.65%	0.9549

Table 4-6 – Classification results applying individual wavenumber selection for the Viagra® dataset

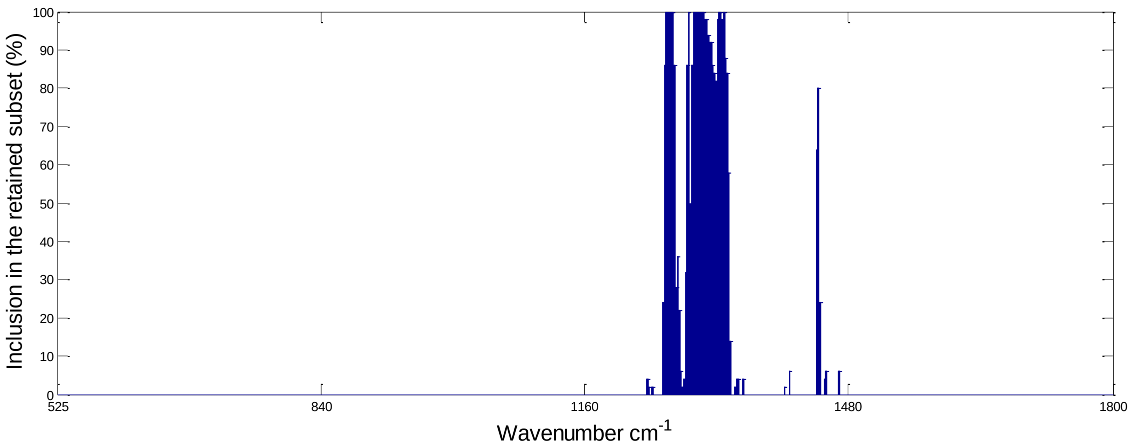


Figure 4-6 – Individual wavenumbers retained applying LDA to the Viagra® dataset

4.5 Conclusion

Although ATR-FTIR spectroscopy is deemed a powerful technique to detect fraudulent medicines, it typically leads to a dataset comprised of a large number of highly correlated and noisy wavenumbers that tend to reduce the performance of several multivariate techniques. Thus, the selection of the most relevant regions of the spectra is a relevant matter to properly discriminate original from counterfeit medicines.

The novelty of this paper relied on a new approach to select the most relevant spectra intervals for sample classification. The suggested framework derived an Interval Importance Index based on the Two Sample Kolmogorov Smirnov Test statistic. Three classification techniques were tested to verify its suitability in the proposed method. Using the recommended interval size, SVM provided the best result by correctly classifying 99.87% and 99.43% of Cialis® and Viagra® testing samples, respectively. The main advantage of the method relies on the easier interpretation of wavenumber intervals when compared to interpreting individual wavenumbers, especially when such individual wavenumbers are spread along the spectra.

Future researches include extending this method to non-equally sized intervals in order to create a framework that does not depend on a previously defined number of intervals. The integration of unsupervised exploratory techniques (e.g., clustering tools) to the proposed III aimed to find the intervals that better group unauthentic samples is also of interest from a forensic perspective.

4.6 REFERENCES

ANZANELLO, M.J., ORTIZ, R.S., LIMBERGER, R., MARIOTTI, K. Performance of some supervised and unsupervised multivariate techniques for grouping authentic and unauthentic Viagra® and Cialis®. **Egyptian Journal of Forensic Science**, 2014. v. 4, p. 83–89.

ANZANELLO, M. J., KAHMANN, A., MARCELO, M. C. A., MARIOTTI, K. C., FERRÃO, M. F., ORTIZ, R. S. Multicriteria wavenumber selection in cocaine classification. **Journal of Pharmaceutical and Biomedical Analysis**, 2015. v. 115, p. 562–569.

BARBON, A.P.A.C., BARBON, S., MANTOVANI, R.G., FUZYI, E.M., PERES, L.M., BRIDI, A.M. Storage time prediction of pork by Computational Intelligence. **Computers and Electronics in Agriculture**, 2016. v. 127, p. 368–375.

COLMAN, E., WAEGEMAN, W., DE BAETS, B., FIEVEZ, V. Prediction of subacute ruminal acidosis based on milk fatty acids: A comparison of linear discriminant and support vector machine approaches for model development. **Computers and Electronics in Agriculture**, 2015. v. 111, p. 179–185.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. **New York: John Wiley, Section**.

FERNANDEZ, F.M., HOSTETLER, D., POWELL, K., KAUR, H., GREEN, M.D., MILDENHALL, D.C., NEWTON, P.N. Poor quality drugs: grand challenges in high throughput detection, countrywide sampling, and forensics in developing countries. **Analyst**, 2011. v. 136, p. 3073–3082.

FERRÃO, M.F., VIERA, M.D.S., PAZOS, R.E.P., FACHINI, D., GERBASE, A.E., MARDER, L. Simultaneous determination of quality parameters of biodiesel/diesel blends using HATR-FTIR spectra and PLS, iPLS or siPLS regressions. **Fuel**, 2011. v. 90, n. 2, p. 701–706.

GARCÍA NIETO, P.J., GARCÍA-GONZALO, E., ARBAT, G., DURAN-ROS, M., RAMÍREZ DE CARTAGENA, F., PUIG-BARGUÉS, J. A new predictive model for the filtered volume and outlet parameters in micro-irrigation sand filters fed with effluents using the hybrid PSO-SVM-based approach. **Computers and Electronics in Agriculture**, 2016. v. 125, p. 74–80.

GROBÉRIO, T.S., ZACCA, J.J., BOTELHO, É.D., TALHAVINI, M., BRAGA, J.W.B. Discrimination and quantification of cocaine and adulterants in seized drug samples by infrared spectroscopy and PLSR. **Forensic Science International**, 2015. v. 257, p. 297–306.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research (JMLR)**, 2003. v. 3, n. 3, p. 1157–1182.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning. **Elements**, 2009. v. 1, p. 337–387.

HUANG, C. L.; WANG, C. J. A GA-based feature selection and parameters optimizationfor support vector machines. **Expert Systems with Applications**, 2006. v. 31, n. 2, p. 231–240.

JUNG, C.R., ORTIZ, R.S., LIMBERGER, R., MAYORGA, P.. A new methodology for detection of counterfeit Viagra® and Cialis® tablets by image processing and statistical analysis. **Forensic Science International**, 2012. v. 216, p. 92–96.

KAHMANN, A., ANZANELLO, M.J., MARCELO, M.C.A., POZEBON, D. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in Agriculture**, 2017. v. 140, p. 348-360.

MARCELO, M.C.A., MARTINS, C.A., POZEBON, D., FERRÃO, M.F. Classification of yerba mate (Ilex paraguariensis) according to the country of origin based on element concentrations. **Microchemical Journal**, 2014. v. 117, p. 164–171.

MORA-LÓPEZ, L., MORA, J.. An adaptive algorithm for clustering cumulative probability distribution functions using the Kolmogorov-Smirnov two-sample test. **Expert Systems with Applications**, 2015. v. 42, p. 4016–4021.

ORTIZ, R.S., MARIOTTI, K. C., FANK, B., LIMBERGER, R.P., ANZANELLO, M.J., MAYORGA, P. Counterfeit Cialis® and Viagra® fingerprinting by ATR-FTIR spectroscopy with chemometry: Can the same pharmaceutical powder mixture be used to falsify two medicines? **Forensic Science International**, 2013. v. 226, p. 282–289.

ORTIZ, R.S., MARIOTTI, K. C., LIMBERGER, R.P., MAYORGA, P. Physical profile of counterfeit tablets Viagra® and Cialis®. **Brazilian Journal of Pharmaceutical Science**, 2012. v. 48, p. 487–495.

RAKOTOMAMONJY, A. Variable Selection Using SVM-based Criteria. **Journal of Machine Learning Research**, 2003. v. 3, p. 1357–1370.

REBOLO, S., PEÑA, R.M., LATORRE, M.J., GARCÍA, S., BOTANA, A.M., HERRERO, C. Characterisation of Galician (NW Spain) Ribeira Sacra wines using pattern recognition analysis. **Analytica Chimica Acta**, 2000. v. 417, n. 2, p. 211–220.

RODOMONTE, A.L., GAUDIANO, M.C., ANTONIELLA, E., LUCENTE, D., CRUSCO, V., BARTOLOMEI, M., BERTOCCHI, P., MANNA, L., VALVO, L., ALHAIQUE, F., MULERI, N. Counterfeit drugs detection by measurement of tablets and secondary packaging colour. **Journal of Pharmaceutical and Biomedical Analysis**, 2010. v. 53, n. 2, p. 215-220.

SACRÉ, P.Y., DECONINCK, E., DASZYKOWSKI, M., COURSELLE, P., VANCAUWENBERGHE, R., CHIAP, P., CROMMEN, J., DE BEER, J.O. Impurity fingerprints for the identification of counterfeit medicines-A feasibility study. **Analytica Chimica Acta**, 2001. v. 701, n. 2, p. 224-231.

SOARES, F.; ANZANELLO, M. J.; MARCELO, M. C. A.; FERRÃO, M. F. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, 2017, v.167, n.15, p. 171-178.

SOARES, I.P., REZENDE, T.F., FORTES, I.C.P. Study of the Behavior Changes in Physical-Chemistry Properties of Diesel/Biodiesel (B2) Mixtures with Residual Oil and Its Quantification by Partial Least-Squares Attenuated Total Reflection-Fourier Transformed Infrared Spectroscopy (PLS/ATR-FTIR). **Energy & Fuels**, 2009. v. 23, p. 4143–4148.

SONG, X., YAN, Y.H.H., XIONG, Y., MIN, S. A novel algorithm for spectral interval combination optimization. **Analytica Chimica Acta**, 2016. v. 948, p.19-29.

TAN, C., LI, M. Mutual information-induced interval selection combined with kernel partial least squares for near-infrared spectral calibration. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, 2008. v. 71, n. 4, p. 1266-1273.

XIAO, Y. A fast algorithm for two-dimensional Kolmogorov-Smirnov two sample tests. **Computacional Statistics & Data Analysis**, 2017. v. 105, p. 53–58.

XIAOBO, Z., JIEWEN, Z., POVEY, M.J.W., HOLMES, M., HANPIN, M. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, 2010. v. 667, n. 1–2, p. 14–32.

XIE, L.; YING, Y.; YING, T. Classification of tomatoes with different genotypes by visible and short-wave near-infrared spectroscopy with least-squares support vector machines and other chemometrics. **Journal of Food Engineering**, 2009. v. 94, n. 1, p. 34–39.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and**

Intelligent Laboratory Systems, 2013. v. 128, p. 17–24.

ZHANG, Y., ZHENG, L., LI, M., DENG, X., JI, R. Predicting apple sugar content based on spectral characteristics of apple tree leaf in different phenological phases. **Computers and Electronics in Agriculture**, 2015. v. 112, p. 20–27.

5 Considerações finais

5.1 CONCLUSÕES

A presente tese tem como objetivo principal a proposição de novas metodologias para seleção de comprimentos de onda para aplicação em bancos de dados de espectroscopia no infravermelho oriundos de amostras de naturezas distintas. Nesta tese, as proposições são divididas em três artigos que propões novas abordagens de seleção de comprimentos de onda para a resolução de problemas específicos.

Dentre as principais contribuições do primeiro artigo, destaca-se a proposição de um novo índice de importância de comprimentos de onda, baseado em um problema de programação quadrática composto pelos valores da Informação Mútua entre os pares de comprimentos de onda e entre os comprimentos de onda e a variável resposta. Tal índice é utilizado em uma metodologia de seleção de comprimentos de onda voltada à categorização de amostras de erva mate de acordo com seu país de origem. Através da retenção média de 28% das variáveis originais foi possível categorizar 95,74% das amostras de teste, resultado superior quando comparados a outras metodologias.

Através de uma abordagem multicriterial para seleção de comprimentos de onda, o segundo artigo apresenta uma adaptação da metodologia proposta no primeiro artigo voltada à predição da concentração de cocaína e adulterantes em amostras de cocaína. Para a predição da concentração de cocaína, a regressão por mínimos quadrados parciais apresentou os melhores resultados, tendo um erro médio absoluto de 0,0879 alcançado através da retenção média de 2,03% dos comprimentos de onda originais. Por sua vez, a regressão linear múltipla apresentou os melhores resultados para a predição da concentração de adulterantes. Retendo em média 2,35% dos comprimentos de onda originais esta técnica atingiu um erro médio absoluto de 0,408.

Por fim, o terceiro artigo tem como principal contribuição a apresentação de uma nova metodologia para seleção de intervalos de comprimentos de onda, explorando a comparação de tal abordagem com a seleção de comprimentos de onda individuais. Utilizado para identificar falsificações de remédios para disfunção erétil, o método utiliza a estatística do teste de Kolmogorov-Smirnov para duas amostras para encontrar os intervalos do espectro com maior poder de separação entre as classes “original” e “falsificado”. Entre os bancos de dados analisados, a acurácia nas porções de teste foi de 99,65%, sendo necessária a retenção média de 18,12% do espectro original. Quando comparado à seleção individual de comprimentos de onda, a seleção de intervalos apresentou menor variabilidade dentre as faixas retidas do espectro.

Para atingir o objetivo principal, objetivos específicos foram determinados, os quais foram executados ao longo dos três artigos: dois novos índices de importância de comprimentos de onda foram propostos, indicando o cumprimento do primeiro objetivo específico; o segundo objetivo específico foi atingido no terceiro artigo, onde há a comparação entre métodos de seleção de comprimentos de onda individuais e de seleção de intervalos de comprimentos de onda; já a validação dos resultados dos artigos 1 e 3, através da comparação dos resultados aos resultados de outras metodologias demonstra a consecução do terceiro objetivo específico; por fim, a aplicação dos métodos em bancos de dados com diferentes origens e tipos de variáveis conduz ao quarto e último objetivo específico. Portanto, infere-se que todos os objetivos específicos determinados foram alcançados, permitindo igualmente afirmar que o objetivo principal deste trabalho foi obtido.

5.2 SUGESTÕES PARA TRABALHOS FUTUROS

Como possíveis extensões do estudo apresentado nesta tese, sugerem-se as seguintes frentes para pesquisas futuras:

a) Desenvolvimento de novas abordagens de análise multivariada voltadas à seleção de comprimentos de onda;

b) Abordagens para a identificação de observações críticas para a melhora do poder de categorização de amostras; e

c) Desenvolvimento de novos índices de importância voltados à análise de variáveis com diferentes características.