UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA FUNDAMENTAL

CAROLINA SANTOS SILVA

Espectroscopia no Infravermelho para Aplicações Forenses: documentoscopia e identificação de sêmen em tecidos

TESE DE DOUTORADO

Recife
2017

CAROLINA SANTOS SILVA

Espectroscopia no Infravermelho para Aplicações Forenses: documentoscopia e identificação de sêmen em tecidos

Tese apresentada ao Programa de Pós-graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco como requisito para a obtenção do título de Doutora em Química.

Orientadora:
Mª Fernanda Pimentel (UFPE, Brasil)
Co-orientadores:
Ricardo Saldanha Honorato (Polícia Federal, Brasil) e José Manuel Amigo (Universidade de Copenhague, Dinamarca)

Recife
2017

Catalogação na fonte
Bibliotecário Jefferson Luiz Alves Nazareno CRB 4- 1758

S586e Silva, Carolina Santos.
Espectroscopia no infravermelho para aplicações forenses: documentoscopia e identificação de sêmen em tecidos. / Carolina Santos Silva. – 2017
139 f. fig., tab.

Orientadora: Maria Fernanda Pimentel.
Tese (Doutorado) – Universidade Federal de Pernambuco. CCEN. Química Fundamental. Recife, 2017.
Inclui referências e apêndices.

1. Química analítica 2. Espectroscopia de infravermelho 3. Análise multivariada I. Pimentel, Maria Fernanda. (Orientadora) II. Título

543 CDD (22. ed.) UFPE-FQ 2017-58

CAROLINA SANTOS SILVA

Espectroscopia no Infravermelho para Aplicações Forenses: documentoscopia e identificação de sêmen em tecidos

Tese apresentada ao Programa de Pós-graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco como requisito para a obtenção do título de Doutora em Química.

Aprovada em: 31/07/2017

Banca Examinadora

_____

Prof. Maria Fernanda Pimentel
Universidade Federal de Pernambuco
Departamento de Engenharia Química

_____

Prof. Marcelo Martins Sena
Universidade Federal de Minas Gerais

_____

Prof. João Bosco Paraíso
Universidade Federal de Pernambuco
Departamento de Química Fundamental

_____

Prof. Claudete Fernandes Pereira
Universidade Federal de Pernambuco
Departamento de Química Fundamental

_____

Dr. André Filipe dos Ramos Martins Braz
Universidade Federal de Pernambuco
Departamento de Química Fundamental

Aos meus pais.

AGRADECIMENTOS

Em 2013, um artigo da *Nature* colocou em números a grande diferença de gênero existente na ciência ao redor do mundo. Não somente as diferenças salariais, mas também os constantes desafios enfrentados por mulheres na academia foram apontados como principais fatores da grande diferença entre o número de mulheres e homens na academia. Um dos fatores importantes citados na matéria é a ausência de mulheres ocupando cargos superiores que sirvam de modelo e inspiração para gerações mais novas. Neste contexto, eu início meus agradecimentos às principais mulheres que me serviram de modelo e inspiração, influenciando a minha jornada até aqui. Principalmente à minha orientadora e à minha mãe.

À professora Maria Fernanda Pimentel agradeço profundamente a orientação, o carinho e os conselhos. Você teve um papel fundamental na minha formação, me guiando desde a iniciação científica até minha formação como doutoranda. Mostrou diariamente a todos nós como criar um ambiente de trabalho em que todos se ajudam e discutem sobre os mais diferentes assuntos. Você é um exemplo de profissional que se envolve e se preocupa não somente com os assuntos acadêmicos, mas todo o contexto em que estamos inseridos e nosso papel na sociedade como acadêmicos. Pelo seu incentivo, apoio, conselhos, discussões e, principalmente, pela paciência, eu lhe agradeço.

Ao professor Jose Manuel Amigo, agradeço profundamente pela confiança. Seu entusiasmo nas discussões de trabalho sempre tão animador, capaz de levantar meu ânimo mesmo quando eu duvidava da minha capacidade. Pela parceria que vai além da sala de aula e chega até a mesa de bar, obrigada.

Ao perito e co-orientador Ricardo Honorato por todas as discussões, ideias e entusiasmo durante todo o período que trabalhamos juntos. Suas ideias sempre nos levaram a excelentes parcerias e seu bom humor e leveza nas discussões sempre foram grandes incentivadores.

Ao professor Peter Wentzell que me recebeu na Universidade de Dalhousie, pelas excelentes contribuições, paciência e atenção ao me ensinar novas técnicas. Fez muito mais que sua obrigação de supervisor durante minha estadia no Canadá. Agradeço.

Ao professor Célio Pasquini pelas contribuições e participação nos projetos desta tese.

Aos professores Carmen García-Ruiz e Fernando Ortega por terem me recebido na Universidade de Alcalá.

À Polícia Científica de Madrid por disponibilizar as amostras de Documentos utilizadas no projeto de datação.

Ao Professor André Mariano do ANDROLAB (UFRPE) por fornecer as amostras de sêmen utilizadas no projeto de fluidos.

Ao colega Fran pela sempre excelente parceria e discussões nos trabalhos realizados e pela ajuda na aquisição das imagens utilizadas. Também a Cristiane Vidal da Unicamp, pela ajuda também na aquisição das imagens.

À CAPES, pelas bolsas de doutorado e de doutorado sanduíche concedidas. Ao INCTAA, NUQAAPE e FACEPE pelo incentivo ao projeto, à UFPE pelo suporte institucional e ao Laboratório de Combustíveis (LAC) por proporcionar a efetivação da pesquisa.

A absolutamente todos os integrantes do LAC, por fazerem do ambiente de trabalho um lugar excelente de se trabalhar. Em especial àqueles que fazem parte do nosso dia a dia (Rafa, Jéssica, Paula, Fabrícia, Cláudio, Carol, Fernanda Honorato, Claudete e todos os outros por esses representados).

Aos amigos queridos Ali, Edu, Nei, Leandro, Sara e Vitor, por todas as discussões quimiométricas e não-quimiométricas, sempre regadas de muita parceria.

Aos meus pais, Ascendino e Fátima, agradeço profundamente por todo apoio, paciência e conselhos que me deram nessa jornada. Não só me incentivando e contribuindo para a minha formação pessoal, mas também profissional. Suas posições e o engajamento de ambos dentro do setor acadêmico e das políticas que o envolvem foram fundamentais para me ensinar o papel institucional e social da universidade. Vocês são meus principais influenciadores e devo a vocês cada conquista que tive até aqui. Por esses e por tantos outros motivos pessoais, obrigada.

A Danilo, pelo companheirismo, carinho, paciência, cafés, cervejas e conselhos de tantos anos.

A minha família, Eduardo, Julia, Renato, Dona Jovem, Tia Gó e Finha pelos conselhos, fofocas, momentos de trela, de diversão, vinho e cerveja.

Aos amigos inesquecíveis que fazem parte do Mammeta. A Rodrigo, Thalles, Tássia, Anaís e Juliana pela amizade ímpar.

A todos que contribuíram de alguma forma para esse trabalho, agradeço.

“Science and everyday life cannot and should not be separated”

Rosalind Franklin

RESUMO

A crescente necessidade de se estabelecer metodologias confiáveis e cientificamente embasadas de análise em laboratórios forenses levou a uma crescente demanda de estudos na área. Neste sentido, o presente trabalho propõe a união da espectroscopia na região de infravermelho e análise multivariada na solução de três diferentes problemas da ciência forense. Dentre eles: a diferenciação de tintas de canetas; a datação de documentos; e a identificação e diferenciação de manchas de sêmen humano em tecidos. Para o problema da diferenciação de tintas, técnicas não supervisionadas como Análise de Componentes Principais (PCA) e *Projection Pursuit* (PP), foram avaliadas. Modelos PP foram construídos utilizando a mínima curtose para encontrar agrupamentos de amostras que estivessem representando diferentes marcas de tintas de canetas. Para a construção de modelos PP, foi necessário utilizar uma ferramenta de redução de dimensionalidade, como PCA. Entretanto, o nível de compressão realizado pode afetar os modelos PP e, para contornar esse problema, a Análise de Procusto foi utilizada para monitorar a estabilidade dos modelos PP em diferentes níveis de compressão de forma que ainda fossem capazes de fornecer projeções de interesse. Os mapas de Procusto produzidos apresentaram uma região de estabilidade para aquelas projeções que apresentaram estruturas semelhantes e informativas, capazes de identificar o nível de compressão adequado para os conjuntos estudados. Ainda na área de documentoscopia, diferentes documentos naturalmente envelhecidos foram estudados empregando Infravermelho Médio para construir modelos de regressão para datação dos mesmos. Diferentes técnicas de pré-processamento como Mínimos Quadrados Generalizados Ponderados (GLSW) e Correção Ortogonal de Sinais (OSC) foram utilizadas no sentido de atenuar a variabilidade de documentos de mesma idade. Além disso, a técnica de Mínimos Quadrados Parciais Esparsos (sPLS) foi aplicada para avaliar o potencial de seleção de variáveis na datação de documentos. Os pré-processamentos apresentaram importante influência nos resultados, minimizando as diferenças entre documentos de um mesmo ano e fornecendo modelos de regressão mais relacionados com a idade do documento. Já na área de identificação de fluidos biológicos, manchas de sêmen em diferentes tecidos absorventes foram analisadas e metodologias presuntivas e confirmatórias foram propostas. Para a abordagem presuntiva, modelos PCA e de Resolução Multivariada de Curvas com Mínimos Quadrados Alternados (MCR-ALS) foram propostos para identificar possíveis manchas de sêmen em diferentes tecidos. Em seguida, modelos classificatórios foram propostos como abordagem confirmatória, de forma a viabilizar a diferenciação de sêmen e falsos-positivos. Os modelos construídos como abordagem presuntiva identificaram as manchas independentemente do tecido utilizado, porém os modelos PCA mostraram grande influência da textura do substrato. Já os modelos classificatórios apresentaram resultados adequados de sensibilidade e especificidade, e apenas uma abordagem não apresentou falsos negativos para sêmen. Com os resultados obtidos, foi possível observar o grande potencial das metodologias analíticas, mais especificamente a espectroscopia no Infravermelho, associadas à análise multivariada na resolução de problemas de natureza forense.

Palavras-chave: Química Analítica. Espectroscopia no Infravermelho. Análise Multivariada.

ABSTRACT

The growing need to establish reliable and scientific based methodologies for analysis in forensic laboratories led to a high demand for studies in the field. In this sense, the present work proposes the association of infrared spectroscopy and multivariate analysis to solve three different problems of forensic science. Among them: discrimination of pen inks from different types and brands by means of unsupervised techniques; a document dating problem through regression models; and the identification and differentiation of semen stains on fabrics using classification techniques and hyperspectral images. Regarding the ink discrimination problem, unsupervised techniques such as Principal Component Analysis (PCA) and Projection Pursuit (PP) were evaluated. PP models were built by means of kurtosis minimization in order to find clusters of samples that represent different brands of pen inks. To build PP models, it is required a smaller number of variables than samples and therefore a technique for variable compression, such as PCA, is needed prior the PP models. However, the level of compression can affect the projections from PP models, therefore, Procustes Analysis was used to monitor the stability of PP models in order to verify wether the different levels of compressions are still able to provide projections of interest. The Procrustes maps obtained showed a stability region to those projections which provided similar and informative structures, which ones were able to indicate the adequate level of compression to the dataset. Still in documentoscopy, different natural aged documents were evaluated with Mid Infrared spectroscopy to build regression models for dating purposes. Different preprocessing techniques such as Generalized Least Squares Weighting (GLSW) and Orthogonal Signal Correction (OSC) were applied in order to attenuate the variability within documents from the same age. In addition, Sparse Partial Least Squares technique (sPLS) was applied to evaluate the potential of variable selection for dating documents. Preprocessing step showed to be the most important step of the analysis, minimizing the differences between documents from the same year and providing regression models related to cellulose changes. In the field of identification of biological fluids, semen stains in different absorbent fabrics were analyzed and presumptive and confirmatory methodologies were proposed. As a presumptive approach, PCA and Multivariate Curve Resolution – Alternate Least Squares (MCR-ASL) models were proposed to identify the possible presence of semen stains on different fabrics. Afterwards, classification models were proposed as a confirmatory approach, in order to differentiate semen from false-positives. The presumptive models built identified the stains independently of the fabric used, however the PCA models showed high influence from the texture. The classification models provided adequate results for sensitivity and specificity, and only one approach did not show any false negative cases for semen. According to the results obtained, it was possible to state the great potential of analytical methodologies, more specifically the infrared spectroscopy, associated with the multivariate analysis, to solve forensic problems.

Keywords: Analytical Chemistry. Infrared Spectroscopy. Multivariate Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

%CC	% de Classificação Correta
ALS	Mínimos Quadrados Alternados
ATR	Refletância Total Atenuada
CLS	Mínimos Quadrados Clássicos
DA	Análise Discriminante
DP	Grau de Polimerização
FN	Falso-negativo
FoM	Figuras de Mérito
FP	Falso-positivo
FTIR	Infravermelho com Transformada de Fourier
GLSW	Mínimos Quadrados Generalizados Ponderados
HCA	Análise de Agrupamentos Hierárquicos
HSI	Imagem Hiperespectral
IR	Infravermelho
LDA	Análise Discriminante Linear
LV	Variáveis Latentes
MCR	Resolução Multivariada de Curvas
MIR	Infravermelho Médio
MSC	Correção Multiplicativa de Sinais
NIPALS	Mínimos Quadrados Parciais Iterativos Não Linear
NIR	Infravermelho Próximo
OSC	Correção Ortogonal de Sinais
PCA	Análise de Componentes Principais
PCR	Regressão por Componentes Principais
PLS	Mínimos Quadrados Parciais
PP	*Projection Pursuit*
RMSE	Raiz do Erro Quadrático Médio

RMSEC	Raiz do Erro Quadrático Médio de Calibração
RMSECV	Raiz do Erro Quadrático Médio de Validação
RMSEP	Raiz do Erro Quadrático Médio de Previsão
ROC	Característica de Operação do Receptor
ROI	Região de Interesse
SIMCA	Modelagem Independente e Flexível por Analogia de Classe
sLV	Variáveis Latentes Esparsas
Sn	Sensibilidade
SNV	Padronização Normal de Sinal ou Variação Normal Padrão
Sp	Especificidade
sPLS	Mínimos Quadrados Parciais Esparsos
SV	Vetores de Suporte
SVD	Decomposição por Valores Singulares
SVM	Máquinas de Vetores de Suporte
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
VIP	Importância das Variáveis na Projeção

SUMÁRIO

1 INTRODUÇÃO E FUNDAMENTAÇÃO TEÓRICA

1.1 INTRODUÇÃO

Problemas provenientes de análises de pouca confiabilidade ou má conduta de peritos podem ser evitados utilizando metodologias objetivas que não dependam apenas da experiência e opinião do investigador forense. Em 2014, Widner e colaboradores publicaram um artigo sobre os problemas enfrentados em muitos laboratórios de polícia científica nos Estados Unidos, onde a má conduta de analistas e análises não confiáveis de evidências foram encontradas levando, muitas vezes, à punição severa de inocentes (WIDENER; DRAHL, 2014). Este artigo enfatiza a necessidade de validar e estabelecer metodologias que devem seguir determinados procedimentos padrões e guias advindos de um comitê superior formado por cientistas qualificados. Neste contexto, não só metodologias analíticas confiáveis são úteis na solução de problemas forenses, mas também a atribuição de confiança estatística a uma análise de evidências se tornaram fundamentais para garantir que conclusões e inferências corretas podem ser realizadas a partir dessas análises.

Métodos objetivos de análises em cenas de crime são especialmente relevantes se são rápidos, não destrutivos, não invasivos e portáteis. Atualmente, técnicas de análise da área forense realizadas em campo, além de apresentarem falsos positivos ou negativos podem ser laboriosas e destrutivas. Para contornar esses problemas, métodos analíticos empregando espectroscopia vibracional são potenciais alternativas (MURO et al., 2015), tais como as espectroscopias Raman e Infravermelho (IR: *Infrared*). Identificação de fibras (THOMAS et al., 2005), detecção de cocaína em saliva (HANS et al., 2014), identificação de resíduos de explosivos (BANAS et al., 2010), visualização de impressões digitais (CRANE et al., 2007), discriminação de tintas de canetas e papéis (KHER et al., 2006; KUMAR; KUMAR; SHARMA, 2017; SILVA et al., 2012), diferenciação de fluidos biológicos (SIKIRZHYTSKI; VIRKLER; LEDNEV, 2010) e identificação de pigmentos em obras de arte (CASADIO et al., 2010) são exemplos do quanto as técnicas de Raman e IR têm atendido às demandas da área forense nos últimos anos.

Embora venha ganhando cada vez mais aplicações, uma das maiores desvantagens de instrumentos portáteis de espectroscopia Raman é que suas análises são pontuais e a conversão dos dados para um mapeamento de superfícies

relativamente grandes pode ser demorada. Neste cenário, a espectroscopia no IR aparece como uma alternativa a ser explorada.

A espectroscopia na região do MIR é geralmente rápida e fornece picos mais definidos e fáceis de interpretar do que a região NIR. Entretanto, a região NIR possui a grande vantagem de ser extremamente versátil e os equipamentos são mais facilmente miniaturizados. Além disso, os equipamentos NIR são mais baratos do que equipamentos MIR e Raman e podem ser facilmente implementados em sistemas de imagens capazes de varrer grandes áreas (GRIFFITHS, 2009). Além disso, as espectroscopias NIR e MIR, assim como a espectroscopia Raman, são técnicas analíticas, que quando associadas a um tratamento quimiométrico dos dados, são capazes de fornecer resultados objetivos que podem ser utilizados para fins forenses sem depender somente da opinião do analista.

Geralmente, as técnicas espectroscópicas fornecem uma grande quantidade de dados, principalmente quando se trata de sistemas de imagens. Particularmente, a espectroscopia NIR fornece espectros que são resultado de sobreposição de bandas, o que dificulta o entendimento do conjunto de dados, caso não haja o tratamento multivariado. Diferentes abordagens podem ser adotadas na análise dos espectros de infravermelho, dependendo do objetivo final da análise. Neste sentido, técnicas de calibração e reconhecimento de padrões (supervisionadas e não supervisionadas) podem ser empregadas, de acordo com o problema proposto.

1.2 OBJETIVOS E METAS GERAIS

Este trabalho possui como principal objetivo propor métodos analíticos empregando a espectroscopia na região do infravermelho e técnicas quimiométricas para solução de problemas na área da química forense, envolvendo a documentoscopia e a identificação de resíduos de sêmen em tecido.

1.3 FUNDAMENTAÇÃO TEÓRICA

1.3.1 Espectroscopia no Infravermelho

A espectroscopia na região do Infravermelho (IR: *Infrared*) é uma técnica de espectroscopia vibracional baseada em absorção molecular em que a energia, quando absorvida por uma determinada molécula, promove transições vibracionais e

rotacionais. Essa região espectral pode ser subdividida em três: (Figura 1), infravermelho próximo, NIR (NIR: *Near Infrared*), infravermelho médio, MIR (MIR: *Middle Infrared*), e distante (FAR: *Far Infrared*) que estão contidas na faixa de 12800-10 cm^{-1} (SKOOG; HOLLER; CROUCH, 2009). Em particular, as regiões MIR e NIR, que serão abordadas neste trabalho, fornecem informações a respeito de vibrações fundamentais, sobretons e combinações de vibrações fundamentais.

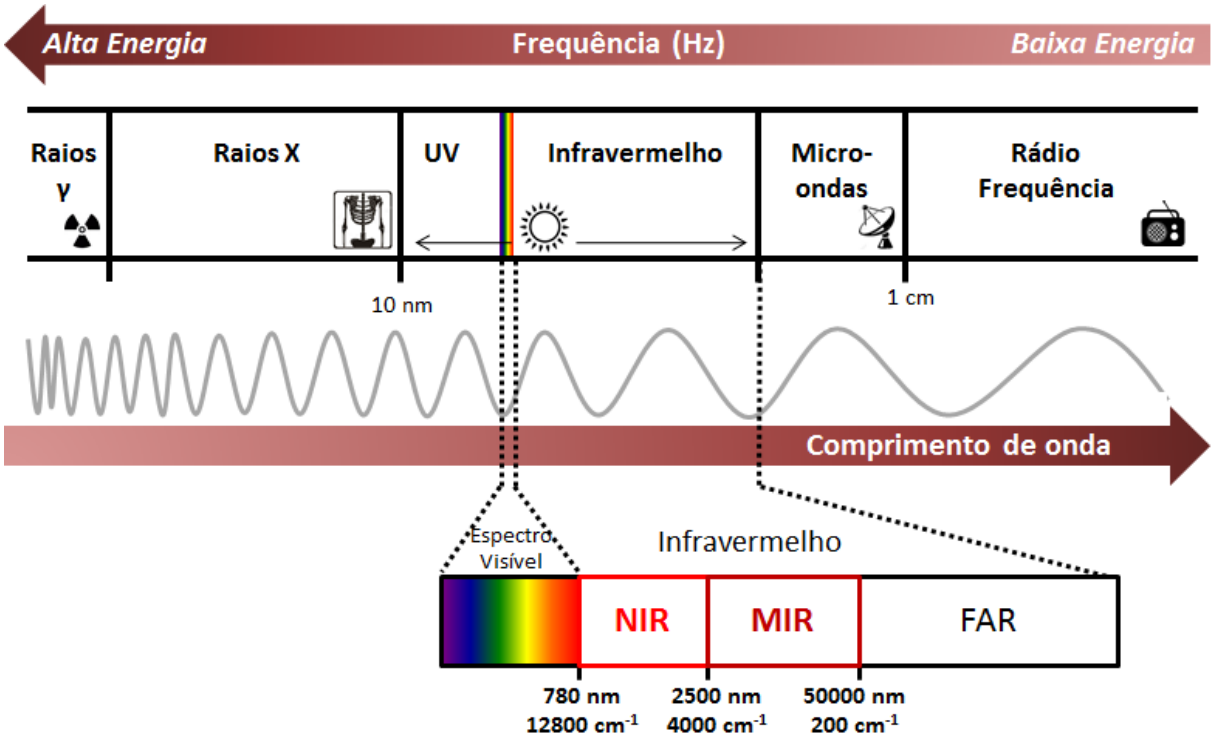

Figura 1 Espectro eletromagnético. Em destaque a região do visível ao Infravermelho (ESKILDSEN, 2016).

A grande vantagem da espectroscopia na região do IR reside no fato de que essa técnica analítica é extremamente versátil, capaz de analisar amostras nos três estados da matéria (gasoso, líquido e sólido) com o mínimo (ou nenhum) preparo de amostra. Após os cromatógrafos, os espectrômetros de IR são os equipamentos de maior procura para aplicação em indústrias, devido à facilidade de implementação, rapidez na aquisição de espectros e custo relativamente baixo no que diz respeito ao suporte e manutenção (COATES, 2010).

A região MIR é a região espectral compreendida entre 5000 a 400 cm^{-1}, e geralmente o espectro é expresso em unidades de número de onda, enquanto na região NIR, o espectro é, em geral, expresso em unidades de comprimento de onda e

a região é definida entre a região 750 a 2500 nm, aproximadamente. Enquanto a região MIR é capaz de fornecer informações diretas a respeito dos grupos funcionais, na região NIR a presença desses grupos só pode ser identificada por inferência, devido à grande complexidade dessa região espectral (COATES, 2010). Além disso, as absorções na região do MIR apresentam, em geral, maior intensidade, pois são provenientes das transições fundamentais, enquanto que as vibrações na região NIR fornecem sinais mais fracos (GRIFFITHS, 2009).

Outra vantagem do NIR é o fato de que seus equipamentos são geralmente mais simples, o que leva a um relativo baixo custo, robustez e facilidade em ser miniaturizado. De qualquer forma, a escolha do método de análise sempre dependerá das amostras e aplicações, e levando em conta as vantagens e desvantagens de cada técnica, o analista deverá decidir qual se adequa melhor aos seus propósitos.

1.3.1.1 Fundamentos da Espectroscopia IR

Qualquer molécula cuja temperatura seja maior do que o zero absoluto irá apresentar movimentos vibracionais e rotacionais. Quando as frequências desses movimentos são iguais às frequências da radiação incidente, a molécula pode absorver energia. A absorção só ocorre quando os movimentos em questão ocasionarem numa variação do momento de dipolo da molécula, resultando em uma variação na amplitude dos movimentos de estiramentos e deformações moleculares (Figura 2) (SKOOG; HOLLER; CROUCH, 2009).

Modelo do oscilador harmônico: Inicialmente, o modelo harmônico foi proposto para entender as absorções moleculares. Esse modelo pode ser entendido considerando que a ligação entre átomos em uma molécula diatômica heteronuclear é análoga a uma mola unindo duas massas (átomos), cuja constante de força está relacionada com a força da ligação entre os átomos da molécula (Figura 3). De acordo com esse modelo, a frequência de vibração dependerá das massas dos átomos envolvidos na ligação e da constante de força da ligação que os une. Neste modelo, as transições são somente permitidas entre níveis de energias adjacentes, explicando apenas os modos de vibração fundamentais. Assumindo que, em condições ambiente, a maioria das moléculas ocupam os estados fundamentais de energia, apenas as transições para o primeiro estado excitado serão permitidas (n=0 para n=1). Desta forma os espectros representarão, majoritariamente, absorções dos modos vibracionais e uma

vez que a maioria dos compostos orgânicos apresentam bandas de absorção na região de 4000-200 cm^{-1}, a região do MIR é considerada extremamente sensível às estruturas moleculares sendo comumente aplicada na identificação de funções orgânicas características do composto analisado (BURNS; CIURCZAK, 2009).

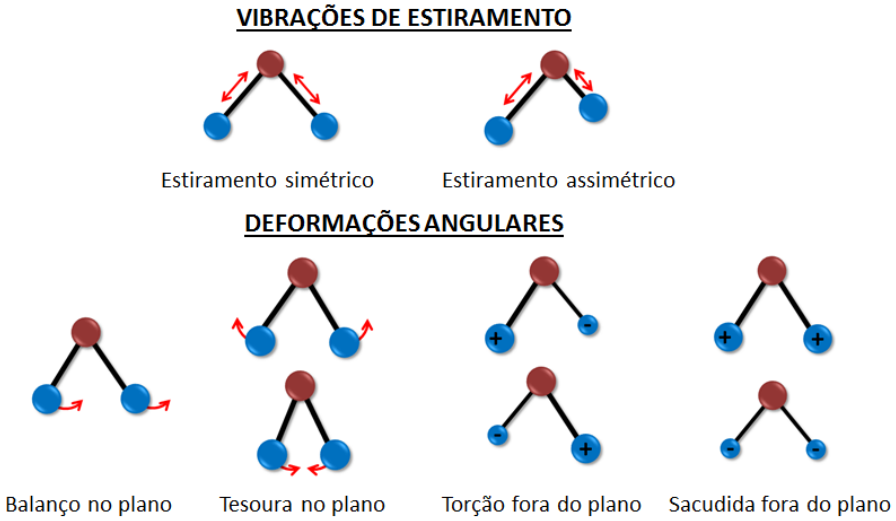

Figura 2 Movimentos de estiramento e deformações moleculares que são ativos no IR. Adaptado de SKOOG, 2009.

Modelo do oscilador anarmônico: embora o modelo de oscilação harmônica seja bem-sucedido ao explicar as transições fundamentais, é preciso considerar que existem forças repulsivas entre os átomos envolvidos numa determinada ligação e que, além disso, existe a possibilidade de dissociação de uma ligação, caso um determinado estiramento exceda um limite. Neste caso, as energias das ligações em sistemas moleculares irão obedecer ao comportamento representado pelo modelo de oscilação anarmônica. De acordo com o modelo em questão, a energia potencial vibracional não varia periodicamente com a variação da distância entre dois núcleos, como sugere o modelo harmônico. Na medida em que dois núcleos se aproximam, a repulsão entre suas nuvens eletrônicas aumenta no sentido de restaurar a distância da ligação. Por outro lado, quando os núcleos se distanciam, há um aumento da energia potencial até o momento em que ocorre a dissociação da ligação (Figura 3). Para esse modelo, as transições ativas não só obedecem à regra de seleção $\Delta u=\pm1$ (relativa aos modos normais de vibração), mas também às regras de seleção $\Delta u=\pm2$ e $\Delta u=\pm3$, que explicam os sobretons e as bandas de combinação (PASQUINI, 2003).

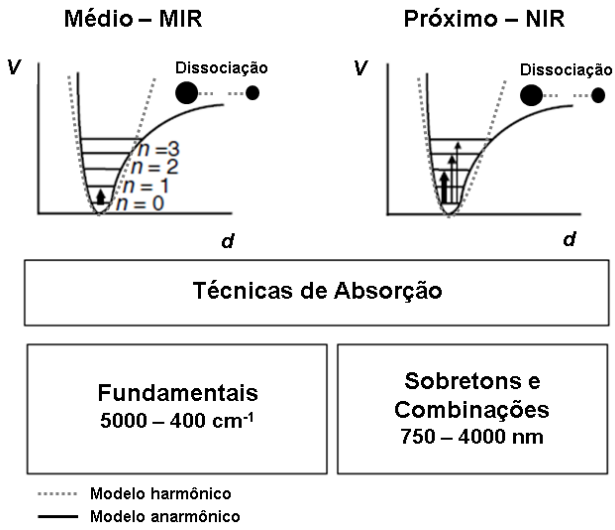

Figura 3 Gráfico das energias potenciais (V) da ligação em função da distância (d) entre os átomos para o modelo do oscilador anarmônico (BURNS; CIURCZAK, 2009)**.**

1.3.1.2 Espectroscopia MIR usando acessório de ATR

Na região MIR, o acessório de Refletância Total Atenuada (ATR: *Attenuated Total Reflectance*) é bastante utilizado para análise de amostras, principalmente sólidas. A aquisição de espectros ocorre quando um feixe de radiação infravermelha incide em um cristal com um ângulo crítico específico, de forma que esse feixe é totalmente refletido no interior do cristal (Figura 4). Nesse processo de reflexão, a radiação penetra na amostra com uma profundidade muito pequena, mas suficiente para fornecer informações a respeito da composição química da amostra analisada. Essa radiação penetrante é conhecida como onda evanescente (SKOOG; HOLLER; CROUCH, 2009). É necessário garantir o total contato da amostra analisada com o cristal do acessório, devido à pequena penetração do feixe na amostra, e por isso, uma pressão é exercida para garantir o melhor contato possível na interface amostra/cristal.

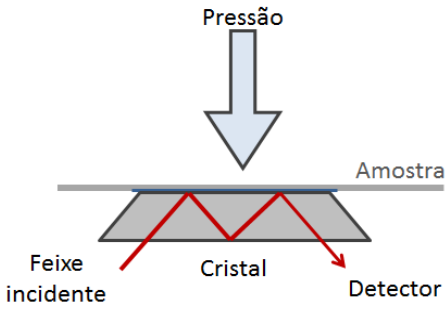

Figura 4 Esquema de aquisição espectral utilizando o acessório de ATR.

1.3.1.3 Imagens Hiperespectrais

Outro método importante de aquisição de espectros são as chamadas Imagens Hiperespectrais (HSI: *Hyperspectral Image*). As imagens digitais são formadas por unidades menores definidas por coordenadas espaciais e uma informação que está relacionada com o(s) canal(is) de cores; as diferentes intensidades de cor para cada unidade dá ao observador as sensações de texturas, sombras, brilhos, etc, que compõem a imagem final. Essas unidades menores podem ser chamadas de *pixels*.

Numa imagem digital na escala de cinza, por exemplo, cada pixel está associado a um valor de intensidade nessa escala, que, quando juntos, são capazes de formar uma imagem. Com o avanço da tecnologia e a capacidade de gerar imagens digitais coloridas, os pixels passaram a representar a combinação de diferentes canais de cores, como CMYK (ciano, magenta, amarelo e preto) e RGB (vermelho, verde e azul). Para imagens digitais em RGB, por exemplo, cada pixel possui então três valores numéricos associados a diferentes intensidades nas escalas de vermelho, verde e azul (GRAHN; GELADI, 2007). Assim, as imagens digitais podem ser interpretadas como matrizes de dados, de forma que, quando em escala de cinza, tem-se uma matriz simples bidimensional com coordenadas espaciais x e y (Figura 5a); em RGB, tem-se uma matriz tridimensional em que z é a dimensão dos diferentes canais de cores (Figura 5b) (PRATS-MONTALBÁN; DE JUAN; FERRER, 2011).

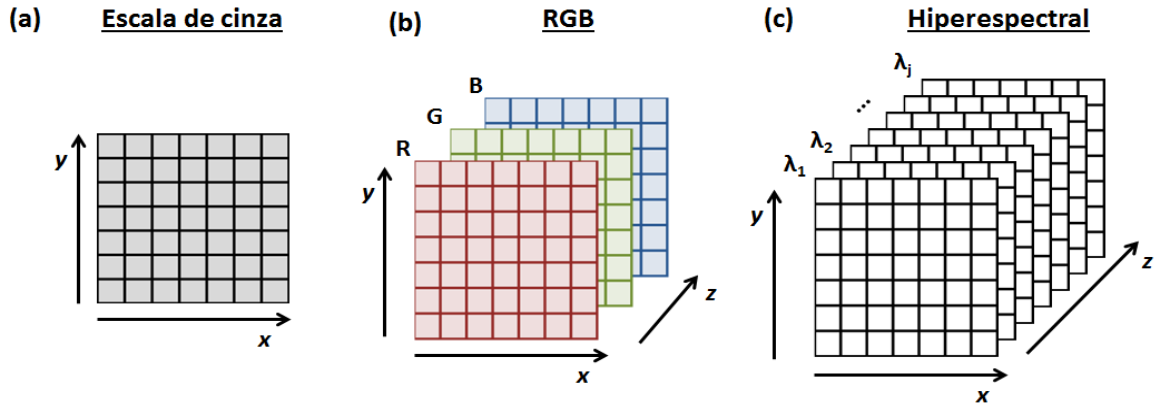

Figura 5 Matriz de dados de imagens (a) em escalas de cinza, (b) em RGB e (c) hiperespectrais.

Além das imagens em escala de cinza e em canais de cores, existe um tipo de imagem muito particular capaz de fornecer informações sobre a composição química de acordo com as coordenadas espaciais. São as chamadas Imagens Hiperespectrais. Dessa forma, cada pixel estará associado a um espectro (Figura 5c), que pode ser adquirido a partir de diversas técnicas analíticas, como espectroscopia Raman, Infravermelho, etc. A matriz de dados de uma imagem hiperespectral é chamada de hipercubo e resulta da concatenação de mapas de absorbâncias para cada comprimento de onda λ_i (DE JUAN et al., 2009).

Para transformar as imagens hiperespectrais em um conjunto de dados que possa ser facilmente manipulado, é preciso realizar um desdobramento do hipercubo para gerar uma matriz de dados bidimensional. Dessa forma, cada pixel será considerado como uma amostra, formando uma matriz de dados clássicos, como sugerido na Figura 6.

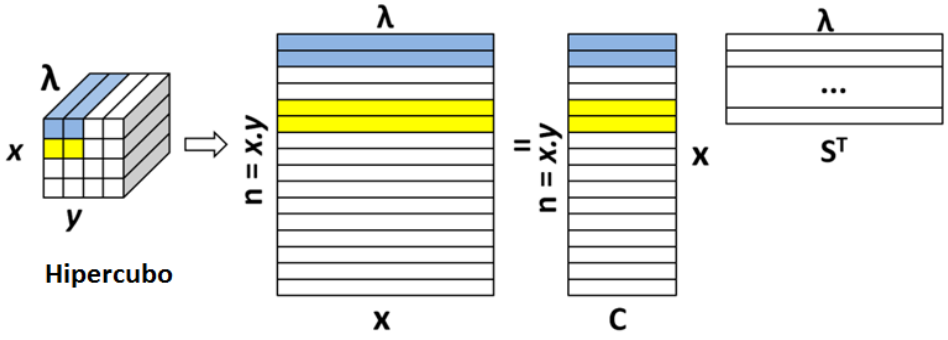

Figura 6 Desdobramento da matriz tridimensional de dados em uma matriz bidimensional e sua decomposição em perfis de concentração relativa e espectros puros.

A combinação dos sistemas de imagens com as técnicas espectroscópicas formam uma poderosa ferramenta, pois fornecem não só informações a respeito da composição química de uma amostra, mas também a distribuição espacial dos compostos nesta amostra (AMIGO; BABAMORADI; ELCOROARISTIZABAL, 2015; ESBENSEN; GELADI, 1989).

Os sistemas de imagem NIR (HSI-NIR: *Hyperspectral Images – Near Infrared*) são, em geral, robustos e versáteis e podem ser adaptados para atender às exigências de portabilidade necessárias para aplicações forenses (DE LA OSSA; AMIGO; GARCÍA-RUIZ, 2014; EDELMAN et al., 2013; ISHIKAWA et al., 2013; PAYNE et al., 2005).

As HSI geralmente fornecem uma grande quantidade de dados, o que torna seu entendimento laborioso sem o uso de metodologias analíticas multivariadas. A adoção dessas ferramentas estatísticas permitiu o desenvolvimento de estudos mais completos das imagens hiperespectrais, que podem ser realizados com diferentes abordagens (AMIGO; BABAMORADI; ELCOROARISTIZABAL, 2015; PRATS-MONTALBÁN; DE JUAN; FERRER, 2011)

1.3.2 Quimiometria

Os avanços tecnológicos e o desenvolvimento de técnicas analíticas inovadoras permitiram a aquisição de uma grande quantidade de dados em um período relativamente curto de tempo. O grande desafio que nasceu junto com esses avanços, foi transformar essa grande quantidade de informação em conhecimento (GASTEIGER; ENGEL, 2003). Neste contexto surgem as técnicas de aprendizado de máquinas (*Machine Learning Techniques*), ou seja, métodos e ferramentas matemáticas utilizadas para desenvolver algoritmos capazes de aprender e se aperfeiçoar através da experiência (MITCHELL, 1997).

Para aplicações na área de química, a análise de dados não se trata apenas da análise de informações primárias (dados espectrais, cromatográficos, etc), mas também da criação de informações secundárias, como modelos de calibração e classificação, de forma que seja possível prever o comportamento de outros dados no modelo criado e conhecer suas características (GASTEIGER; ENGEL, 2003). Em todo caso, independente da ferramenta utilizada, as técnicas quimiométricas, quando

aplicadas para dados espectrofotométricos, são baseadas na lei de Beer e partem do pressuposto que existe uma relação linear entre a absorção espectral de um analito e sua concentração.

Visto isso, é possível decompor qualquer conjunto de dados, **X**, adquirido a partir de medidas físicas (tais como espectros de infravermelho) de acordo com a Equação 1. E qualquer parâmetro de interesse, que se deseja estimar a partir das medidas adquiridas (como concentração, classe, qualquer propriedade de interesse, etc.) será representado pela matriz **Y** (ou vetor **y**).

$$\boldsymbol{X = CS^T + E} \qquad \textbf{Equação 1}$$

Em que **C** é a matriz de concentração do analito, $\mathbf{S}^T$ é a matriz das intensidades de sinal dos compostos puros e **E** é a matriz residual.

Na análise multivariada, as ferramentas de estudo de dados podem ser divididas em grupos: as técnicas de (i) análise exploratória, (ii) calibração e (iii) reconhecimento de padrões.

A técnica de análise exploratória mais utilizada é a Análise de Componentes Principais (PCA: *Principal Component Analysis*), uma técnica não supervisionada baseada na variância dos dados. As técnicas de calibração têm como objetivo prever um parâmetro (que pode ser uma concentração, atividade biológica, etc.) a partir de uma série de medidas (que podem ser espectros, por exemplo). As técnicas de reconhecimento de padrões podem ser subdivididas em supervisionadas e não supervisionadas. As técnicas não supervisionadas consistem basicamente na Análise de Agrupamentos Hierárquicos (HCA: *Hierarchical Cluster Analysis*). Essas ferramentas tentam representar semelhanças e diferenças entre as amostras sem conhecimento prévio das classes às quais pertencem. As técnicas não supervisionadas de reconhecimento de padrões diferem das técnicas de análise exploratória no sentido de que, as primeiras têm como objetivo realizar agrupamentos para observar semelhanças, já as ferramentas de análise exploratória não visam, necessariamente, a formação de agrupamentos, embora a PCA possa ser usada com esse objetivo (BRERETON, 2003).

As técnicas supervisionadas de reconhecimento de padrões compreendem, majoritariamente, técnicas classificatórias, em que um modelo é construído a partir de

um conjunto de amostras de classes conhecidas e, em seguida, uma amostra de classe desconhecida pode ser classificada de acordo com a semelhança/diferença entre ela e as amostras do conjunto de treinamento.

Antes de utilizar qualquer tipo de ferramenta de análise multivariada de dados, é fundamental estudar o conjunto de dados para remover informações que não estão associadas com o analito. Qualquer etapa prévia de tratamento matemático dos dados que preceda a etapa de análise multivariada é chamada de pré-processamento de dados e é de fundamental importância para eliminar ou atenuar essas flutuações do conjunto de dados (KOWALSKI; BEEBE, 1987).

1.3.3 Técnicas de Pré-processamento

A aquisição de dados espectrais não fornece apenas informações relevantes sobre a presença, ausência e concentração de compostos químicos. Dependendo da técnica de aquisição de espectros, equipamentos, condições experimentais, acessórios utilizados, etc., uma grande quantidade de informação relativa a fenômenos físicos, erros aleatórios e sistemáticos também estarão presentes no conjunto de dados adquirido. Para que essas informações não encubram a informação que está verdadeiramente relacionada com o analito ou com a propriedade que se deseja estudar, uma série de ferramentas matemáticas pode ser utilizada. Essas são as técnicas de pré-processamento que podem operar sobre as amostras ou sobre as variáveis (Figura 7).

Embora alguns trabalhos na literatura já propuseram ferramentas de otimização para pré-processamentos (DEVOS; DUPONCHEL, 2011; JARVIS; GOODACRE, 2005), é importante enfatizar que encontrar o pré-processamento mais adequado para um determinado conjunto de dados é um processo de tentativa e erro. O conhecimento da composição química dos dados e da forma de aquisição dos espectros é extremamente importante para encontrar a estratégia mais adequada para o pré-processamento dos dados (FERREIRA, 2015).

As técnicas de pré-processamento de objetos operam nas amostras, ao longo de todas as variáveis. Normalizações (como por exemplo, autoescalonamento, SNV, normalização vetorial), correção de linha de base, suavização, derivadas, padronização pelo desvio padrão, correção multiplicativa de sinal são exemplos de

pré-processamentos de amostras extremamente úteis. Já as técnicas de pré-processamento de variáveis são aplicadas às colunas da matriz de dados, para cada variável; são exemplos desse conjunto de técnicas o autoescalonamento (ou escalamento pela variância), a centragem na média, etc. (FERREIRA, 2015; KOWALSKI; BEEBE, 1987).

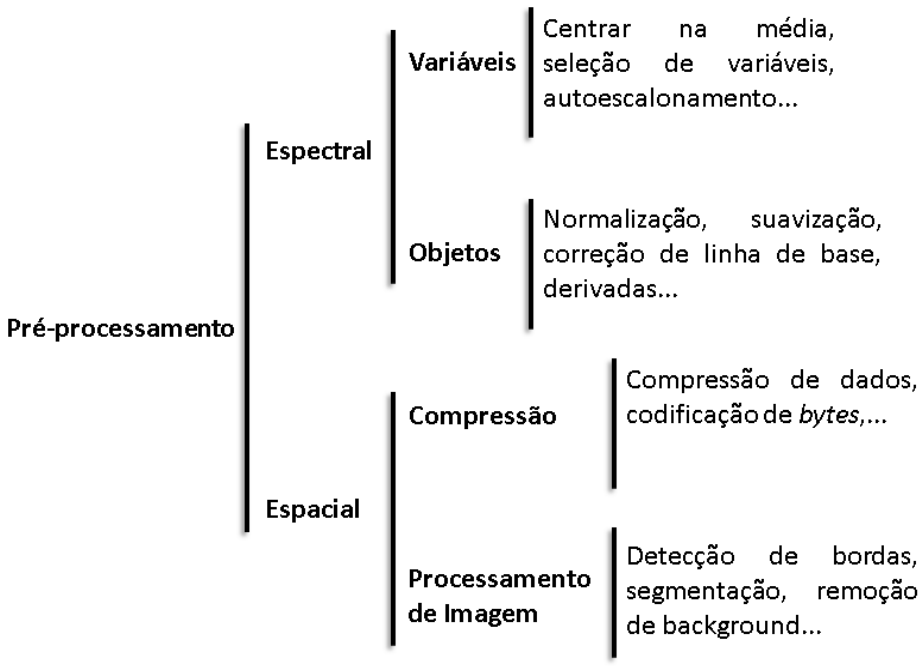

Figura 7 Esquema de técnicas de pré-processamento.

Ainda para o tratamento de Imagens Hiperespectrais, ferramentas de pré-processamento espacial também podem ser aplicadas a fim de corrigir defeitos nas imagens. Essencialmente, as correções espaciais das HSI podem ser caracterizadas como pré-processamento de amostras, visto que os pixels das imagens são considerados amostras independentes. Porém, como a natureza dessas ferramentas é bastante específicas para imagens, a Figura 7 mostra esse conjunto de técnicas separadamente. Aqui, serão apresentadas apenas as discussões sobre as técnicas utilizadas nos trabalhos desenvolvidos nesta tese.

Pré-processamento de Variáveis

1.3.3.1 Centragem na média

A Centragem na média (MC: *Mean Centering*) consiste na subtração dos valores individuais de uma coluna pela sua média, como pode ser visto na Equação 2.

$$\boldsymbol{x_{ijcorr}} = \boldsymbol{x_{ij}} - \bar{x}_j \qquad \textbf{Equação 2}$$

Em que x_{ijcorr} é o i-ésimo elemento da j-ésima coluna corrigido; x_{ij} é o i-ésimo elemento da j-ésima coluna inicial e $\bar{x}_j$ é a média dos elementos da j-ésima coluna. Geometricamente, a centralização na média é uma operação de translação em que o centro de gravidade dos objetos se torna a origem dos eixos. Essa operação faz com que os novos valores dos objetos sejam agora desvios em relação à média.

Pré-processamento de Amostras

1.3.3.2 Padronização Normal de Sinal (SNV):

A Padronização Normal de Sinal é uma ferramenta que tem como objetivo corrigir os efeitos aditivos e multiplicativos, geralmente causados por espalhamento de radiação. SNV realiza uma normalização pelo desvio-padrão dos valores de intensidade espectral de cada espectro precedido de uma centralização na média do próprio espectro (FEARN et al., 2009). Esta operação pode ser matematicamente representado pela Equação 3:

$$\boldsymbol{x_{corr}} = \frac{\boldsymbol{x_{org}} - a_0}{a_1} \qquad \textbf{Equação 3}$$

Em que a_0 e a_1 são, respectivamente, o valor médio e o desvio-padrão dos valores de intensidade do espectro que será corrigido (RINNAN; BERG; ENGELSEN, 2009). É possível perceber que, como o SNV não depende de um espectro de referência, o resultado da correção não irá depender do conjunto de dados.

1.3.3.3 Derivadas e suavização

Uma outra ferramenta de pré-processamento que opera nas amostras ao longo das variáveis, é a derivada. Essa técnica também é capaz de corrigir efeitos aditivos e multiplicativos dos espectros e pode ser usada para enfatizar picos ou bandas que não estão claros. Em contrapartida, a capacidade de enfatizar informação também pode acentuar ruídos e dificultar a análise dos dados, como pode ser observado na

Figura 8a. Por esse motivo, a derivada com filtro Savitzky-Golay ganhou popularidade, pois uma etapa de suavização precede a etapa de derivação (RINNAN; BERG; ENGELSEN, 2009; SAVITZKY; GOLAY, 1964) .

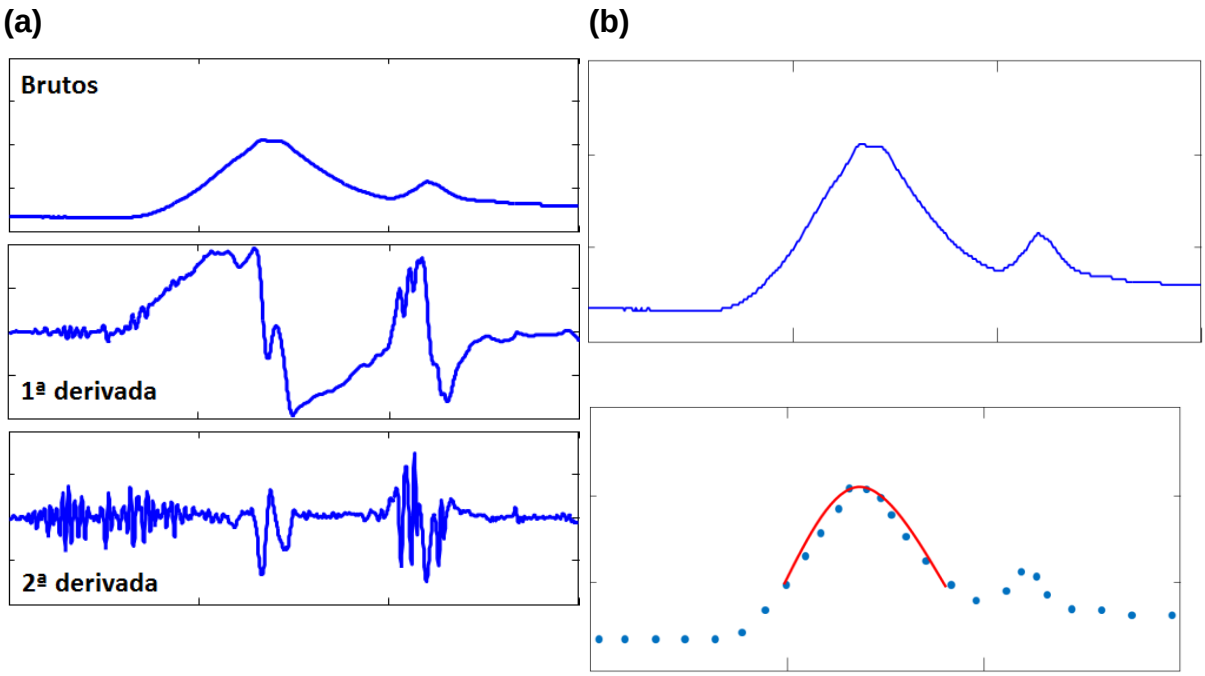

Figura 8 Efeito da derivada com filtro Savitzky-Golay sobre os espectros: (a) acentuação de ruído e (b) etapa de suavização. Adaptado de RINNAN et al, 2009.

Para obter a derivada no centro de um determinado ponto *i*, ajusta-se um polinômio numa região simétrica ao redor do i-ésimo ponto denominada de janela (Figura 8b). O tamanho da janela, escolhida pela quantidade de pontos ao redor de *i*, irá influenciar na adição ou atenuação de ruído presente, mas também poderá encobrir informações úteis. Dessa maneira, é preciso escolher com cuidado os parâmetros mais adequados para pré-processar um determinado conjunto de dados com derivada. Caso os dados apresentem apenas efeitos aditivos, a 1ª derivada é suficiente para corrigir esse problema, se efeitos multiplicativos são observados, a 2ª derivada pode ser útil para a correção dos dados. Além disso, a quantidade de ruído e a intensidade das bandas de interesse irão influenciar diretamente a escolha da janela do polinômio de suavização.

1.3.3.4 Correção Ortogonal de Sinais (OSC):

Algumas vezes, o conjunto de dados possui informações de interferentes que não podem ser eliminadas pelas técnicas usuais e, por essa razão, alguns filtros de pré-processamento mais avançados podem ser utilizados.

A Correção Ortogonal de Sinais (OSC: *Orthogonal Signal Correction*) foi desenvolvida por Wold (WOLD et al., 1998) e se baseia na ideia de que grande parte da variabilidade do conjunto de dados possui baixo valor preditivo, ou seja não está correlacionada com o parâmetro que se quer prever. Dessa forma é possível encontrar e remover a variabilidade do conjunto de dados que é ortogonal ao parâmetro de interesse.

Inicialmente, a ferramenta propunha uma correção em que as etapas de ortogonalização e regressão eram realizadas de forma iterativa (WOLD et al., 1998), porém posteriormente essa proposta foi modificada de forma que o número de componentes ortogonais ao parâmetro de interesse é definido antes da etapa de regressão (FEARN, 2000). Esse tipo de correção permite eliminar variações de interferentes entre amostras semelhantes e foi inicialmente aplicado em problemas de transferência de calibração (SJÖBLOM et al., 1998).

1.3.3.5 Mínimos Quadrados Generalizados Ponderados (GLSW):

A ferramenta de pré-processamento Mínimos Quadrados Generalizados Ponderados (GLSW: *Generalized Least Squares Weighting*) também consiste num filtro para atenuação de interferentes. O GLSW estima uma matriz de interferentes para os grupos de amostras que deveriam ser similares e utiliza essa informação para realizar uma ponderação, minimizando a atuação dos interferentes entre amostras similares como, por exemplo, espectros de uma mesma amostra que foram adquiridos em equipamentos diferentes.

Para aplicar GLSW para modelos de regressão, é necessário, primeiramente, calcular uma matriz diferença $\mathbf{X_d}$, que representa as diferenças entre as amostras similares. No caso de modelos de regressão, amostras que possuem valores de $\mathbf{y_i}$ semelhantes, devem apresentar perfis espectrais semelhantes. Portanto, é possível considerar o vetor **y** como um descritor de similaridade entre as amostras da matriz **X**, de forma que a informação dos interferentes ($\mathbf{X_d}$) seja, idealmente, ortogonal à **y**. Para isso, amostras similares são organizadas na matriz de forma que fiquem juntas. Assim, a diferença entre elas é calculada a partir de uma derivada de Savitzky-Golay nas colunas.

Essencialmente, as amostras da matriz corrigida resultante ($\mathbf{X_d}$) são uma medida das diferenças entre amostras que estão em suas proximidades, ou seja, amostras semelhantes. Da mesma maneira que na matriz **X**, uma derivada também é calculada para estimar as diferenças em **y**, fornecendo um vetor $\mathbf{y_d}$. Amostras significativamente diferentes contidas na matriz **X** podem perder parte das diferenças relevantes no processo de derivação, por isso, o vetor $\mathbf{y_d}$ será utilizado para ponderar essas diferenças, restituindo as informações significativas. Para isso, uma matriz de pesos **W** é construída utilizando o vetor $\mathbf{y_d}$ e os desvios-padrão de $\mathbf{y_d}$. A partir de então, a matriz de covariância é calculada a partir da Equação 4:

$$C = X_d^t W X_d \qquad \textbf{Equação 4}$$

Em seguida, essa matriz de covariância é submetida à uma decomposição por valores singulares de acordo com a Equação 5, os valores singulares (**S**) obtidos são então modificados a partir da Equação 6, em que $\mathbf{1_D}$ é uma matriz diagonal contendo apenas a unidade, e α é um parâmetro que mede o efeito do filtro. Assim, a matriz filtro **G** (calculada como mostrado na Equação 7) é utilizada para corrigir os dados iniciais.

$$\boldsymbol{C = VS^2V^t} \qquad \textbf{Equação 5}$$

$$\boldsymbol{D = \sqrt{\frac{S^2}{\alpha} + 1_D}} \qquad \textbf{Equação 6}$$

$$\boldsymbol{G = VD^2V^t} \qquad \textbf{Equação 7}$$

Em que **V** contém os autovetores obtidos na decomposição.

É importante mencionar que o parâmetro α é utilizado para aumentar ou diminuir o efeito do filtro. Quando o valor de α cresce, o efeito do filtro diminui e quando seu valor é pequeno, o efeito do filtro aumenta. A determinação do valor de α depende da ordem de grandeza das variáveis originais e deve ser escolhido com cautela, pois aumentando o efeito do filtro, é possível remover variabilidade informativa dos dados. Isso faz com que, caso a variabilidade do interferente tenha uma ordem de grandeza similar à variabilidade da medida analítica, o valor de α a ser escolhido deve ser mais alto do que o usual (geralmente, <1), para evitar perda de informação relevante (WISE et al., 2006).

1.3.4 Técnicas de Análise Exploratória

1.3.4.1 Análise de Componentes Principais

Uma matriz **X** contendo dados espectroscópicos de *n* amostras em *λ* variáveis diferentes pode conter uma grande quantidade de ruído e informações redundantes. Isso significa que a informação relevante do conjunto de dados que compõe **X** pode ser descrita em um espaço de dimensionalidade reduzida com *k* variáveis, em que $k \leq \lambda$. Matematicamente esse conceito, que é a base da Análise de Componentes Principais (PCA: *Principal Component Analysis*), pode ser expresso na forma de decomposição da matriz **X** como exemplificado na Equação 8 (SMILDE; BRO; GELADI, 2004).

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{\boldsymbol{t}} + \boldsymbol{E} \qquad \textbf{Equação 8}$$

Em que **T** é a matriz de escores e possui dimensão *n* x *k*, $\mathbf{P^t}$ é a matriz de pesos (ou *loadings*) com dimensões *k* x *λ*, e **E**, a matriz residual, apresenta as mesmas dimensões da matriz **X**, *n* x *λ*.

A PCA é, provavelmente, a ferramenta quimiométricas mais bem aceita e conhecida pela comunidade científica. É uma ferramenta de análise exploratória que pode ser utilizada como uma técnica de reconhecimento de padrões não supervisionada. Nesta análise, um novo espaço de variáveis ortogonais formadas a partir das variáveis originais é construído. As novas variáveis são obtidas no sentido de maximizar a variância dos dados e as amostras possuem novas coordenadas nesse novo espaço. A essas coordenadas dá-se o nome de escores. A esse par de escores e pesos é dado o nome de Componentes Principais (BRO; SMILDE, 2014).

A matriz dos escores da PCA está intimamente associada com as concentrações dos compostos das amostras, enquanto os pesos estão relacionados com as variáveis responsáveis pela maior variabilidade dos dados. O principal objetivo desta técnica é extrair informações sobre o conjunto de dados através de: (i) número de PCs independentes que melhor descrevem as informações no conjunto de dados, ou seja, encontrar as fontes de variação dos dados; (ii) o comportamento dos escores, ou seja, as similaridades e diferenças entre amostras; e (iii) o perfil dos pesos, ou seja, as variáveis originais que mais contribuem para a variabilidade e como estão relacionadas com as amostras (BRERETON, 2003; BRO; SMILDE, 2014).

1.3.4.2 Projection Pursuit

Devido à grande quantidade de informação que é possível adquirir por meio das atuais técnicas analíticas, ferramentas de redução de dimensionalidade, como PCA, se tornaram altamente atrativas para comprimir essas informações relevantes em um número reduzido de dimensões. Na química analítica, ferramentas não supervisionadas de análise multivariada como PCA (WOLD; ESBENSEN; GELADI, 1987) e HCA são amplamente utilizadas para acessar as informações mais relevantes em um conjunto de dados. A PCA e HCA descrevem o conjunto de dados de acordo com a variância e a distância entre as amostras, respectivamente, porém nem sempre essa é a melhor forma de acessar a informação em que se está interessado.

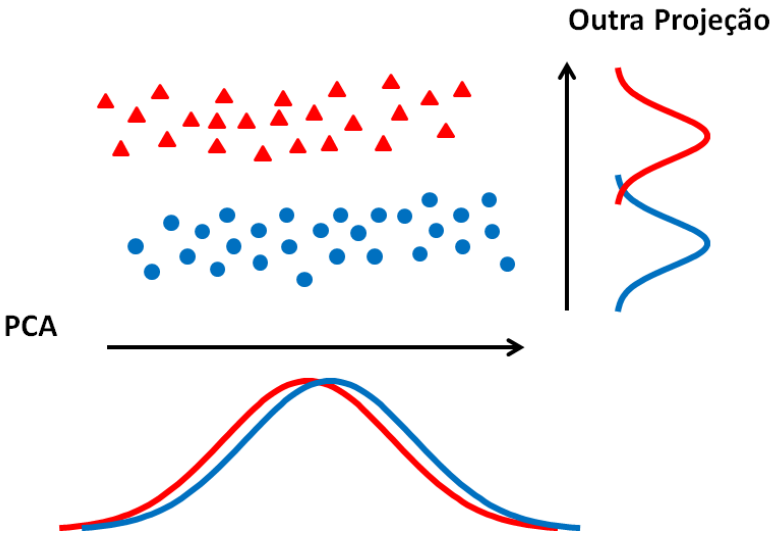

Figura 9 Diferença entre uma projeção no sentido da máxima variância (PCA) e uma projeção de interesse.

É possível notar que, no caso exemplificado na Figura 9, a projeção que representa maior variabilidade dos dados, PCA, não fornece informações significativas a respeito das duas diferentes classes. De fato, não é possível enxergar a separação dos dois grupos de amostras usando a projeção da PCA. Porém utilizando uma outra projeção, é possível encontrar estruturas interessantes e que sejam capazes de evidenciar a separação das amostras. Dessa forma, *Projection Pursuit* (PP) se apresenta como uma alternativa à PCA para encontrar a informação desejada. Friedman e Tuckey implementaram, pela primeira vez, PP como ferramenta de análise exploratória (FRIEDMAN; TUCKEY, 1974). Essa técnica tem como objetivo representar um conjunto de dados de alta dimensionalidade em um espaço de projeções com dimensões reduzidas capaz de revelar características de interesse.

Para isso, é necessário otimizar uma função que deve estar relacionada com a informação de interesse, a qual é chamada de índice de projeção.

Na literatura, é possível encontrar diferentes funções utilizadas como índices de projeção (FRIEDMAN; TUCKEY, 1974; HALL, 1989; HUBER, 1985; JONES; SIBSON, 1987; PEÑA; PRIETO, 2001b). De fato, qualquer função que esteja diretamente relacionada com a normalidade de uma dada distribuição pode ser utilizada como um índice de projeção, inclusive a variância, o que faz com que a PCA seja considerada como um caso particular de PP. Um dos índices de projeções que podem ser utilizados é a curtose.

Apesar de ser uma técnica bastante apropriada para análise exploratória dos dados, PP não é tão bem estabelecida como PCA. Isso ocorre porque os algoritmos disponíveis para a otimização da curtose são, em geral de difícil otimização. Recentemente, um novo algoritmo foi proposto por Hou e Wentzell (HOU; WENTZELL, 2011), que propõe uma forma mais rápida e eficiente de otimização.

Partindo do pressuposto que as amostras obedecem à distribuição normal, os valores devem estar dispostos em torno de uma média ($\bar{x}$), como mostrado na Figura 10a. A dispersão desses valores em torno dessa média pode ser descrita pela curtose, que é expressa matematicamente pela Equação 9:

$$K = \frac{1/n \sum_{i=1}^{n} (x_i - \bar{x})^4}{(1/n \sum_{i=1}^{n} (x_i - \bar{x})^2)^2} \qquad \textbf{Equação 9}$$

Em que *K* é a curtose, *n* é o número de amostras (ou objetos), x_i é o valor do parâmetro para uma amostra individual e $\bar{x}$ a média de todas as amostras (HOU; WENTZELL, 2011).

Peña e Pietro mostraram como a minimização e maximização da curtose pode ser capaz de revelar agrupamentos e amostras anômalas (*outliers*), respectivamente. Suponha então que se tem como objetivo visualizar diferentes agrupamentos em uma projeção unidimensional. Projetam-se então as amostras em uma direção e observa-se se é possível separar as amostras em dois grupos ao longo dessa projeção. Em seguida, procura-se outra direção de projeção, de forma que seja possível checar se, na matriz deflacionada, existem agrupamentos nesta segunda projeção, e assim sucessivamente (PEÑA; PRIETO, 2001a, 2001b).

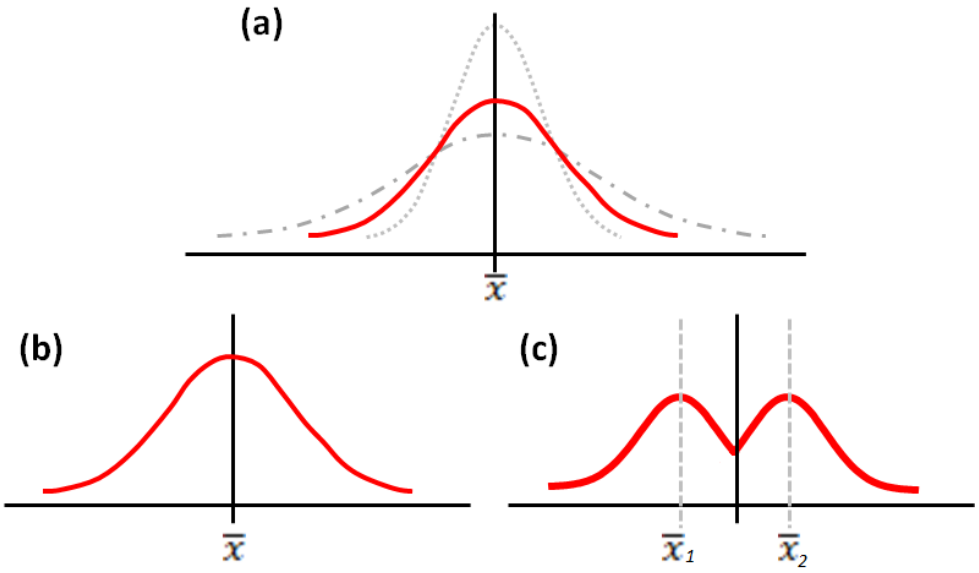

Figura 10 Figuras esquemáticas de (a) distribuições normais com diferentes curtoses; (b) distribuição normal unimodal; (c) distribuição normal bimodal.

Para encontrar essas projeções informativas, que revelam a formação de diferentes agrupamentos, é necessário que os objetos projetados numa dada direção estejam agrupados em torno de diferentes médias ($\bar{x}$) e que essas médias estejam significativamente separadas uma da outra, formando uma distribuição bimodal, como mostrado na Figura 10c. Distribuições bimodais tendem a apresentar valores pequenos de curtose, e portanto, para maximizar a formação de uma distribuição bimodal pode-se minimizar a curtose de uma dada distribuição (HOU; WENTZELL, 2011).

Na medida em que a curtose diminui, a distribuição se torna mais achatada, forçando os objetos a se separarem em dois agrupamentos distintos. Se a curtose é maximizada, a distribuição normal assume uma forma mais acentuada e revela objetos mais distintos da média (amostras anômalas ou *outliers*).

A extração dos vetores de uma análise PP pode ocorrer de duas formas, utilizando a curtose univariada ou multivariada. A extração dos vetores de projeção utilizando o algoritmo univariado ocorre por etapas, de forma que um vetor de projeção unidimensional na direção da curtose mínima é inicialmente obtido. Em seguida, as estruturas da projeção obtida são removidas da matriz original, gerando uma matriz deflacionada da qual um outro vetor unidimensional na direção da curtose mínima é

obtido, e assim sucessivamente até a obtenção do número desejado de vetores de projeção. O caso limite (sobreajuste) dessa distribuição ocorre com a separação em agrupamentos igualmente populosos nos vértices de um quadrado (projeção bidimensional) ou nas arestas de um paralelepípedo (projeção tridimensional). Já no caso do algoritmo multivariado, todos os vetores de projeção são extraídos simultaneamente e, em seu caso limite, as amostras tendem a ocupar o perímetro de uma circunferência de uma elipse (projeção bidimensional) ou na superfície de um elipsóide (projeção tridimensional) (HOU; WENTZELL, 2014).

As projeções das amostras pela minimização da curtose univariada utilizando um único vetor de projeção, forçam a separação das amostras em dois agrupamentos diferentes, como ilustrado na Figura 11a. No caso sobreajustado, utilizando dois vetores, é possível observar que as amostras são forçadas a ocupar os vértices de um quadrado, formando 4 agrupamentos distintos (Figura 11b). Para dados multivariados, a distribuição normal toma a forma indicada pela Figura 11c e a minimização da curtose, quando em seu caso limite, leva a um formato circular (Figura 11d). Ou seja, as amostras podem ser forçadas a ocupar o perímetro de uma circunferência, quando dois vetores de projeção são utilizados, e uma esfera, quando três vetores de projeção são utilizados.

Esses agrupamentos podem ser extremamente informativos e devem ser estudados com bastante cuidado. Os problemas de sobreajuste de uma análise PP ocorrem justamente quando o caso limite é atingido. Se muitas variáveis estiverem sendo utilizadas, o algoritmo pode encontrar correlações aleatórias entre as variáveis e fornecer projeções que não refletem nenhuma informação relevante. Assim, agrupamentos oportunísticos das amostras serão formados de forma que a curtose atinja seu valor mínimo (HOU; WENTZELL, 2014).

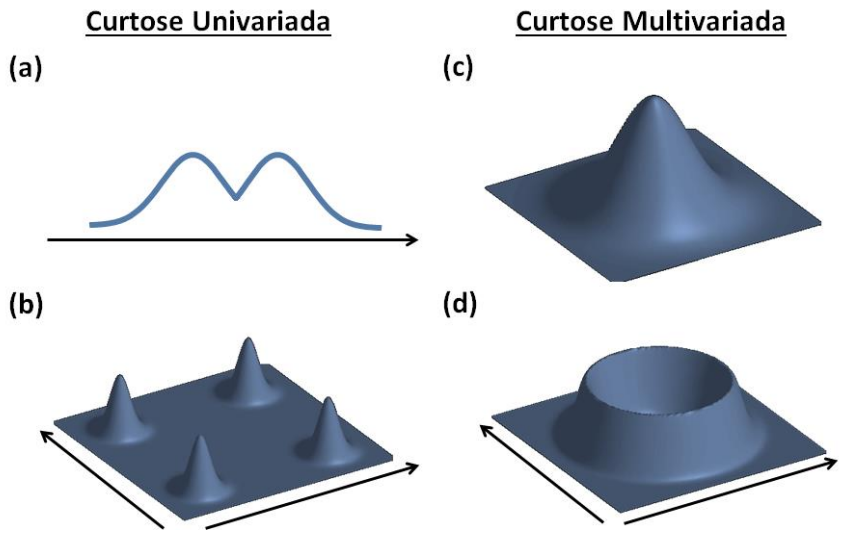

Figura 11 Distribuição normal utilizando: (a) curtose univariada mínima, com apenas 1 vetor de projeção; (b) curtose univariada mínima com 2 vetores de projeção (c) distribuição normal multivariada; (d) curtose mínima para uma distribuição normal multivariada.

Pode-se notar que, para tentar separar de 2 a 4 grupos diferentes de amostras, pode-se fazer uso de um algoritmo que otimize a curtose univariada. Porém, se o objetivo é identificar a separação de 5 ou mais classes, a utilização de um algoritmo utilizando a curtose multivariada é mais indicado.

Assim como PCA, PP também trabalha com projeção de amostras em um novo espaço e, portanto, também fornece matrizes de escores e pesos nessas projeções. Essas matrizes podem ser interpretadas exatamente como na PCA, com a exceção de que não teremos informação sobre a variância do conjunto de dados.

Foi observado que o desempenho da técnica PP é altamente eficiente quando o número de amostras é significativamente maior que o número de variáveis, geralmente numa razão em torno de 8:1, podendo variar com a natureza dos dados. Na medida em que o número de variáveis aumenta, a eficiência da análise em encontrar projeções realmente informativas diminui. Isso acontece porque com o aumento do número de variáveis, a probabilidade de o algoritmo encontrar correlações aleatórias entre variáveis pode aumentar, elevando também a possibilidade de encontrar agrupamentos de amostras que minimizem a curtose, porém não apresentem informações significativas, que é o caso de sobreajuste (HOU; WENTZELL, 2014).

Na química analítica, os conjuntos de dados utilizados apresentam, geralmente, uma quantidade de variáveis muito maior do que a quantidade de amostras. Por isso, para solucionar esse problema e ainda assim utilizar toda a informação relevante contida no conjunto de dados, uma PCA pode ser utilizada como um método de redução de dimensionalidade. Assim, é possível obter um conjunto de dados com uma razão de amostras/variáveis apropriada para a construção de um modelo PP eficiente. Uma vez que a razão amostras/variáveis adequada pode mudar com a natureza dos dados, é preciso encontrar uma forma de determinar melhor o número de variáveis latentes a ser usado na análise. Portanto, é possível realizar uma comparação entre espaços sucessivos obtidos a partir de modelos PP com diferentes níveis de compressão (diferentes números de PCs) para identificar quais os espaços que mais se assemelham. As projeções com n PCs encontradas a partir de agrupamentos oportunísticos não deverão se assemelhar a projeções formadas com n-1 e n+1 PCs e, portanto, deve-se encontrar uma ferramenta capaz de comparar o grau de similaridade entre esses espaços de projeções.

1.3.4.2.1 Análise de Procusto

Na mitologia grega, Procusto era um malfeitor que foi derrotado por Teseu em uma de suas aventuras. Procusto era o dono de uma estalagem que possuía leitos de ferro. Ele costumava amarrar os viajantes nos leitos e fazê-los caber exatamente no espaço disponível. Se os viajantes eram menores que os leitos, ele os esticava e, quando maiores, ele cortava os membros para ajustá-los exatamente aos comprimentos dos leitos (BULFINCH, 2002).

A mitologia grega emprestou o nome de Procrusto para a matemática, cujo termo “Análise de Procusto” foi primeiramente empregado por Hurley e Cattell em 1962 (HURLEY; CATTELL, 1962). Essa ferramenta pode ser utilizada para comparar dois conjuntos de dados realizando uma série de operações matemáticas, como rotações, translações, etc., e, em seguida, calculando um índice de dissimilaridade entre esses dois conjuntos.

Suponha que duas matrizes de escores **T** e **Z** são obtidas de dois conjuntos de dados originais contendo *n* amostras e *m* variáveis, **X** e **Y**, a partir de alguma ferramenta de análise multivariada de dados. A análise de Procusto irá realizar uma

série de transformações (Figura 12) na matriz **Z** para ajustá-la à matriz **T** (ANDRADE et al., 2004).

Essa série de operações matemáticas tem como objetivo minimizar a soma dos quadrados das distâncias entre os elementos correspondentes de **Z** e **T**. Quanto menor a soma dos quadrados, mais semelhantes os conjuntos podem ser considerados e o índice de dissimilaridade entre os dois é menor (ANDRADE et al., 2004; KRZANOWSKI, 2000).

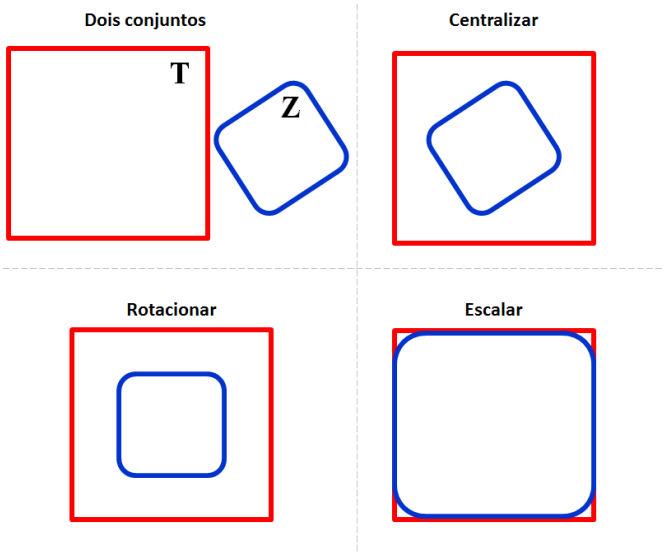

Figura 12 Sequência de transformações matemáticas que fazem parte da Análise de Procusto.

Matematicamente, a análise de Procusto tenta representar a matriz **T** a partir de uma transformação linear realizada em **Z**, de acordo com a Equação 10, em que *a* é um escalar, **R** é a matriz de rotação/reflexão, **B** uma matriz *n*×*m* de translação, $\widehat{\boldsymbol{T}}$ é a matriz resultante da transformação linear (ou **Z** ajustada) e **E** é a matriz de resíduos. O índice de dissimilaridade *d* entre **T** e **Z** pode ser calculado a partir da Equação 11.

$$\boldsymbol{T} = a\boldsymbol{Z}\boldsymbol{R} + \boldsymbol{B} + \boldsymbol{E} = \widehat{\boldsymbol{T}} + \boldsymbol{E} \qquad \textbf{Equação 10}$$

$$d = \frac{\sum_i \sum_j \left(t_{ij} - \hat{t}_{ij}\right)^2}{\sum_i \sum_j \left(t_{ij} - \bar{t}_j\right)^2} \qquad \textbf{Equação 11}$$

Em que t_{ij} e $\hat{t}_{ij}$ são elementos da matriz **T** e $\widehat{\mathbf{T}}$, respectivamente e $\bar{t}_j$ é a média dos valores da variável j em **T**.

1.3.5 Técnicas de Calibração e Análise Quantitativa

Os métodos de calibração multivariada consistem em um conjunto de técnicas cujo objetivo principal é quantificar determinados parâmetros a partir de um modelo que relacione dois conjuntos de variáveis. Os equipamentos espectrofotométricos fornecem informações a respeito dos analitos em termos de magnitude de sinal e não em concentrações. Dessa forma, o objetivo principal da calibração é encontrar um modelo matemático capaz de prever a concentração de um determinado analito em uma dada amostra a partir das variáveis espectrais (BRERETON, 2003; SANCHEZ; KOWALSKI, 1988).

1.3.5.1 Regressão por Mínimos Quadrados Parciais

Na análise por componentes principais, um modelo é criado a partir da variabilidade da matriz **X**, contendo espectros adquiridos para as amostras em estudo. Em geral, é comum a necessidade de avaliar o comportamento dos espectros de amostras de acordo com uma ou mais propriedades **Y** aparentemente independentes. Assim, a técnica dos Mínimos Quadrados Parciais tem como objetivo construir um modelo capaz de considerar as variações de cada matriz de dados (**X** e **Y** individualmente) e considerar variações conjuntas, relacionando **X** com **Y** (WOLD; SJÖSTRÖM; ERIKSSON, 2001). Dessa forma, o modelo é construído a partir da correlação entre uma matriz **X** contendo um conjunto de variáveis de previsão e a matriz **Y**, que contém variáveis dependentes que são parâmetros de interesse. Essa matriz **Y** pode assumir a forma de um vetor **y** (*n*x1), quando apenas um parâmetro de interesse está sendo avaliado (PLS-1) e pode assumir a forma de uma matriz **Y**(*n*x*m*), quando *m* parâmetros estão sendo avaliados (PLS-2) (BRERETON, 2003).

Matematicamente, as matrizes são decompostas em matrizes de escores e pesos, como exemplificado na Equação 12 e na Equação 13, obedecendo a relação estabelecida pela Equação 14. Essa relação entre os as matrizes dos escores **T** e **U** estabelece que as novas variáveis do modelo devem ser obtidas de forma a maximizar a correlação entre as matrizes **X** e **Y**.

$$\boldsymbol{X = TP^t + E} \qquad \textbf{Equação 12}$$

$$\boldsymbol{Y = UL^t + F} \qquad \textbf{Equação 13}$$

$$\boldsymbol{U = TW} \qquad \textbf{Equação 14}$$

Em que $\mathbf{P^t}$ e $\mathbf{L^t}$ são as matrizes dos pesos (ou *loadings)* de **X** e **Y**, respectivamente, **W** é a matriz de pesos e **E** e **F** as matrizes residuais de **X** e **Y**, respectivamente.

As equações acima podem ser combinadas para fornecer a matriz dos coeficientes de regressão do modelo $\hat{B}$, que será usada para prever uma propriedade $\hat{\boldsymbol{y}}$ de uma amostra quando seu espectro $\boldsymbol{x_i}$ ($\boldsymbol{x_i} \in \boldsymbol{X}$) for conhecido (GEMPERLINE; KALIVAS, 2006).

$$\widehat{\boldsymbol{B}} = \boldsymbol{P}(\boldsymbol{P^t P})^{-1}\boldsymbol{WL} \qquad \textbf{Equação 15}$$

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{B}} \qquad \textbf{Equação 16}$$

É possível perceber que as novas variáveis dos modelos PLS, chamadas de Variáveis Latentes (LV: *Latent Variables*), são definidas na direção da maior correlação entre **X** e **Y** e não são, necessariamente, ortogonais entre si, como nos modelos PCA.

Mínimos Quadrados Parciais Esparsos – sPLS

Extensões dos modelos PLS e PLS-DA (seção 1.3.6.1 deste capítulo) também estão disponíveis na literatura. Dentre eles, o modelo dos Mínimos Quadrados Parciais Esparsos (sPLS: *Sparse Partial Least Squares*) propõe uma metodologia em que uma quantidade otimizada de variáveis originais é forçada a zero por um termo de penalidade (CAO et al., 2008; CAO; BOITARD; BESSE, 2011). O principal objetivo dos métodos esparsos é reduzir o ruído gerado por variáveis que não são informativas, forçando-as a assumir o valor de zero. Para isso, o operador Lasso (LASSO: *Least Absolute Shrinkage and Selection Operator*) pode ser utilizado para introduzir o termo de penalidade, reduzindo e selecionando o grau de esparsidade do modelo final (RASMUSSEN; BRO, 2012).

Para a construção desses modelos é necessário otimizar o número de variáveis latentes (sLV: *sparse Latent Variables*) e o número de variáveis originais que serão incluídas na construção de cada sLV (CALVINI; ULRICI; AMIGO, 2015; FILZMOSER; GSCHWANDTNER; TODOROV, 2012). A otimização desses parâmetros para modelos sPLS pode ser efetuada pela análise das Figuras de mérito RMSEP e R^2, enquanto para modelos sPLS-DA, é possível avaliar a eficiência dos modelos (ver seções 1.3.8.1 e 1.3.8.2).

1.3.6 Técnicas de Classificação

1.3.6.1 Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA)

A técnica de Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA: *Partial Least Squares – Discriminant Analysis*) consiste essencialmente na aplicação da técnica PLS para fins de discriminação (BARKER; RAYENS, 2003). E assim, da mesma forma que descrita na seção 1.3.5.1, o objetivo desta ferramenta é prever a classe de uma determinada amostra utilizando os coeficientes de regressão definidos na Equação 15. Nesse caso, a matriz **Y** (ou vetor **y**) não mais contém os parâmetros para previsão, mas as classes a que pertencem cada amostra do conjunto de Treinamento. Quando apenas duas classes são estudadas, **y** é um vetor e a abordagem PLS1 é utilizada; entretanto, se 3 ou mais classes são estudadas em um mesmo modelo, **Y** é uma matriz e a abordagem PLS2 pode ser utilizada (BRERETON; LLOYD, 2014). Neste caso, a matriz **Y** (ou vetor **y**) deve ser construída em forma de uma matriz binária capaz de indicar as amostras que pertencem a cada uma das classes. Após a etapa de treinamento (análoga à calibração), o valor de $\hat{y}$ é previsto como descrito na Equação 16. Entretanto, esse valor previsto não assumirá a forma de números inteiros, mas valores reais que estarão próximos de 1, quando pertencer à uma determinada classe, e de 0 quando não pertencer (Figura 13). Desta forma, um limiar deve ser estabelecido de forma a realizar atribuição das classes para as amostras desconhecidas (FERREIRA, 2015).

A escolha desse limiar vai depender das densidades de probabilidade das classes e, obedecendo à teoria Bayesiana, assume que as funções de densidade de cada classe obedecem à distribuição normal e o limiar de decisão será definido de forma a minimizar o risco de classificar uma amostra fora de sua classe (Figura 13).

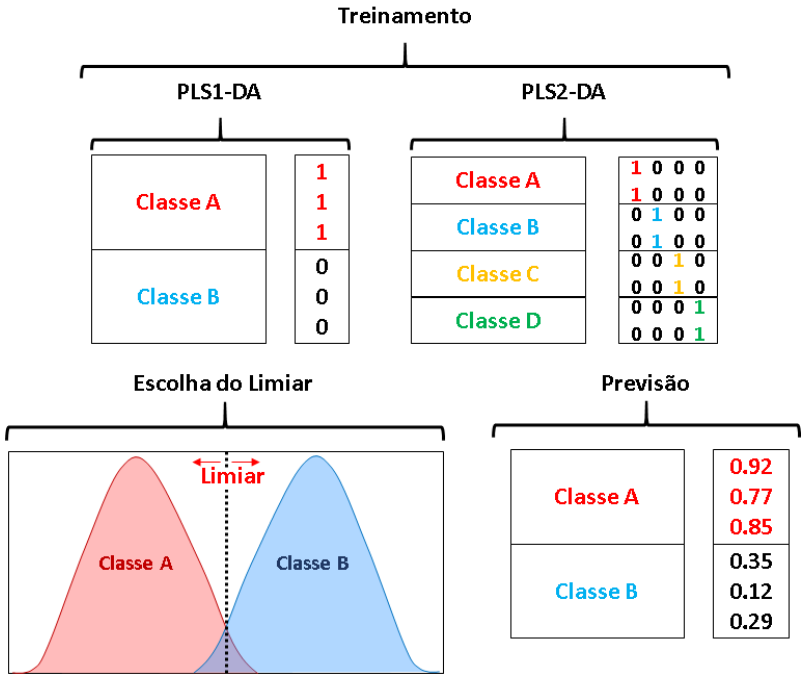

Figura 13 Esquema de modelos PLS-DA: etapa de treinamento (acima), escolha do limiar (abaixo à esquerda) e conjunto de previsão (abaixo à direita).

1.3.6.2 Máquinas de Vetores de Suporte (SVM-DA)

As Máquinas de Vetores de Suporte (SVM: *Support Vector Machines*) são, por definição, uma técnica linear supervisionada que pode ser utilizada para construir modelos de classificação ou de regressão e é capaz de lidar com dados não-lineares. No contexto linear classificatório, a ideia principal dessa ferramenta é encontrar um plano (hiperplano ou reta) que maximize a distância entre as amostras (x_i) mais semelhantes de duas classes (assumindo que o problema envolva apenas duas classes, A definida como +1 e B definida como -1). A fronteira entre as classes é otimizada e definida como equidistante entre as amostras mais extremas de cada classe. Essas amostras extremas são utilizadas para traçar a fronteira e são chamadas de Vetores de Suporte (BRERETON, 2009)

A função que rege o critério de classificação é mostrada na Equação 17, em que α é o multiplicador de Lagrange, um parâmetro que pode ser utilizado para otimização de funções que contenham restrições, *c* é a classe atribuída, *s* os vetores de suporte, $\boldsymbol{x_I}$ a amostra a ser classificada e o *bias* (ou viés) uma medida do erro sistemático do modelo.

$$g(x_l) = sgn(\sum_{i \in SV} \alpha_i c_i s_i x_i^t + bias) \qquad \textbf{Equação 17}$$

Para casos não lineares, o hiperplano não será construído no espaço definido pelas variáveis originais, mas em um novo espaço de maior dimensionalidade que será gerado a partir de uma função característica φ(x). Nesse novo espaço, define-se uma fronteira linear que, quando projetada no espaço de variáveis original, assume uma forma não linear, como sugerido pela Figura 14.

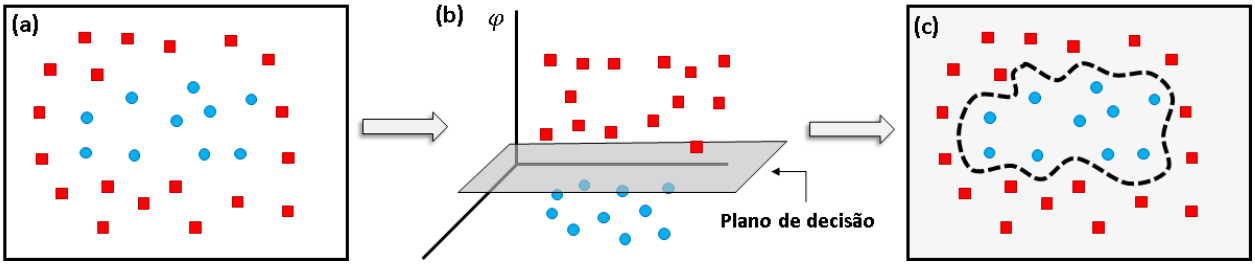

Figura 14 Esquema da construção das fronteiras para um modelo SVM para classificação; (a) espaço com classes com limites não lineares; (b) expansão para um espaço de maior dimensionalidade a partir de φ; (c) projeção das fronteiras não-lineares. (BRERETON, 2009)

Neste caso, a Equação 17 sofre uma mudança em que um produto interno *kernel* irá substituir o produto interno $s_i x_i^t$ (Equação 18). As funções do tipo *kernel* são comumente utilizadas para realizar um mapeamento sobre um espaço de maior dimensionalidade de forma implícita, ou seja, sem a necessidade de calcular as coordenadas desse espaço (SEMOLINI, 2002).

$$g(x_l) = sgn(\sum_{i \in SV} \alpha_i c_i K(s_i, x) + bias) \qquad \textbf{Equação 18}$$

Em que $K(s_i,x)$ é o produto interno Kernel, sendo os mais comuns: (i) Função de Base Radial (RBF: *Radia Function Basis*); (ii) Função polinomial; e (iii) Sigmoidal (SMOLA; SCHÖLKOPF, 2004).

1.3.7 Técnicas de Resolução de Curvas

1.3.7.1 Resolução Multivariada de Curvas – Mínimos Quadrados Alternados (MCR-ALS):

A técnica de Resolução Multivariada de Curvas por Mínimos Quadrados Alternados (MCR-ALS: *Multivariate Curve Resolution – Alternating Least Squares*) consiste numa ferramenta que tem como objetivo decompor a matriz de dados em

perfis químicos dos componentes puros e suas respectivas contribuições de uma forma iterativa. A Equação 1 é a principal equação que governa os modelos MCR.

$$\boldsymbol{X = CS^T + E} \qquad \textbf{Equação 19}$$

Para buscar as soluções para esse problema, estimativas iniciais para os perfis espectrais (**S**$^{\mathbf{T}}$) ou para as concentrações (**C**) são necessárias. Essas estimativas iniciais podem ser obtidas por PCA, técnicas como SIMPLISMA (WINDIG; STEPHENSON, 1992) ou até mesmo fornecidas quando disponíveis. Uma vez obtidos, por exemplo, os perfis espectrais (**S**$^{\mathbf{T}}$), as estimativas iniciais são utilizadas para calcular as concentrações **C*** a partir da Equação 20, assumindo que o resíduo do modelo é igual a zero. Com as concentrações estimadas **C***, é possível calcular novos perfis espectrais **S**$^{\mathbf{T*}}$ (Equação 21) e assim sucessivamente, utilizando o algoritmo ALS para otimizar os valores de **C*** e **S**$^{\mathbf{T*}}$ a se ajustarem a **X** (DE JUAN; JAUMOT; TAULER, 2014).

$$\boldsymbol{C^* = XS^{t+}} \qquad \textbf{Equação 20}$$

$$\boldsymbol{S^{t*} = C^+X} \qquad \textbf{Equação 21}$$

Em que **C**$^{+}$ e **(S**$^{\mathbf{T}}$**)**$^{+}$ são as pseudoinversas das matrizes **C** e **S**$^{\mathbf{T}}$, respectivamente.

Matematicamente, o MCR possui um problema de ambiguidade rotacional, ou seja, é possível encontrar diversas soluções para esse problema e, por isso, diversas restrições matemáticas podem ser utilizadas para melhorar o modelo e fornecer soluções quimicamente viáveis. Restrições de não negatividade, unimodalidade, fechamento, seletividade entre outras são exemplos que estão descritos na literatura (JAUMOT; DE JUAN; TAULER, 2015; TAULER; SMILDE; KOWALSKI, 1995).

Uma outra vantagem da técnica de MCR-ALS é a possibilidade de lidar com um conjunto de dados aumentados, ou seja, é possível concatenar diferentes matrizes de dados para aumentar a capacidade do modelo de resolver os componentes (DE JUAN; TAULER, 2003; TAULER, 1995).

Outra vantagem da técnica de MCR-ALS que é especialmente útil para aplicações forenses é o fato de que na maioria dos problemas, o espectro do composto de interesse é conhecido. Desta forma, essa informação pode ser utilizada

como referência e restrições de igualdade podem ser impostas na resolução da mistura.

1.3.8 Validação e Figuras de Mérito

Para avaliar a qualidade e confiabilidade dos modelos construídos, é possível avaliar uma série de parâmetros estatísticos, chamados de Figuras de Mérito. Esses parâmetros dependem não só da abordagem utilizada (exploratória, classificatória ou calibração), mas também do modelo utilizado em cada abordagem. Neste tópico, serão discutidos os parâmetros mais comuns encontrados na literatura obtidos na etapa de validação.

1.3.8.1 Figuras de Mérito para Calibração

Os parâmetros a serem avaliados na calibração são indicativos do quanto da variância total dos dados é explicada pelo modelo. O coeficiente de determinação (R^2), por exemplo, mostrado na Equação 22 é um parâmetro que mostra o quão bem o modelo se ajusta aos dados, pois quando $R^2 \approx 1$ significa que o resíduo é pequeno e que a soma quadrática da regressão se assemelha à soma quadrática total.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} \qquad \textbf{Equação 22}$$

Em que $\hat{y}_i$ é o valor da variável dependente estimado pelo modelo, $\bar{y}$ é a média dos valores da variável dependente, e y_i é o valor medido da i-ésima variável dependente.

Outro parâmetro importante para a escolha do melhor modelo é a Raiz do Erro Quadrático Médio (RMSE: *Root Mean Square Error*), que pode ser calculado para a calibração (RMSEC), validação cruzada (RMSECV) e previsão (RMSEP). Em geral, é importante considerar sempre o RMSEP quando um conjunto de previsão estiver disponível.

O RMSEP é definido de acordo com a Equação 23 e é uma medida do erro do modelo e deve ser minimizado na etapa de validação.

$$RMSEP = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{N}} \qquad \textbf{Equação 23}$$

Em que N é o número de amostras de previsão (para o cálculo do RMSECV, N será o número de amostras do conjunto de calibração) (FERREIRA, 2015).

Quando o modelo é construído, também é possível identificar a presença de erros sistemáticos, indicada pelo parâmetro *bias* (ou viés). E

$$bias = \frac{\sum(y_i - \hat{y}_i)}{N} \qquad \textbf{Equação 24}$$

Em que *N* é o número de amostras de calibração ou de previsão, dependendo da etapa de construção do modelo.

1.3.8.2 Figuras de Mérito para Classificação

No caso dos métodos de classificação, como PLS-DA, o valor de RMSECV não tem utilidade para validar o modelo. Esse valor representa apenas os desvios de $\hat{Y}$ (ou $\hat{\boldsymbol{y}}$) para a matriz binária; quanto maiores os desvios da matriz binária, maior será a contribuição para o valor de RMSECV. É importante notar que esse desvio não fornece nenhuma informação a respeito das fronteiras de cada classe e, portanto, outras figuras de mérito são necessárias para avaliar os modelos de classificação (KJELDAHL; BRO, 2010).

Uma vez que o limiar para uma determinada classe é escolhido, uma série de parâmetros úteis para os modelos de classificação podem ser avaliados. A Figura 15 mostra a chamada Matriz de Confusão, que é gerada para os modelos classificatórios. Nela, o número de amostras de Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN) para a classe A é mostrada como exemplo.

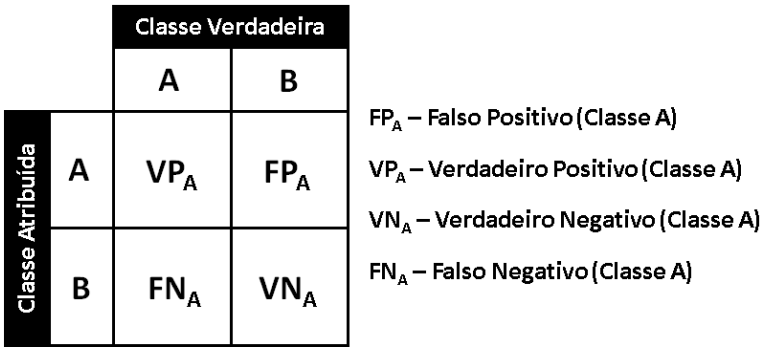

Figura 15 Matriz de confusão para um modelo com duas classes.

A partir desses valores, alguns parâmetros importantes podem ser definidos. A sensibilidade (Sn: *Sensitivity*) de um modelo fornece informações a respeito da capacidade de um determinado modelo de evitar falsos negativos, ou seja, reflete a taxa de verdadeiros positivos e pode ser definido pela Equação 25. Já a chamada especificidade, ou seletividade, (Sp: *Specificity*) pode ser interpretada como a taxa de verdadeiros negativos, ou seja, reflete a capacidade do modelo de evitar falsos positivos e está expressa na Equação 26. E, finalmente, a taxa de erros do modelo (ER: *Error Rate*), definida na Equação 27.

$$Sn = \frac{VP}{VP + FN} \qquad \textbf{Equação 25}$$

$$Sp = \frac{VN}{VN + FP} \qquad \textbf{Equação 26}$$

$$ER = 1 - \frac{Sn \times Sp}{2} \qquad \textbf{Equação 27}$$

É a partir desses parâmetros que é possível definir as chamadas Curvas ROC (ROC: *Receiver Operating Characteristics*). Essas curvas representam gráficos que mostram a variação de Sn e Sp na medida em que o limiar de classificação varia e são úteis para avaliar o desempenho de um modelo binário (Figura 16).

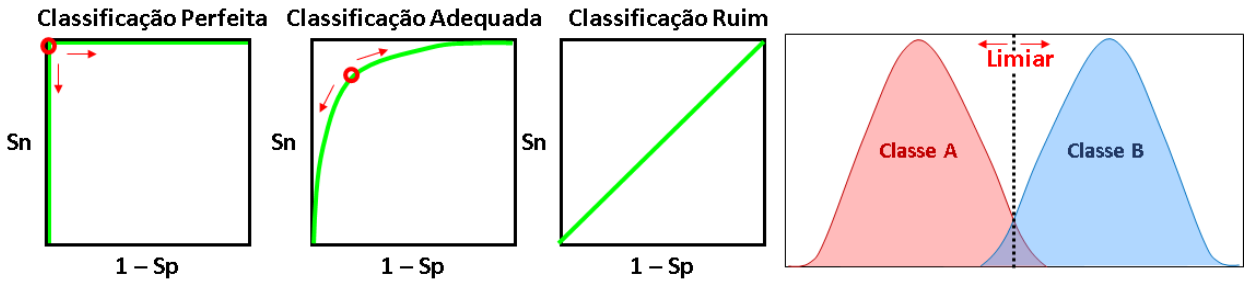

Figura 16 Esquema de construção das Curvas ROC.

Outro parâmetro utilizado para avaliar a qualidade dos modelos de classificação sPLS-DA é a Eficiência e está definida pela Equação 28, (CALVINI; ULRICI; AMIGO, 2015).

$$EFF = \sqrt{Sn \times Sp} \qquad \textbf{Equação 28}$$

2 PROJECTION PURSUIT E ANÁLISE DE PROCUSTO PARA DISCRIMINAÇÃO DE TINTAS DE CANETAS

2.1 INTRODUÇÃO

Um dos grandes problemas que surgem nos departamentos de polícia científica é a identificação de fraudes em documentos. Diversos casos que envolvem adulterações de cheques, carteiras de trabalho, atestados médicos, testamentos, passaporte, entre outros documentos, podem ser solucionados a partir da análise das tintas e do papel utilizados para produzir o documento em questão. A área da ciência forense que trata dos estudos de manipulações de documentos é conhecida como documentoscopia (BRUNELLE; CRAWFORD, 2003). O estudo de tintas de canetas e de outros instrumentos gráficos é importante para identificar adulterações em documentos (EZCURRA et al., 2010) e, muitas vezes, resolver casos litigiosos. Em diversas situações a comparação entre tintas de canetas presentes em um mesmo documento pode ser suficiente para identificar uma possível falsificação. Complementarmente, uma possível correlação de um instrumento escrevente com a tinta de manuscritos, como por exemplo, anotações de pagamentos de propinas, é capaz de servir de prova material segura para ligar o possuidor do instrumento com uma determinada ação delituosa. Por isso, diversos trabalhos podem ser encontrados na literatura abordando o problema de diferenciação de tintas de canetas.

As técnicas de espectroscopia Raman, IR e fluorescência de Raios-X (XRF) foram empregadas na tentativa de discriminar canetas pretas e azuis de diferentes tipos (ZIEBA-PALUS et al., 2016). Sessenta e nove canetas pretas e azuis dos tipos gel e esferográfica foram adquiridas e usadas para escrever pequenos textos em papéis brancos. Espectros Raman foram adquiridos diretamente das tintas depositadas sobre o papel e os espectros de infravermelho e fluorescência de Raios-X foram adquiridos das tintas secas depositadas sobre uma pequena placa de vidro. Os autores relataram uma grande dificuldade de obter espectros bem resolvidos no Raman, especialmente para tintas azuis do tipo gel, devido à sua alta fluorescência. A partir dos espectros de infravermelho das tintas, os autores foram capazes de identificar uma série de pigmentos orgânicos característicos. Os compostos inorgânicos presentes também foram caracterizados. Os espectros foram comparados com o objetivo de diferenciar as tintas estudadas. Para as canetas esferográficas

azuis, 95% das canetas puderam ser discriminadas de acordo com os espectros Raman e IR. Esse percentual de discriminação aumentou para, aproximadamente, 99% quando os dados de XFR foram adicionados na análise. Para as canetas pretas, 98% das canetas puderam ser discriminadas por Raman e infravermelho, porém a caracterização por XFR não mostrou melhoras nos resultados. Para as canetas gel (pretas e azuis), 97% das tintas foram discriminadas por meio dos métodos utilizados.

Nos trabalhos citados, a diferenciação de canetas ocorre por meio da análise dos perfis químicos de cada tinta de caneta. Apesar de fornecer resultados confiáveis, são análises trabalhosas, que dificultam a comparação quando muitas amostras devem ser analisadas e, além disso, dependem da experiência do analista e do seu domínio da técnica aplicada. Por esse motivo, as ferramentas de análise multivariada começaram a ser aplicadas com o objetivo de criar metodologias cada vez menos subjetivas.

Dentre essas ferramentas, destacam-se as técnicas de classificação, como LDA, PLS-DA, SIMCA, etc. Kher e colaboradores utilizaram cromatografia líquida de alta eficiência (HPLC) e IR para diferenciar canetas esferográficas azuis de diferentes marcas. Oito marcas de canetas foram selecionadas e seis canetas de cada marca adquiridas. Cada caneta foi usada para registrar uma linha ou um pequeno texto em papel branco, cuja tinta foi removida com uma solução de acetonitrila. A avaliação dos dados cromatográficos associada à PCA e LDA foi capaz de discriminar 96,4% e 97,9% das canetas, respectivamente (KHER et al., 2006). De acordo com os autores, a PCA dos dados obtidos por infravermelho não apresentou um bom desempenho e o modelo LDA apresentou um percentual de classificação correta de 62,5%.

Silva e colaboradores propuseram uma metodologia não destrutiva utilizando espectroscopia na região do visível e análise multivariada para discriminar diferentes marcas e tipos de canetas pretas (DA SILVA et al., 2014). Os autores adquiriram vinte e cinco canetas pretas em estabelecimentos comerciais que foram usadas para registrar um traço em papel branco. Sessenta espectros usando a faixa espectral de 400 a 1000 nm foram adquiridos de cada traço, fornecendo um conjunto de dados de 1500 espectros. Estudos de variabilidade entre diferentes lotes de canetas e diferentes papéis também foram realizados. Um modelo PLS-DA foi construído para cada tipo e cada marca de caneta, utilizando 40 espectros de cada caneta para

compor o conjunto de treinamento e 20 para o de previsão. Todas as amostras foram corretamente classificadas, inclusive as amostras produzidas em diferentes papéis e usando canetas de diferentes lotes.

Borba e colaboradores exploraram o potencial da espectroscopia Raman para diferenciação de tintas de canetas azuis por marcas e modelos. Três canetas de cada modelo foram adquiridas e usadas para registrar sete linhas em papel branco, de onde foi coletado um espectro cada, fornecendo, assim, 249 espectros Raman para análise. Os modelos PCA e Análise de Agrupamento Hierárquico (HCA: *Hierarchical Cluster Analysis*) foram construídos para avaliar o comportamento dos dados, porém não foi possível distinguir todas as tintas estudadas. Dessa forma, um modelo PLS-DA foi construídos para cada modelo de caneta, fornecendo percentuais de classificação correta maiores que 97% (BORBA; HONORATO; DE JUAN, 2015).

Nosso grupo de pesquisa estudou o potencial da espectroscopia no infravermelho médio associada à PCA e LDA para classificar tintas de canetas azuis (SILVA et al., 2012). Diferentes ferramentas de seleção de variáveis foram utilizadas em associação com a técnica de LDA para classificar canetas azuis por tipo e por marca em diferentes papéis. Foram adquiridas 5 marcas de canetas esferográficas, 2 marcas do tipo *rollerball* e 3 marcas do tipo gel. Para cada marca foram adquiridas 10 canetas diferentes, sendo 5 provenientes do mesmo lote e 5 de lotes diferentes. Cada caneta foi usada para registrar círculos em papéis de diferentes tipos (duas marcas de papel branco e uma marca de papel reciclado), de forma que os espectros foram adquiridos diretamente de cada círculo, sem necessidade de preparo de amostras. A PCA dos dados mostrou uma tendência de separação de algumas canetas, porém apenas os modelos LDA associados à seleção de variáveis foram capazes de distinguir as canetas, apresentando um percentual de classificação correta de 100% para todos os casos (classificação por tipo e marca nos três tipos de papel), com exceção da classificação por marcas no papel reciclado que apresentou um percentual de classificação correta de 91,3%. O uso das técnicas de seleção de variáveis neste contexto se mostrou extremamente relevante, pois houve uma grande influência do papel nos espectros de Infravermelho.

Calcerrada e colaboradores revisaram as recentes metodologias propostas na análise de documentos, incluindo não só a discriminação de tintas, análise de papéis,

cruzamento de traços, etc (CALCERRADA; GARCÍA-RUIZ, 2015). Apesar dos resultados promissores dos trabalhos apresentados, essas metodologias dependem do conhecimento prévio das classes das canetas, ou seja, para uma aplicação futura, seria necessária a criação de um banco de dados suficientemente representativo para conseguir diferenciar as tintas. Assim, surge a importância de estudar o potencial de técnicas exploratórias e não supervisionadas para diferenciação de tintas.

As técnicas HCA e PCA estão entre as ferramentas mais utilizadas entre as técnicas não supervisionadas e, além de diversas aplicações em outras áreas de conhecimento, também foram utilizadas na tentativa de discriminar tintas de canetas (ADAM; SHERRATT; ZHOLOBENKO, 2008; BORBA; HONORATO; DE JUAN, 2015; SILVA et al., 2012; THANASOULIAS; PARISIS; EVMIRIDIS, 2003). Payne e colaboradores também utilizaram Imagens Hiperespectrais na região do visível (400-1100nm) e PCA para discriminar tintas de canetas pretas e azuis (PAYNE et al., 2005). Os autores conseguiram discriminar 94% das tintas estudadas.

Nos trabalhos citados, nem todas as marcas de canetas são facilmente diferenciadas pelas ferramentas utilizadas, o que não significa que a informação necessária para diferenciar as amostras não esteja presente no conjunto de dados. Na química analítica, métodos não supervisionados de análise multivariada de dados como PCA e HCA são amplamente utilizados para acessar as informações mais importantes de um conjunto de dados e visualizá-lo em dimensões reduzidas. HCA e PCA buscam descrever os conjuntos em termos de distâncias e variância, respectivamente, porém essas nem sempre são as melhores formas de abordar o problema estudado. A informação relevante para a análise do conjunto de dados nem sempre está diretamente relacionada com a variância ou com a distância entre as amostras. Por essa razão, a técnica de *Projection Pursuit* (PP) se apresenta como alternativa neste tipo de análise (HOU; WENTZELL, 2011).

Como mencionado anteriormente, essa ferramenta procura por vetores de projeções que sejam capazes de mostrar estruturas de interesse nos conjuntos de dados, indicando a presença de informação relevante. Essas estruturas de interesse vão depender do objetivo do problema abordado e serão caracterizadas pelo índice de projeção. O desempenho da técnica de PP pode alcançar uma alta eficiência quando o número de amostras é maior do que o número de variáveis, quando essa

razão diminui, a eficiência da técnica pode diminuir significativamente (HOU; WENTZELL, 2014). Gou e colaboradores empregaram Algoritmo Genético na seleção de variáveis informativas para a construção de modelos PP, as variáveis foram selecionadas avaliando os gráficos dos escores de PP que forneciam as estruturas de interesse para a análise (GUO et al., 2001).

Na química analítica, especialmente quando dados espectroscópicos ou cromatográficos são objetos de estudos, o número de variáveis é geralmente maior que o número de amostras. Assim, para contornar esse problema e utilizar toda a informação disponível, um modelo PCA pode ser utilizado como um método de redução de dimensionalidade para obter um conjunto de dados com uma razão de amostras/variáveis apropriada para a construção de um modelo PP eficiente. A grande questão é determinar o número apropriado de variáveis latentes a ser usado para a construção do modelo. Se o número for menor que o ideal, há perda de informação. Se um número elevado de componentes for empregado ocorre o problema de sobreajuste. Para encontrar o número apropriado de componentes é necessário realizar uma série de análises, que podem consumir bastante tempo. Desta maneira, é possível realizar uma comparação entre espaços sucessivos obtidos a partir de modelos PP com diferentes números de PCs e encontrar regiões de estabilidade que permitam a escolha apropriada do número de componentes que deve ser usado em um determinado conjunto de dados para a construção de um modelo PP eficiente.

A Análise de Procusto (GOODALL, 1991) consiste numa ferramenta que reúne um conjunto de operações matemáticas que pode ser usada para comparar dois conjuntos de dados diferentes. Assim, é possível utilizar a Análise de Procusto para facilitar o uso da técnica de *Projection Pursuit*. Desta forma, o presente trabalho propõe uma nova metodologia para facilitar o uso de *Projection Pursuit*, usando a Análise de Procusto com ferramenta para o diagnóstico dos modelos PP e aplicar esta estratégia na diferenciação de tintas de canetas (WENTZELL et al., 2015).

2.2 OBJETIVOS

O objetivo deste trabalho consiste em avaliar a utilização de uma nova metodologia por reconhecimento de padrões não supervisionados para diferenciação

de tintas de canetas por infravermelho médio. Para isso, do ponto de vista quimiométrico e de aplicação, serão avaliados, respectivamente:

- O potencial da Análise de Procusto como ferramenta para o diagnóstico de análises de *Projection Pursuit*.
- A capacidade da espectroscopia no infravermelho associada à técnicas não supervisionadas de discriminar tintas de canetas azuis de diferentes tipos e marcas por meio de técnicas não supervisionadas.

2.3 METODOLOGIA

Para a realização das análises, canetas azuis de diferentes tipos e marcas foram adquiridas em estabelecimentos comerciais em Recife e estão descritas na Tabela 1. Foram utilizadas cinco marcas de canetas do tipo esferográfica, três do tipo gel e duas do tipo *rollerball*. Para dar variabilidade à análise, 10 canetas de cada marca, sendo 5 provenientes do mesmo lote e 5 de lotes diferentes, foram empregadas. Cada caneta foi usada para realizar 5 registros em papel branco (papel sulfite CHAMEQUINHO$^{@}$ tipo A4) e 2 espectros foram adquiridos para cada registro. Dessa forma, ao final da aquisição de todos os espectros, uma matriz contendo 1000 amostras foi usada para análise (SILVA et al., 2012).

Tabela 1 Descrição dos tipos e marcas das canetas empregadas neste trabalho.

Tipo	Marca	Código
Esferográfica	Bic	Ebc
	Compactor	Ecp
	Mitsubishi	Emi
	Pentel	Epe
	Pilot	Epi
Rollerball	Bic	Rbc
	Compactor	Rcp
Gel	Bic	Gbc
	Faber-Castell	Gfc
	Molin	Gmo

Esses dados são os mesmos descritos em trabalhos prévios do grupo, nos quais foram utilizadas técnicas de reconhecimento de padrões supervisionadas, associadas à seleção de varáveis espectrais (SILVA et al., 2012). No presente

trabalho, dos 1000 espectros inicialmente adquiridos, foram usados apenas 10 de cada lote e as análises foram realizadas de duas formas: (i) utilizando apenas 4 marcas (Rcp, Gbc, Gfc e Gmo), que são as mais difíceis de diferenciar, e (ii) utilizando todas as marcas de canetas de uma vez. Dessa forma, as matrizes utilizadas nas análises (i) e (ii) contêm, respectivamente, 240 e 600 amostras antes da remoção de anomalias.

Para a aquisição dos espectros, o acessório de Refletância Total Atenuada (UATR: *Universal Attenuated Total Reflectance*) do espectrômetro de infravermelho *Spectrum* 400 da *Perkin Elmer* foi utilizado. Os espectros foram obtidos na faixa de 4000 a 650 cm^{-1}, com resolução de 4 cm^{-1} e 16 varreduras por espectro. As análises foram realizadas diretamente no papel, sem danificar as amostras e sem necessitar nenhum preparo.

Após a etapa de pré-processamento, o conjunto de dados foi submetido às sucessivas PCAs, utilizando uma quantidade progressiva de PCs que variaram de 3 até 240, para a análise das 4 marcas e de 3 a 100, para a análise de todas as marcas. As análises PP foram realizadas com as matrizes dos escores das PCAs, ou seja, 238 (para as 4 marcas citadas) e 98 (para todas as marcas) análises PP foram realizadas, com diferentes níveis de compressão. O algoritmo univariado foi utilizado para avaliar a separação das quatro marcas e o algoritmo multivariado para todas elas.

Para construir o mapa de Procusto, os gráficos dos escores da análise PP utilizando k PCs e k+1 PCs foram comparados por análise de Procusto e os níveis de similaridade entre eles foram avaliados. Um mapa mostrando a dissimilaridade entre as projeções adjacentes foi construído com o objetivo de identificar regiões de estabilidade.

Todos os tratamentos quimiométricos foram realizados no software Matlab®. As funções para a análise de *Projection Pursuit* que foram utilizadas no presente trabalho foram desenvolvidas por colaboradores do grupo (HOU; WENTZELL, 2011). Para análise Procusto foi utilizada a função disponível no Matlab Statistics Toolbox.

2.4 RESULTADOS E DISCUSSÃO

2.4.1 Análise e Pré-processamento espectral

A Figura 17 mostra os espectros médios de infravermelho das amostras de canetas para cada uma das marcas adquiridas. É possível perceber a grande semelhança presente nos espectros, mesmo provenientes de diferentes marcas. As maiores diferenças entre as marcas se encontram na faixa de 2000-650 cm^{-1}.

Bandas características da celulose podem ser identificadas nos espectros da Figura 17: a região em torno de 3300 cm^{-1} pode ser atribuída à vibração intermolecular do H na ligação OH-O; as bandas referentes à deformação angular de C-H podem ser observadas entre 1500-1300 cm^{-1}; enquanto que as bandas relacionadas com o estiramento de C-O encontram-se em 1030 cm^{-1} (Tabela 2). Portanto, como é possível observar na Figura 17, a região do espectro que contém as informações mais importantes para diferenciar as tintas compreende os números de onda entre 1700-650 cm^{-1} (ALI et al., 2001; HAJJI et al., 2016; ZIEBA-PALUS et al., 2016).

A Figura 18 mostra em maior detalhe a região espectral que fornece as maiores diferenças entre as tintas de diferentes marcas. A atribuição de bandas nesse caso é difícil de ser realizada, pois as composições são desconhecidas e os espectros das tintas apresentam grande sobreposição com o espectro da celulose.

Tabela 2 Atribuição de bandas no IR para a celulose (ALI et al., 2001; ZIEBA-PALUS et al., 2016)**.**

Absorção (cm^{-1})	Atribuição
1030	C-C, estiramento C-OH, anel CH
1105	Estiramento glicosídico C-O-C
1160	Deformações $C^{-}OH$ e $C-CH_2$; estiramento assimétrico C-O-C
1315	Deformação *rocking* CH
1370	Deformações O-H e CH_2; CH dobramento no plano
1430	Deformação tesoura CH_2; deformação CH, OCH dobramento no plano
2900	Estiramento simétrico CH em CH, CH_2, CH_3
3300	Ligação H intramolecular OH-O

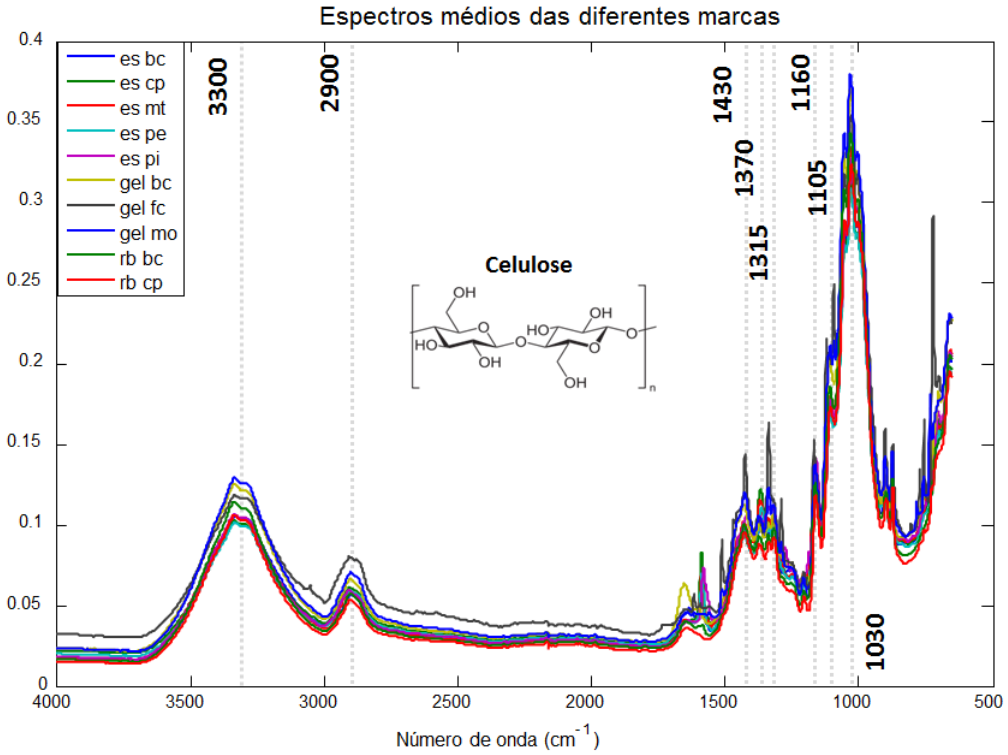

Figura 17 Espectros médios em MIR-ATR das 10 marcas diferentes de canetas.

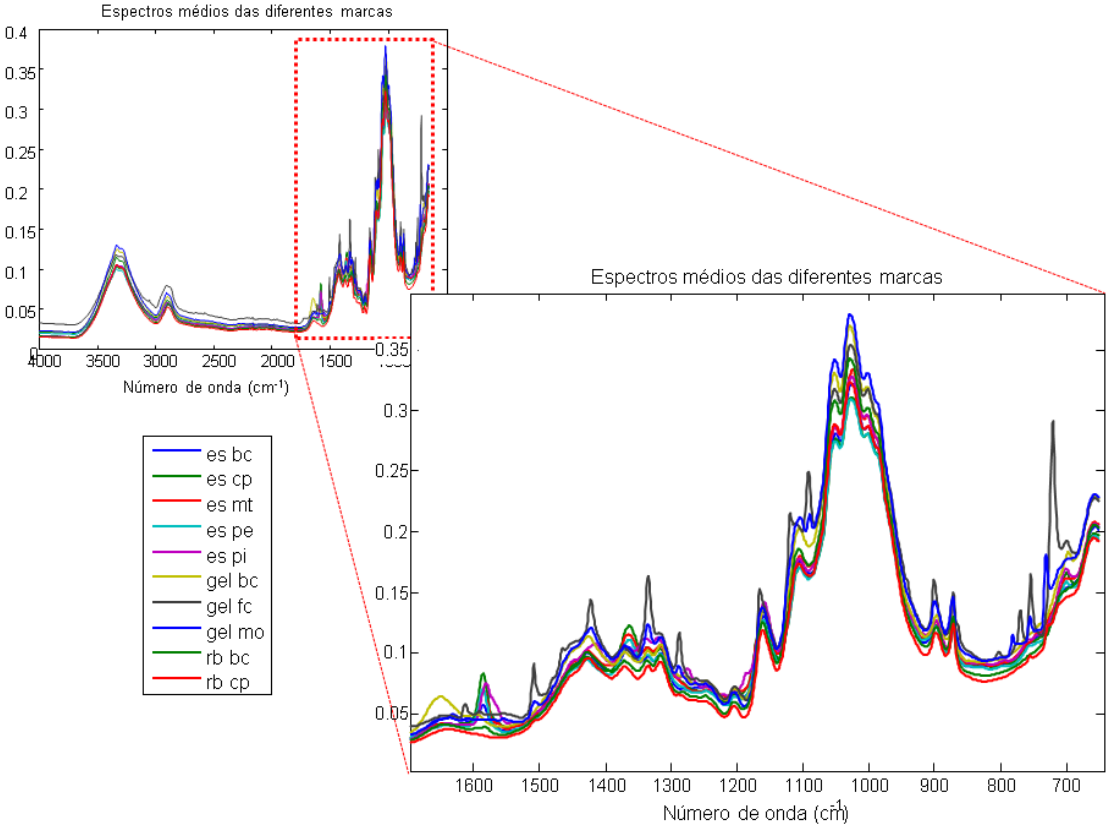

Figura 18 Detalhe dos espectros médios das 10 marcas diferentes de canetas (1700-650cm^{-1}).

Para corrigir efeitos de espalhamento, comumente encontrados em espectros obtidos por técnicas de refletância, técnicas como SNV, MSC (MSC: *Multiplicative Signal Correction*) e derivadas foram avaliadas. As ferramentas de pré-processamento como normalizações, centralização na média, etc., não interferem na construção de um modelo PP (HOU; WENTZELL, 2014), porém os modelos PP serão construídos com base na matriz de escores da PCA, que por sua vez é extremamente sensível a técnicas de pré-processamento. Dessa forma, a análise dos espectros e a escolha adequada do pré-processamento pode ser crucial para a obtenção de resultados adequados.

Ao final das análises, a técnica SNV apresentou os melhores resultados após a obtenção dos vetores de projeção da análise PP. A Figura 19a mostra os espectros brutos e pré-processados com SNV e centrados na média (Figura 19b).

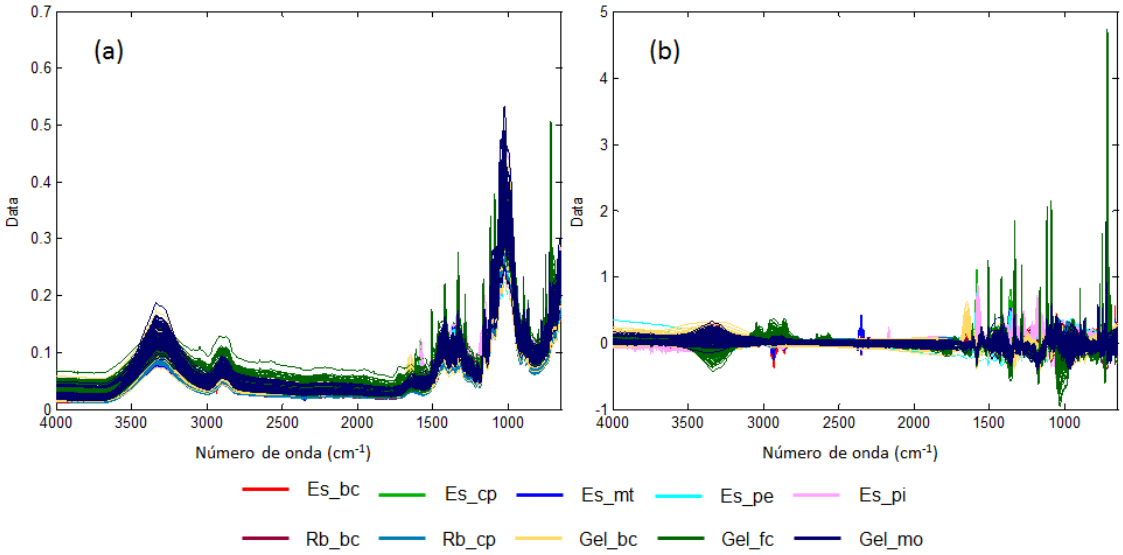

Figura 19 Efeito do pré-processamento SNV nos espectros das canetas. Os espectros (a) brutos e (b) pré-processados com SNV.

Ao observar os espectros brutos, verifica-se que não há uma grande variação de linha base nem efeitos multiplicativos significativos. Entretanto, é possível perceber de imediato a presença de efeitos aditivos, que podem estar relacionados com a própria técnica de aquisição espectral. É muito comum observar esses efeitos de espalhamento de radiação em amostras sólidas quando uma técnica de aquisição de espectros por refletância é empregada. Dessa forma, o SNV é capaz de atenuar essas

variações que não estão relacionadas com as características químicas das tintas e, após centrar na média, diferenças entre algumas marcas se tornam evidentes.

2.4.2 Análise de quatro marcas de canetas

As primeiras análises foram realizadas utilizando apenas 4 marcas de canetas (Rcp, Gbc, Gfc e Gmo). Essas marcas foram escolhidas por serem as mais difíceis de diferenciar entre as demais. A Figura 20a mostra o gráfico dos escores da PCA realizada para as 4 marcas citadas. As duas primeiras PCs representam 96% da variabilidade total dos dados, entretanto não é possível perceber as separações entre amostras de marcas diferentes.

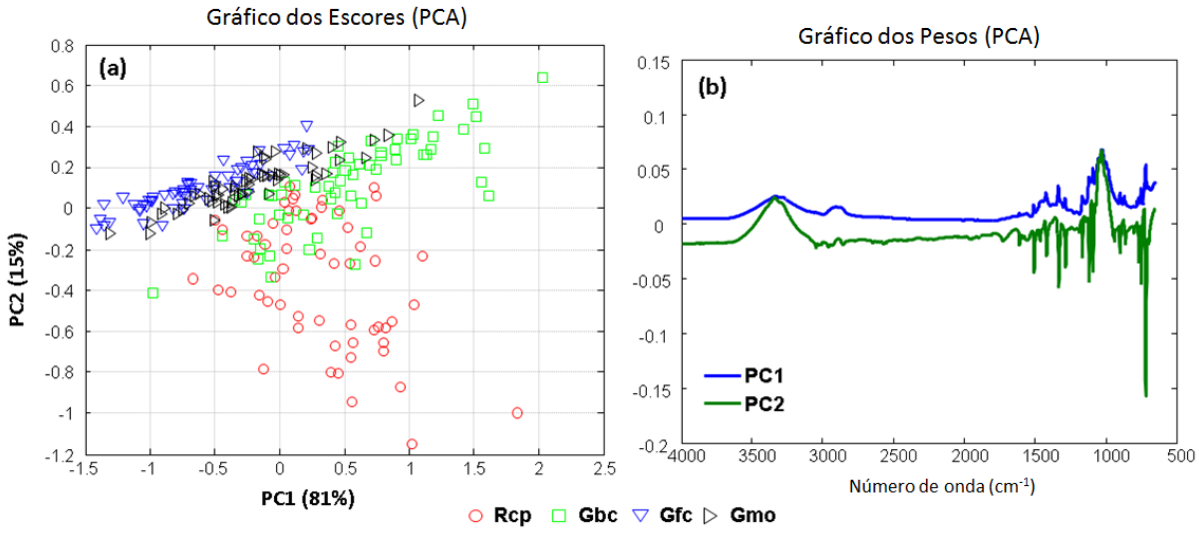

Figura 20 Gráfico dos (a) escores e (b) pesos das duas primeiras PCs, representando, respectivamente, 81 e 15% da variabilidade dos dados.

É importante ressaltar que as contribuições das variáveis originais no modelo PCA, representadas pelo gráfico dos pesos (Figura 20b), fornecem muitas informações relacionadas ao papel, como as bandas em torno de 3300 cm^{-1} (PC2) e a absorção em torno de 1000 cm^{-1} (PC1 e PC2). Uma absorção importante em 715 cm^{-1} pode ser observada na 2ª PC, associada à deformação do tipo *rocking* do grupo CH_2 (ZIEBA-PALUS et al., 2016).

O gráfico da variância explicada (Figura 21) sugere que 3 PCs são suficientes para explicar a maior parte da variabilidade dos dados. Porém, ao observar os gráficos dos escores bi- e tridimensionais nas Figura 20 e Figura 21b, respectivamente, é possível enxergar que a informação a respeito das diferenças entre as amostras das

diferentes classes não está evidenciada nas projeções em questão. Apenas as canetas Rcp parecem se diferenciar das demais.

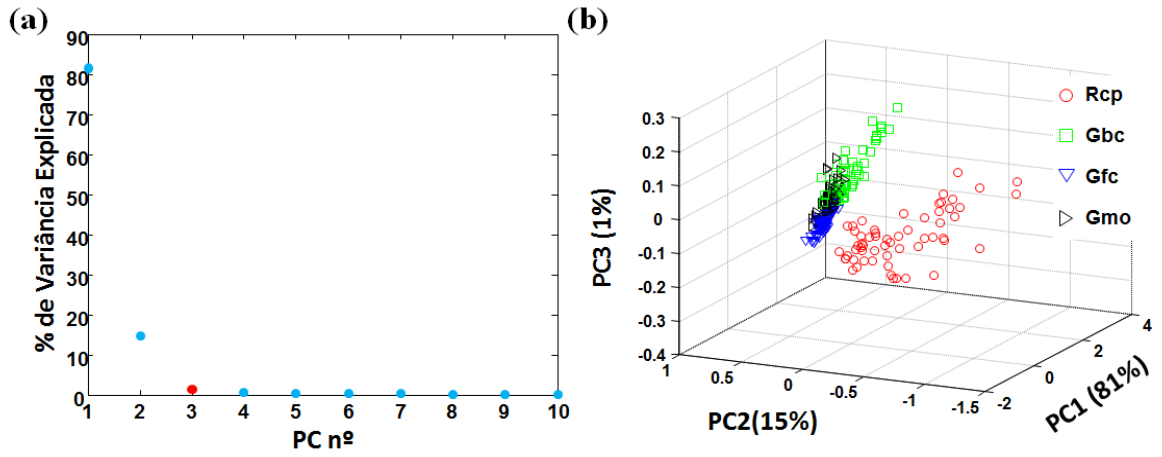

Figura 21 Gráficos (a) da variância explicada por cada PC (observando os valores até a 10ª PC) e (b) dos escores das 3 primeiras PCs.

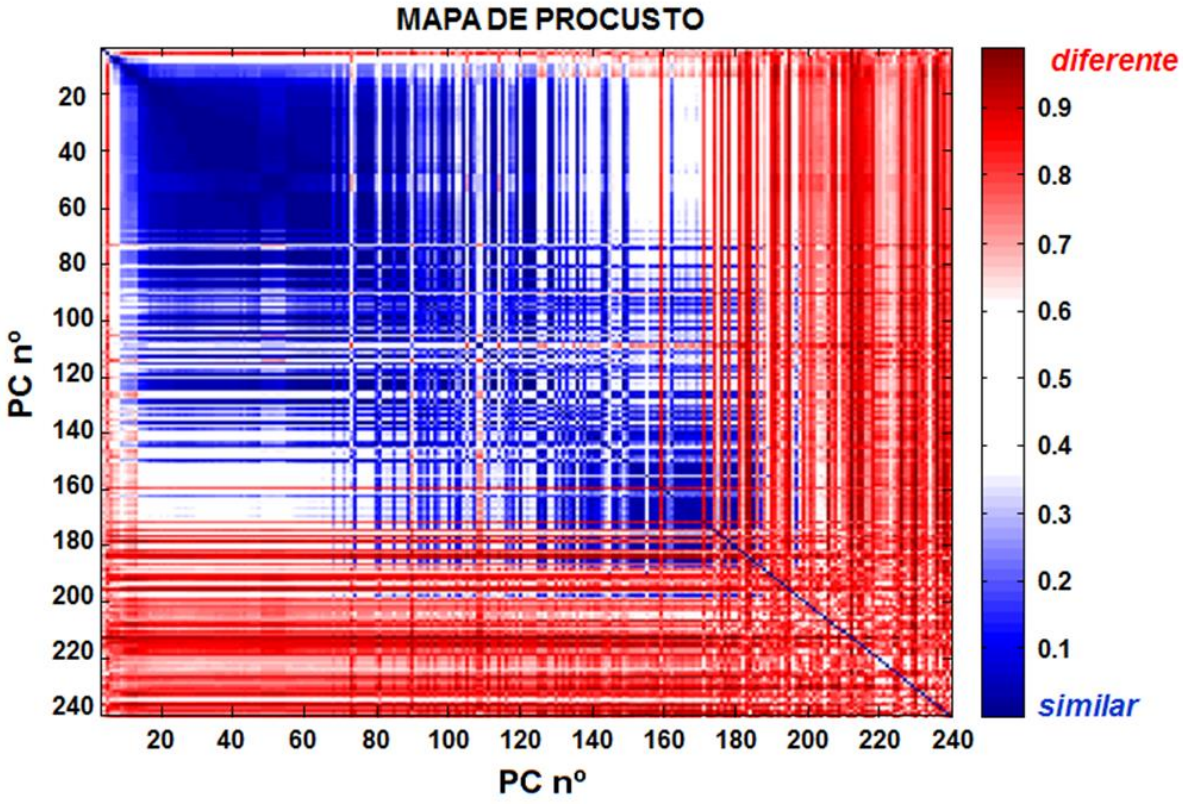

Figura 22 Mapa de Procusto das quatro marcas de canetas.

O mapa de Procusto foi então construído a partir dos escores das sucessivas análises PP, como mostrado na Figura 22. As regiões em azul do mapa mostram as regiões de estabilidade, o que significa que as projeções adjacentes possuem

estruturas semelhantes. As regiões em vermelho representam os valores mais altos de dissimilaridade, mostrando que as projeções podem não ser mais informativas, por se tratarem de projeções aleatórias (casos de sobreajuste).

O mapa construído sugere que as análises PP utilizando cerca de 15 a 70 PCs para compressão possuem estruturas semelhantes e podem ser informativas. É possível notar também que as regiões que utilizam poucas PCs para a compressão estão em regiões de instabilidade, pois não possuem informações suficientes para fornecer estruturas informativas. Esse gráfico funciona como uma ferramenta para monitorar o comportamento das projeções e identificar mais rapidamente as regiões do mapa (em azul) que, a priori, revelariam as estruturas de interesse, ou seja, revelariam os agrupamentos das amostras de mesma marca.

Os gráficos da Figura 23 ilustram a diferença entre as análises com diferentes níveis de compressão. Utilizando 6 PCs, fica claro que a compressão não fornece informação suficiente e a projeção das amostras tem uma distribuição que, apesar de apresentar tendências de separação, mostra uma mistura de todas as marcas. Utilizando 16 PCs (início da zona de estabilidade no mapa de Procusto), já é possível observar uma separação clara dos 4 grupos relativos às 4 marcas diferentes de canetas. Esse comportamento é observado até uma compressão com 67 PCs (final da zona de estabilidade). Quando muitas PCs são incorporadas na análise, informações não relevantes que se encontram nas últimas PCs forçam a separação de agrupamentos sem significados, representando bem o problema de sobreajuste.

Dessa forma, é possível afirmar que a separação das amostras de diferentes classes foi realizada com sucesso, utilizando a análise de Procusto para identificar o número adequado de PCs necessário para obter projeções que revelem os diferentes agrupamentos de amostras.

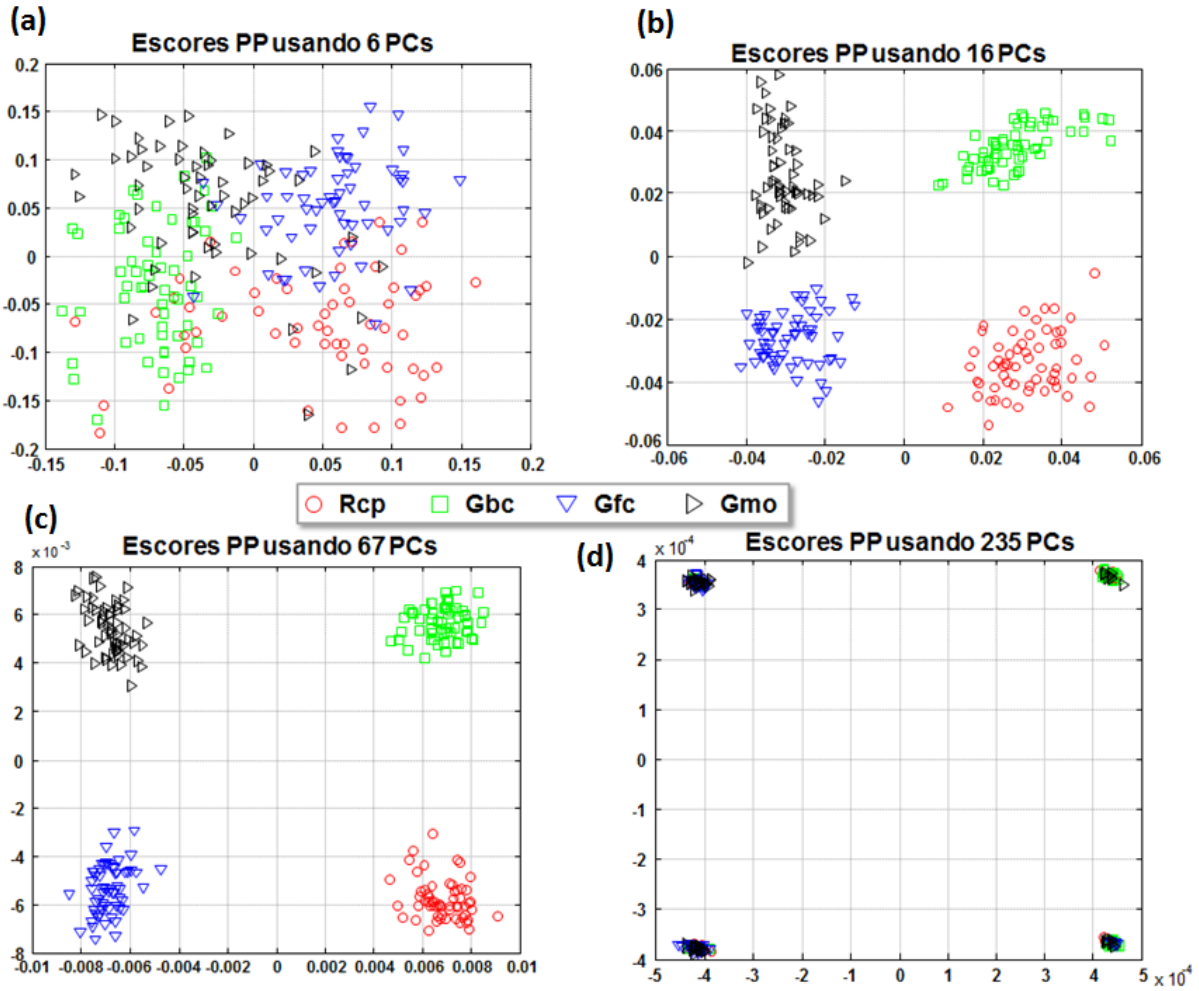

Figura 23 Gráficos de escores das análises PP usando um número diferente de PCs para cada nível de compressão dos dados: (a) 6; (b) 16; (c) 67 e (d) 235 PCs.

A Figura 24 mostra o gráfico dos pesos para os modelos PP construídos que correspondem aos gráficos dos escores apresentados na Figura 23. Percebe-se que, inicialmente, o modelo PP com 6 PCs, que por sua vez não mostra separação entre as marcas, ainda apresenta contribuições relacionadas ao papel (ver PP2 na Figura 24). Na medida em que mais informações vão sendo adicionadas, isto é, mais PCs são usadas para gerar um modelo PP, a importância das absorções em 3300 cm^{-1} e 1005 cm^{-1} vai diminuindo e dando espaço para a região da impressão digital, que apresenta a maior diferença entre os espectros das diferentes tintas. O aumento do número de PCs vai adicionando ruído ao modelo e levando a projeções sobreajustadas, caso do modelo utilizando 235 PCs na Figura 23. Portanto, apesar de todos os modelos construídos com nível de compressão de 16 a 70 PCs apresentarem estruturas semelhantes, o modelo mais adequado seria o mais simples, utilizando 16 PCs.

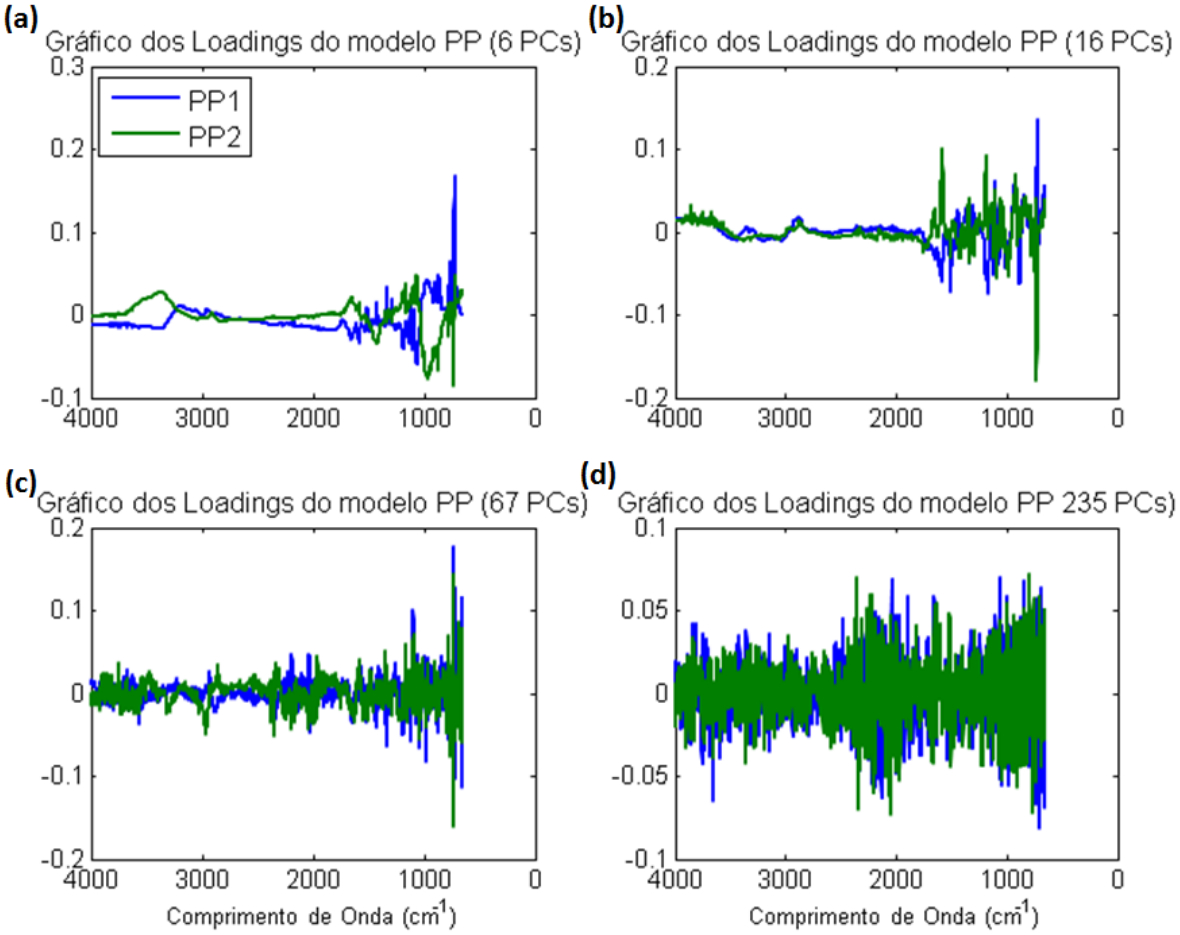

Figura 24 Gráficos de pesos das análises PP usando um número diferente de PCs para cada nível de compressão dos dados: (a) 6; (b) 16; (c) 67 e (d) 235 PCs. Modelo para as 4 marcas.

2.4.3 Análise de Todas as Marcas

A Figura 25 mostra o gráfico dos escores da PCA para todas as marcas de canetas. As 3 primeiras PCs explicam, respectivamente, 35%, 29% e 9% da variabilidade do conjunto de dados. Embora as diferenças entre os espectros das canetas sejam muito pequenas, é possível observar que existe uma tendência de separação das amostras Rcp com relação às demais. Já a semelhança entre os espectros das tintas das outras marcas pode ser identificada porque a informação da caneta é sobreposta pela do papel, como já mencionado.

Outro comportamento importante das amostras no gráfico dos escores está relacionado às amostras da caneta Gbc, que parecem estar levemente distanciadas das demais, mas não é possível distingui-las e observar um agrupamento claro apenas dessas amostras. É importante notar que, novamente, a variância pode não ser a melhor métrica para avaliar o conjunto de dados, pois a PCA não foi capaz de,

em uma projeção de dimensionalidade reduzida, revelar as separações de interesse (diferentes marcas).

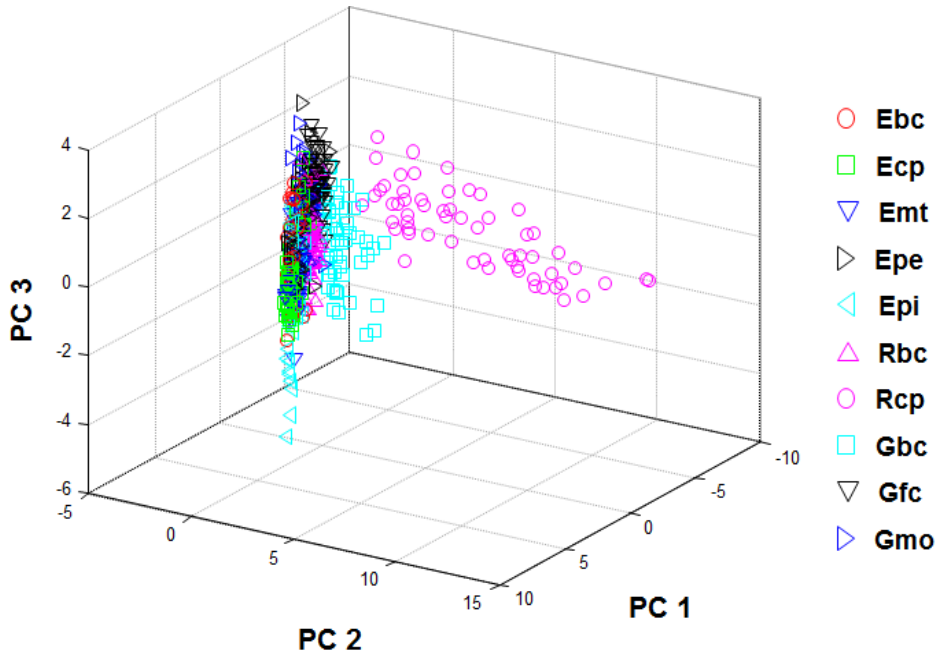

Figura 25 Gráfico dos escores das 3 primeiras PCs utilizando todas as marcas. As 3 primeiras PCs explicam 35,64%, 28,78% e 9,05% da variabilidade dos dados, respectivamente.

Observando os gráficos dos pesos (Figura 26b), é possível observar alguns picos característicos na faixa de 1700 a 650 cm^{-1}. Isso ocorre porque a região de 4000 a 2000 cm^{-1} está basicamente relacionada com a informação da celulose, que é comum a todos os espectros. Assim, a maior variabilidade dos dados está na região de impressão digital.

Observando o gráfico da Figura 26a, é possível notar o percentual de variância explicada por cada PC. O gráfico sugere que o número de componentes principais que explicam a maior variabilidade dos dados é cerca de 10 PCs, ou seja, as informações mais relevantes para retratar o conjunto de dados estariam representadas pelas 10 primeiras variáveis latentes do modelo PCA.

Entretanto, as projeções da PCA não evidenciam separações das amostras de uma forma relevante para o tipo de estudo que está sendo realizado. Ou seja, as projeções no sentido da maior variância não evidenciam a separação das marcas.

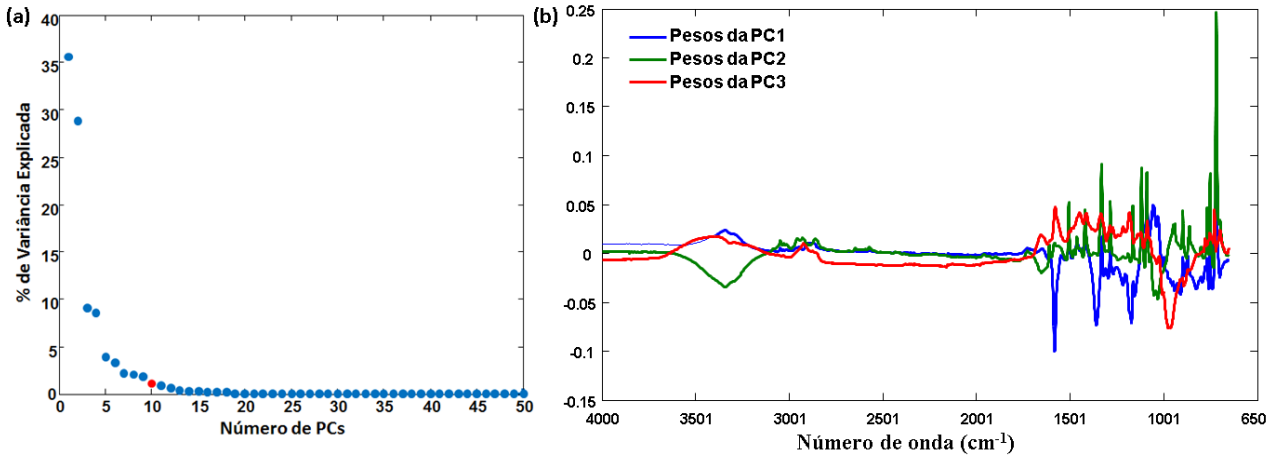

Figura 26 Gráfico (a) da variância explicada e (b) dos pesos das três primeiras PCs para a análise das 4 marcas.

A Figura 27 mostra o mapa da análise de Procusto realizado para comparar as projeções de escores utilizando todas as marcas de canetas. É possível notar algumas regiões de estabilidade no mapa (em azul), algumas menores como a compreendida entre PCs 23-30, 40-48 e uma região maior compreendida entre PCs 48-75, apesar de apresentar menor estabilidade.

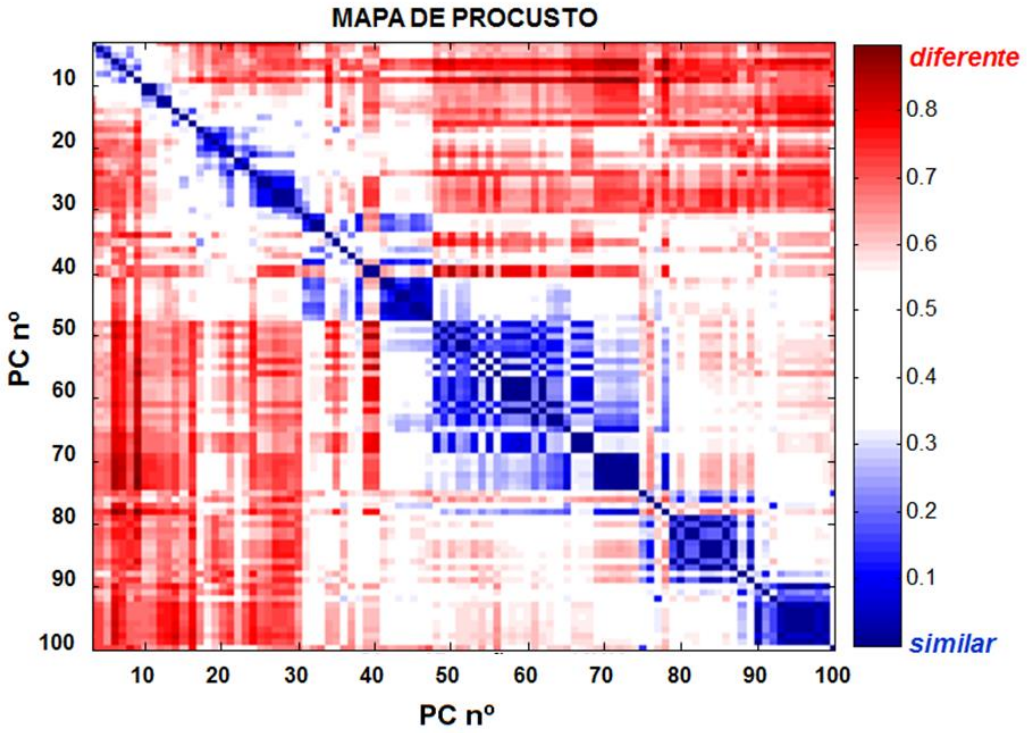

Figura 27 Mapa de Procusto e gráfico para todas as marcas.

Também na Figura 27, é possível perceber que nas regiões de estabilidade em que a análise é realizada com 80 PCs ou mais, o percentual de classificação das

amostras cai, mostrando que as projeções, apesar de semelhantes, não são informativas, pois, aparentemente, geram confusão de classes.

Na Figura 28, observam-se diferentes projeções tridimensionais de análise PP com diferentes níveis de compressão (6, 10, 44 e 98 PCs). A Figura 28a mostra que a projeção utilizando apenas 6 PCs para comprimir os dados não é informativa o suficiente para observar a separação dos agrupamentos. Na medida em que mais PCs são usadas para comprimir a matriz original de dados, a informação desejada parece ser revelada (Figura 28b e Figura 28c) até atingir o caso limite evidenciando projeções sobreajustadas (Figura 28d). Assim, é possível verificar que, a partir da região de estabilidade sugerida pelo mapa de Procusto, identifica-se que as projeções mais informativas estão entre as construídas utilizando de 40 a 70 PCs.

Nota-se que utilizando 10 PCs para a construção das projeções PP, número de PCs sugerido pelo modelo PCA, não é possível identificar uma separação clara das marcas de canetas, embora já exista uma tendência semelhante de comportamento de amostras de uma mesma classe; corroborando a ideia de que a métrica utilizada para construir modelos PCA (variância) não é a melhor maneira de abordar o problema de separação das tintas de canetas de forma não supervisionada. Assim, é possível utilizar a projeção construída com 44 PCs para observar uma melhor separação das marcas das canetas estudadas.

A Figura 29 mostra como os vetores de projeção mudam na medida em que mais PCs são adicionadas ao modelo. Da mesma forma que para a análise das 4 marcas, o modelo tende a ser desestabilizado quando mais PCs são adicionadas, forçando a uma projeção sobreajustada.

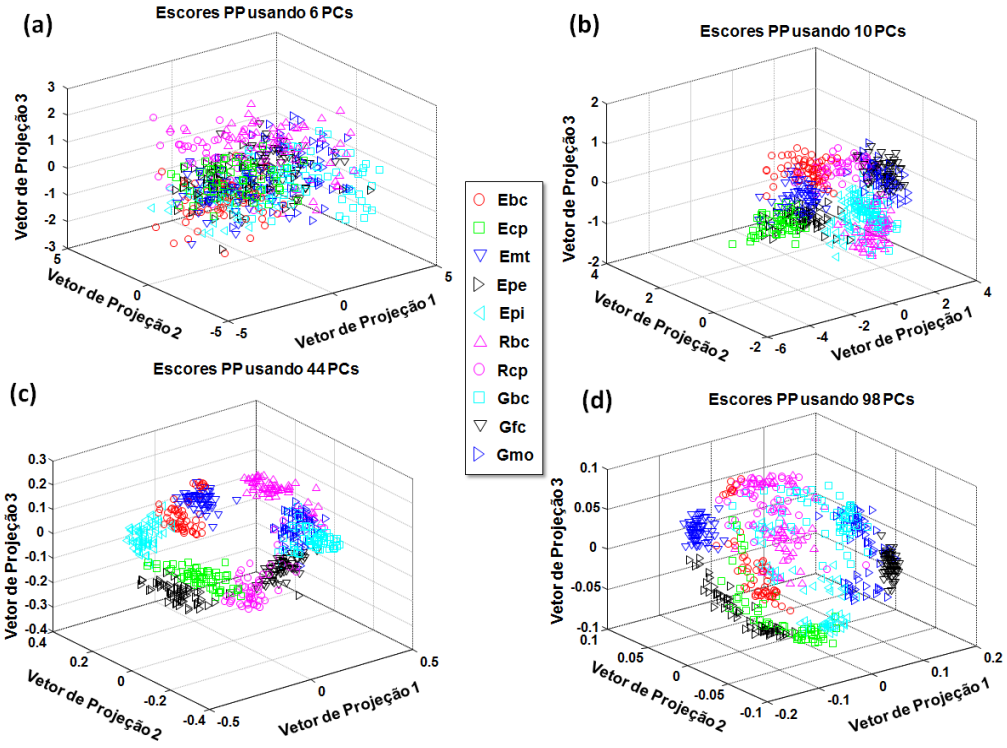

Figura 28 Gráficos de escores das análises PP usando um número diferentes de PCs para compressão dos dados para todas as marcas de canetas: (a) 6, (b) 10, (c) 44 e (d) 98 PCs.

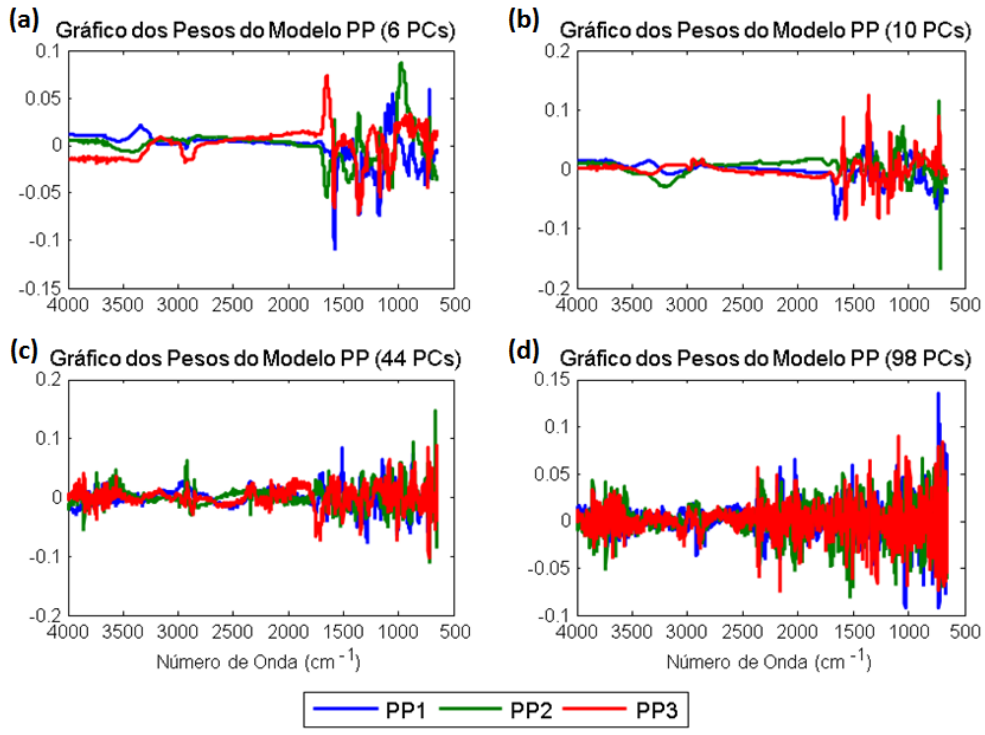

Figura 29 Gráficos de pesos das análises PP usando um número diferente de PCs para cada nível de compressão dos dados: (a) 6, (b) 10, (c) 44 e (d) 98 PCs. Modelo para todas as marcas.

2.5 CONCLUSÃO

Das técnicas de pré-processamento utilizadas na correção dos espectros, SNV foi a que mostrou o melhor desempenho a partir das projeções PP em todos os casos. No modelo PCA construído tanto para as quatro marcas, quanto para todas as marcas, não foi possível observar uma separação clara em projeções tridimensionais.

A análise de PP foi capaz de contornar esse problema, utilizando projeções tridimensionais capazes de revelar tendências de separação de todas as marcas analisadas. Para encontrar o nível de compressão mais informativo para o problema estudado, regiões de estabilidade no mapa de Procusto foram identificadas. A metodologia proposta foi capaz de mostrar o potencial de ferramentas não supervisionadas para identificar a separação de diversas marcas de canetas.

3 MODELOS DE CALIBRAÇÃO PARA DATAÇÃO DE DOCUMENTOS

3.1 INTRODUÇÃO

Outro problema associado com idoneidade de documentos é a datação, que consiste em um dos maiores desafios da área de documentoscopia (EZCURRA et al., 2010). Isso ocorre por conta da grande variedade de tintas e papéis disponíveis no mercado e complexo mecanismo de degradação, ainda desconhecido, fazendo com que o estudo do processo de envelhecimento seja muito complexo. Embora muitos grupos de pesquisa tenham se voltado para o estudo das tintas, a datação de documentos ainda carece de novas metodologias, principalmente no que diz respeito às mudanças ocorridas no papel.

Na medida em que o papel envelhece, diversas mudanças ocorrem em sua estrutura, incluindo degradação de carboidratos, mudanças no grau de polimerização da celulose, degradação por agentes biológicos, oxidação e hidrólise de compostos, entre outros (AREA; CHERADAME, 2011). Avaliar esses processos incluindo a grande variabilidade de compostos inorgânicos adicionados à superfície do papel e as condições de armazenamento propõem certamente um desafio na área de análise de documentos.

O estudo de fatores que atuam na mudança de coloração do papel na medida em que este envelhece, foi realizado por Schedl e colaboradores (SCHEDL et al., 2017). A partir desse estudo, um método utilizando espectrometria de massa foi proposto para quantificar o cromóforo DHAP (2,5-dihidroxiacetofenona) em amostras de celulose, documentos envelhecidos artificialmente e documentos históricos.

Trabalhos preliminares foram desenvolvidos por pesquisadores para avaliar a cinética das reações que ocorrem na composição do papel ao longo do tempo. Esses estudos, além de estabelecer uma relação entre envelhecimento artificial e natural, também estabeleceram relações entre o grau de polimerização da celulose e a mudança de algumas variáveis, como temperatura e acidez (ZOU; UESAKA; GURNAGUL, 1996a, 1996b). Outros grupos tentaram avaliar a mudança não só no grau de polimerização da celulose, mas também na variação dos compostos inorgânicos presentes empregando outras técnicas de análise como espectroscopia de fluorescência, difração de Raios-X e fluorescência de Raios-X por energia

dispersiva (HAJJI et al., 2016; KAČÍK et al., 2009; MARTÍNEZ et al., 2017), embora utilizando muitas vezes metodologias destrutivas.

Devido à relevância de se preservar a integridade do documento questionado, a espectroscopia vibracional se apresenta como uma alternativa aos procedimentos usuais, pois é rápida e não destrutiva (MURO et al., 2015). Alguns trabalhos já podem ser encontrados na literatura utilizando as técnicas de Infravermelho e fluorescência na análise papéis envelhecidos.

Ali e colaboradores (ALI et al., 2001) avaliaram o potencial da espectroscopia nas regiões NIR e MIR para a datação de documentos envelhecidos artificialmente de nove amostras de papéis diferentes. Os autores utilizaram uma razão entre picos característicos relacionados com a cristalinidade da celulose e avaliaram a mudança dessa razão com o tempo de envelhecimento. Antes do estudo de datação, os autores realizaram uma caracterização de diferentes materiais à base de celulose e, utilizando uma técnica discriminante não mencionada no artigo, tentaram identificar e eliminar as regiões espectrais relacionadas com as características dos diferentes materiais para manter apenas os efeitos relacionados com o envelhecimento. A partir do estudo de datação, os autores observaram duas mudanças características: (i) a mudança da cristalinidade com o tempo e (ii) o aumento da absorção do pico carbonila/carboxila ao longo do envelhecimento do papel. Um modelo PCR foi construído para as amostras analisadas na região NIR e obtiveram um SEP de 95h (aproximadamente 3% da faixa analisada).

Hajji e colaboradores (HAJJI et al., 2016) empregaram a espectroscopia na região MIR, além de fluorescência e difração de raios-X, para avaliar amostras de documentos também envelhecidos artificialmente e comparar os resultados com documentos restaurados datados dos séculos 16, 17, 18 e 19 também submetidos a condições extremas de armazenamento. Os autores realizaram um intenso estudo de atribuição de bandas por comparação espectral e investigaram como as condições experimentais afetaram as absorções da celulose ao longo do tempo. Mudanças relacionadas com a hidrólise, oxidação, e cristalinidade da celulose foram identificadas a partir dos espectros de FTIR. Os resultados obtidos com os documentos restaurados tiveram como objetivo evidenciaram mudanças drásticas em condições de alta temperatura e umidade relativa, demostrando que o processo de restauração

empregado não é eficiente para conservar documentos a longo prazo. Embora essa conclusão não tenha necessariamente aplicações forense, o estudo realizado é de grande valia para a identificação e caracterização de mudanças da celulose ao longo do tempo, ainda que as modificações ocorressem sob condições controladas.

Trafela e colaboradores (TRAFELA et al., 2007) publicaram um trabalho propondo o uso da espectroscopia na região MIR para datar documentos históricos e quantificar o grau de polimerização, pH, teores de cinzas, alumínio e de lignina utilizando modelos PLS. De acordo com os autores, mais de 170 amostras foram adquiridas para o conjunto de calibração, em que para o estudo de datação, 174 amostras foram selecionadas para o conjunto de calibração e 30 amostras para o conjunto de validação. Os modelos de regressão construídos para a quantificação dos compostos acima citados forneceram resultados adequados com coeficientes de correlação maiores que 0,90 para todos os compostos, exceto para o teor de alumínio (0,87). Para a construção do modelo de datação, duas faixas cronológicas foram analisadas: (i) documentos pré-1850 e (ii) documentos pós-1850. O erro padrão de previsão encontrado para o modelo de datação construído foi de 8,5 anos. Embora os autores tenham explorado uma grande faixa de intervalo de tempo analisado, diferentes documentos de um mesmo ano não foram utilizados no modelo de regressão. Para aplicações da área forense, é de extrema importância que a variedade de amostras de um mesmo ano seja explorada, para garantir que o modelo de regressão construído leve em consideração apenas as mudanças relacionadas com o envelhecimento e, neste caso, as mudanças que ocorrem na celulose.

Em particular, a degradação da celulose é um importante processo para estimar o processo de envelhecimento do papel, inclusive pelo fato de ser o composto majoritário sua presença independe do fabricante. Portanto, estudos que foquem nas mudanças desse composto podem ser extremamente informativos para a datação.

Alguns trabalhos podem ser encontrados na literatura reportando mudanças significativas no grau de polimerização (DP: *Degree of Polymerization*) da celulose com o tempo. O DP é medido pelo número de moléculas de celulose alinhadamente ligadas entre si. Na medida em que o tempo passa, essas ligações são enfraquecidas e quebradas por hidrólise, gerando formas amorfas das fibras, diminuindo o grau de polimerização (AREA; CHERADAME, 2011; WILLIAMS, 1981). Outros autores

reportaram que essa mudança no DP é refletida na diminuição do sinal das bandas 1425, 1370 e 900 cm^{-1} de seus espectros MIR (HAJJI et al., 2016).

Neste contexto, foi proposto o estudo de datação de documentos de diferentes épocas utilizando espectroscopia na região do infravermelho médio e análise multivariada para estimar a data de amostras de papel. Também foi proposta a avaliação de diferentes documentos de um mesmo ano.

3.2 OBJETIVOS

O principal objetivo do presente trabalho é avaliar o potencial da técnica de espectroscopia MIR-ATR para estimar a idade de um determinado documento desconhecido. Para isso, os seguintes objetivos específicos foram estabelecidos:

- Avaliar o potencial da técnica PCA para identificação de regiões do espectro que possam estar relacionadas com a tendência de envelhecimento do papel.
- Avaliar a importância de técnicas de pré-processamentos e seleção de variáveis para prever a idade de um determinado documento, baseado na análise do papel.
- Avaliar o potencial dos modelos de regressão PLS e sPLS para a datação de documentos com base na análise do papel.

3.3 METODOLOGIA

Para a realização deste estudo, documentos de 15 anos diferentes entre 1985 e 2012 foram fornecidos pela Polícia Científica Espanhola (Comisaría General de Policía Científica, Sección Documentoscopia, Madri, Espanha). Para cada ano, 5 documentos foram fornecidos contendo, em média, 5 folhas cada. Para cada folha de papel, 8 espectros foram adquiridos, sendo duplicatas em cada extremidade da folha (superior, inferior, esquerda e direita), fornecendo um total de aproximadamente 3000 espectros, como indicado na Figura 30.

As amostras foram divididas em conjunto de Calibração e de Previsão, em que um documento inteiro de cada ano e uma folha de cada um dos documentos remanescentes foram selecionados para fazer parte do conjunto de previsão (35,88% do conjunto de dados).

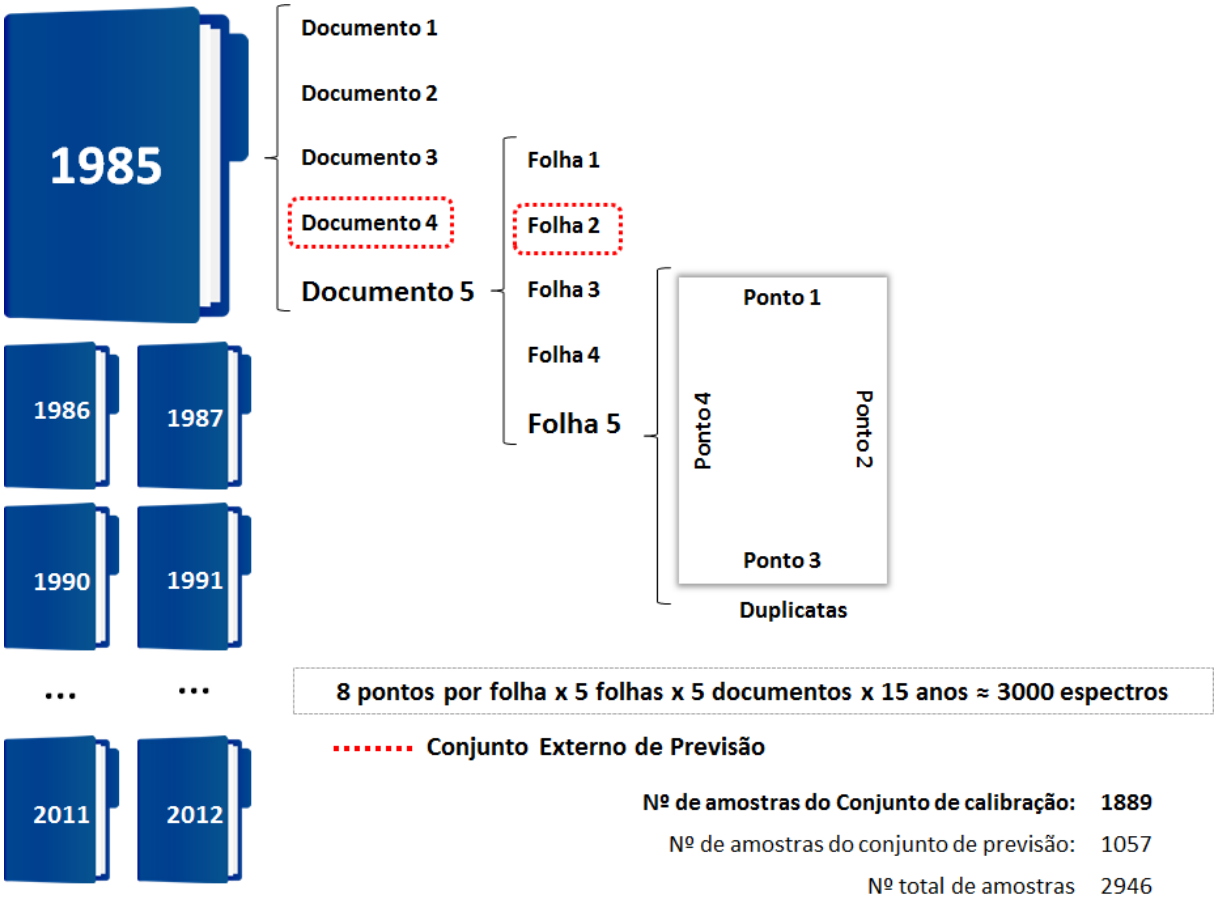

Figura 30 Esquema de aquisição de amostras e indicação do conjunto de Previsão.

Os espectros foram adquiridos na região do infravermelho médio utilizando o Espectrômetro *Nicolet iS10* da *Thermo* com o acessório de ATR *Smart iTR diamond*. Os espectros foram adquiridos na faixa espectral de 4000-650 cm^{-1}, com resolução 4 cm^{-1}, incremento de 0,482 cm^{-1} e 32 varreduras por espectro.

Em seguida, o conjunto de dados foi avaliado, pré-processado e modelos PLS foram construídos com o objetivo de prever o ano de cada documento. Os diferentes modelos foram comparados de forma a identificar diferenças significativas entre eles, utilizando para isto o teste F para comparar os valores de RMSEP e o teste t para identifica se os valores do *bias* de cada modelo são significativos.

Diferentes pré-processamentos foram utilizados com o objetivo de identificar e minimizar as diferenças significativas entre papéis de um mesmo ano. Para isso os filtros OSC, GLSW foram empregados e os modelos comparados. Em seguida, modelos sPLS foram construídos para comparação e visualização das variáveis mais influentes.

Para o emprego dos modelos esparsos, foram construídos modelos com combinações entre o número de variáveis esparsas latentes (sLV) e o número de variáveis incluídas (Var. Incl.). Os modelos foram construídos incluindo de 5 a 150 variáveis em cada sLV, sendo incluída a mesma quantidade de variáveis para cada sLV; e o número de sLV variou de 3 a 20. Superfícies de respostas contento o RMSEP e o R^2 foram avaliadas com objetivo de definir o modelo ótimo.

3.4 RESULTADOS E DISCUSSÃO

3.4.1 Análise Espectral

Após a aquisição dos dados, os espectros foram comprimidos para aumentar o incremento espectral. Essa compressão foi realizada utilizando a média da intensidade num intervalo definido por uma janela de 4 pontos, isto é, uma vez que a matriz de dados consiste em espectros contendo 6.949 canais espectrais, o espectro resultante apresentou 1.737 variáveis. Essa compressão de dados foi realizada para minimizar o ruído gerado devido ao incremento do equipamento que não podia ser alterado. Os espectros adquiridos podem ser observados na Figura 31.

As amostras de papel consistem em uma mistura complexa. Seu composto majoritário é a celulose, porém diversos compostos inorgânicos são adicionados ao papel durante o processo de fabricação, conferindo brilho, maciez e branquidão ao produto final. Dentre os compostos inorgânicos mais comuns utilizados, estão o carbonato de cálcio ($CaCO_3$) e a caulinita ($Si_2Al_2O_5(OH)_4$), cujas absorções na região do MIR são características e podem ser observadas nos espectros da Figura 31.

Nos espectros médios dos documentos é possível observar as absorções que estão relacionadas com os compostos inorgânicos. A banda característica da caulinita em 3660 cm^{-1} pode ser observada na Figura 31 nos espectros dos documentos dos anos 1986 e 1987, ainda com uma pequena contribuição no espectro do ano 1985. De fato, os documentos mencionados apresentam características diferentes dos demais, pois são papéis cujos instrumentos de impressão utilizados na produção dos documentos foi uma máquina de escrever; portanto, como esperado, a composição desses papéis diferiu dos demais. Também é possível notar no detalhe da Figura 31, a contribuição na região em torno de 875 cm^{-1} em todos os documentos, com exceção dos documentos datados de 1985 até1990 e de 1996.

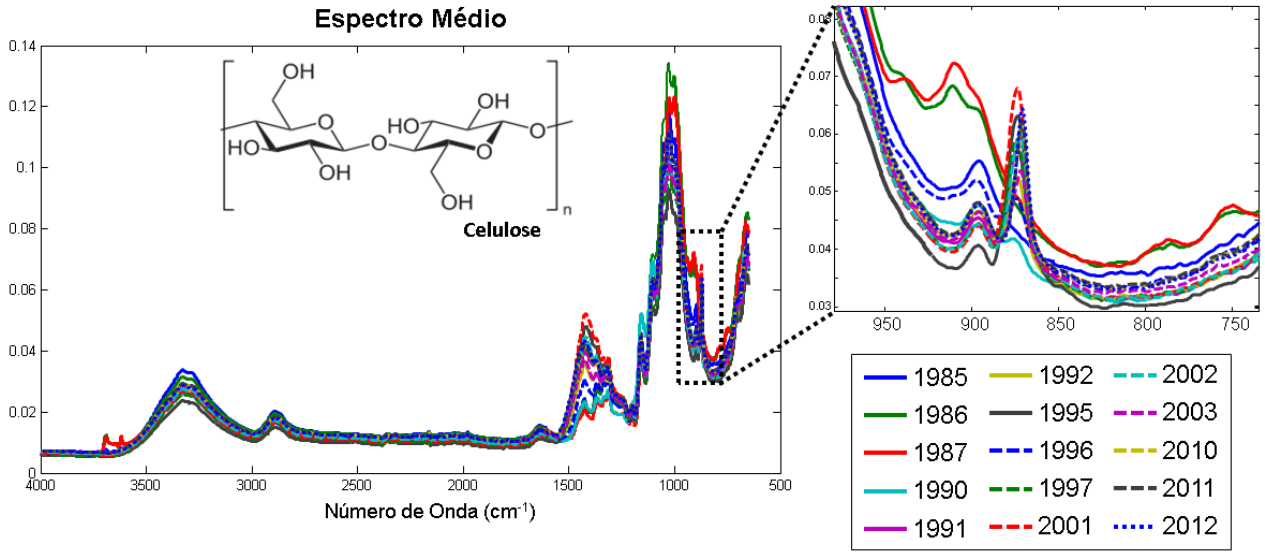

Figura 31 Espectros Médio dos Documentos de cada Ano. Detalhe da absorção relacionada ao carbonato de cálcio.

Além das contribuições da celulose, que foram discutidas na seção 2.4.1 do capítulo 2, é importante notar uma importante banda em torno de 1430 cm^{-1}, também característica da celulose e que está associada à despolimerização desse composto durante o processo de envelhecimento.

3.4.2 Pré-processamento Espectral e PCA

Diferentes técnicas de pré-processamento foram avaliadas neste estudo e o melhor resultado foi escolhido com base nos modelos de regressão PLS. Inicialmente, as técnicas de SNV e suavização (filtro Savitzky-Golay com polinômio de 2ª ordem e janela de 21 pontos) foram aplicadas para corrigir ruído e efeitos de espalhamento de radiação (Figura 32a e Figura 32b). Esse pré-processamento inicial já evidencia um comportamento sistemático das amostras de documentos de um mesmo ano, que pode ser posteriormente avaliado por PCA (Figura 33).

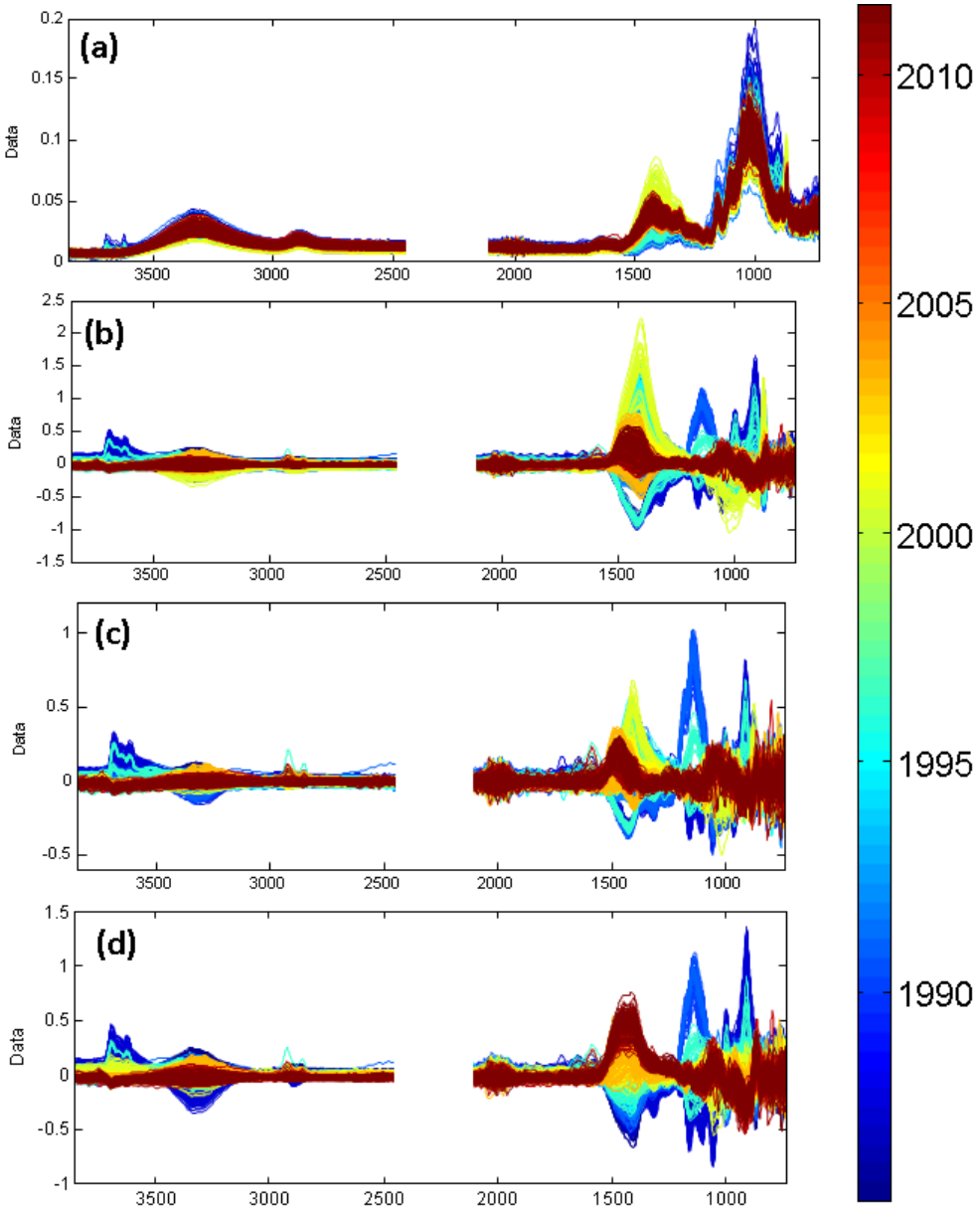

Figura 32 Espectros de acordo com os diferentes pré-processamentos. (a) Espectros brutos e pré-processados com (b) SNV, suavização e centragem na média, (c) SNV, suavização, GLSW (α = 2,9) e centragem na média e (d) SNV, suavização, OSC (1 componente) e centragem na média.

O modelo inicial de Componentes Principais mostra que as duas primeiras PCs explicam, respectivamente, 77,3 e 11,5% da variabilidade total dos dados. Dois agrupamentos podem ser observados no gráfico dos escores da Figura 33, o agrupamento de documentos mais antigos que apresentam valores mais negativos de escores para a 1ª PC e o agrupamento com valores mais positivos, composto pelos documentos mais recentes.

O gráfico dos pesos para o modelo PCA mostra que a banda de absorção em 1430 cm^{-1} está relacionada com os documentos mais recentes. Na 2ª PC, as contribuições negativas em 1000 e 900 cm^{-1} podem estar associadas também à caulinita (UDRIŞTIOIU et al., 2012). Embora esse comportamento possa ser observado, há uma grande variabilidade dos documentos de um mesmo ano. De fato, é de se esperar que, mesmo pertencentes a um mesmo ano, os documentos podem variar de forma significativa, pois compostos orgânicos e inorgânicos podem variar de acordo com as diferentes marcas de papel. Em contrapartida, a mudança da celulose ao longo do tempo não está associada com os componentes inorgânicos encontrados no papel e, portanto, a variabilidade de papéis de um mesmo ano deve ser suprimida de alguma forma.

O gráfico dos escores da Figura 33 mostra detalhes de como os documentos de um mesmo ano se comportam. Por exemplo, o documento do ano de 1991, que aparece selecionado em destaque no gráfico da Figura 33, representa um problema de amostragem importante que deve ser levado em consideração nesse contexto: a grande variabilidade dos diferentes documentos de um mesmo ano. Neste caso, a variabilidade dos documentos de 1991 é tão grande quanto a variabilidade do conjunto de dados total. Essa informação não está necessariamente relacionada a idade dos documentos. Por isso, métodos para suprimir essas informações devem ser empregados antes da construção dos modelos de regressão.

Duas ferramentas avançadas de pré-processamento foram utilizadas com o objetivo de atenuar as diferenças entre os documentos de um mesmo ano: o filtro GLSW e o OSC. E embora o critério de avaliação para a escolha do pré-processamento mais adequado tenha sido baseado nos modelos PLS, é possível observar algumas características interessantes que são geradas em decorrência do efeito do pré-processamento escolhido.

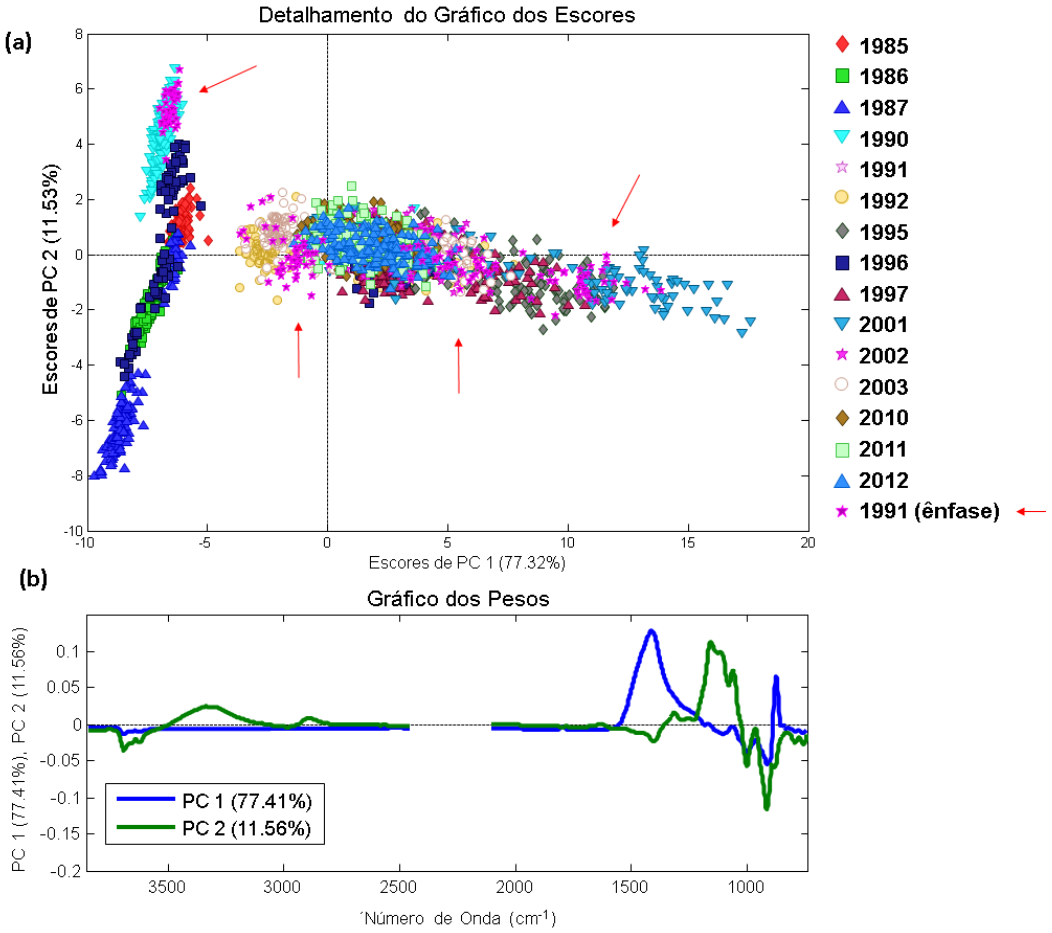

Figura 33 Modelo PCA para todos os documentos. Gráficos dos escores (a) e dos pesos (b).

As Figura 32c e Figura 32d mostram os espectros pré-processados com GLSW e OSC, respectivamente. Os espectros foram pré-processados com GLSW utilizando α=2,9, que é um valor alto para esse tipo de técnica, refletindo numa diminuição do efeito do filtro. Esse valor foi ajustado de forma a evitar que parte variabilidade relevante do conjunto de Calibração fosse removida. Como a variabilidade da interferência (variabilidade entre documentos de um mesmo ano) é da mesma ordem de grandeza da variabilidade total dos dados, o valor de α deve ser maior que o usual, para evitar a remoção de informação analítica útil dos espectros.

Para o emprego da técnica OSC, os modelos foram criados avaliando o número de componentes que remove a variabilidade do conjunto de dados que é ortogonal ao parâmetro estudado (idade). Modelos com 1 e 2 componentes foram criados e avaliados de acordo com os resultados fornecidos por PLS. O melhor resultado foi alcançado quando apenas um componente foi utilizado para representar a máxima

variância do conjunto de Calibração, que é ortogonal à variação das idades de cada documento.

Os espectros pré-processados com OSC já sugerem uma variação sistemática na intensidade de absorção na região de, aproximadamente, 1430 cm^{-1}. É possível perceber que documentos mais recentes apresentam valores mais altos de absorção nesta região e, na medida em que os documentos vão envelhecendo, a intensidade dessa banda diminui consideravelmente (Figura 32d).

3.4.3 Prevendo o Ano do Documento

Para prever a idade dos documentos, os modelos foram construídos utilizando o conjunto de calibração e, subsequentemente testados com o conjunto de previsão. Os modelos foram construídos com os pré-processamentos descritos acima e comparados. Os principais resultados obtidos podem ser vistos na Tabela 3.

Tabela 3 Resumo dos resultados dos modelos PLS para estimar o ano do documento com os diferentes pré-processamentos.

Pré-processamento	LV	RMSECV	RMSEP	$Bias_{CV}$	$Bias_{pred}$	R^2_{CV}	R^2_{pred}
SNV + Suav. +MC	4	4,6	4,4	-0,072	-0,489	0,72	0,75
SNV + Suav. + GLSW (0,48) + MC	2	4,5	3,8	-0,078	-0,472	0,74	0,82
SNV + Suav. + GLSW (1,2) + MC	2	4,5	4,0	-0,054	-0,463	0,73	0,79
SNV + Suav. + GLSW (2,9) + MC	2	4,6	4,5	-0,030	-0,366	0,71	0,74
SNV + Suav. + OSC (1) + MC	1	4,9	3,8	0,003	-0,173	0,71	0,81
SNV + Suav. + OSC (2) + MC	1	5,3	4,3	0,037	-0,392	0,66	0,75

*Entre parêntesis o valor de α (para GLSW) e o número de componentes (para OSC); RMSEP e RMSECV em anos.

É possível perceber de imediato a mudança no número de variáveis latentes do modelo PLS sem e com os filtros OSC e GLSW. Isso ocorre porque, como discutido anteriormente, os filtros removem parte da variância da matriz **X**, ou seja, há uma simplificação dos dados analisados, podendo refletir nessa diminuição de variáveis latentes.

Os modelos utilizando o filtro GLSW mostram uma pequena melhoria na medida em que o valor de α diminui (aumentando o efeito do filtro). Os valores de R^2_{prev}, bias e o RMSEP apresentem valores aparentemente melhores quando α = 0,48, e estatisticamente diferentes quando comparado com os modelos construídos com α = 1,2 e 2,9. Porém, os espectros pré-processados resultantes apresentam baixa razão sinal/ruído, não sendo considerado um modelo adequado (Figura 34).

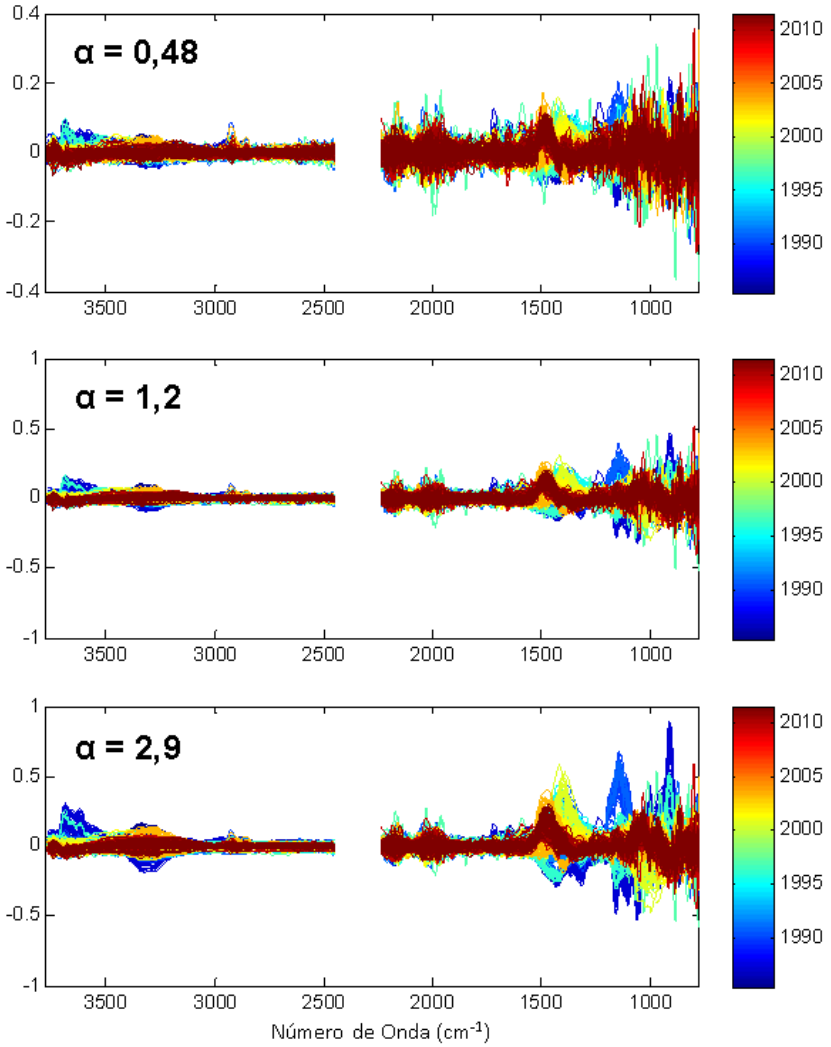

Figura 34 Efeito do α nos espectros pré-processados.

Como discutido anteriormente, a variabilidade dos documentos de um mesmo ano interfere significativamente no desempenho do filtro GLSW, pois a variabilidade dos documentos de alguns anos é da mesma magnitude da variabilidade total dos dados. Por esse motivo, quando o valor de α é baixo (= 0,48), ocorre uma remoção de informação útil. Dessa forma, outros pré-processamentos devem ser avaliados.

A Figura 35 mostra os gráficos de valores previstos *versus* reais para as amostras do conjunto de previsão e os gráficos de importância das variáveis (VIP) nos respectivos modelos. No que diz respeito ao teste entre os valores de RMSEP, os modelos construídos com os filtros GLSW e OSC são estatisticamente semelhantes, porém diferentes do modelo construído apenas com SNV e suavização. Em contrapartida, quando comparado com o modelo utilizando GLSW, o efeito do filtro OSC nas amostras (particularmente os documentos de 1990, 1996 e 2003 na Figura

35e) apresenta uma melhora. Percebe-se que o filtro faz com que as amostras de um mesmo ano se tornem mais próximas apresentando um efeito final mais eficiente do que o filtro GLSW. Isso significa que o desempenho da técnica OSC em lidar com as variações entre amostras semelhantes é melhor do que o desempenho do filtro GLSW.

Outro indício de que o modelo utilizando OSC é mais adequado que os demais, é o fato de os VIPs escores (Figura 35f) só apresentarem influência de variáveis que estão associadas à celulose. De fato, as variáveis que apresentam uma maior importância, em todos os modelos estão associadas às absorções em 1420 e 910 cm^{-1}. De acordo com a literatura (HAJJI et al., 2016), na medida em que o papel envelhece, a celulose perde cristalinidade e as bandas de absorção em 1430 e 900 cm^{-1} apresentam grande sensibilidade a essas mudanças.

A Figura 32d corrobora o gráfico dos VIPs escores, mostrando que documentos mais antigos apresentam uma menor absorção na banda de 1430 cm^{-1}, pois possuem baixa cristalinidade. Já os documentos mais novos, apresentam maiores absorções nessa região, ocasionada pela maior cristalinidade da celulose.

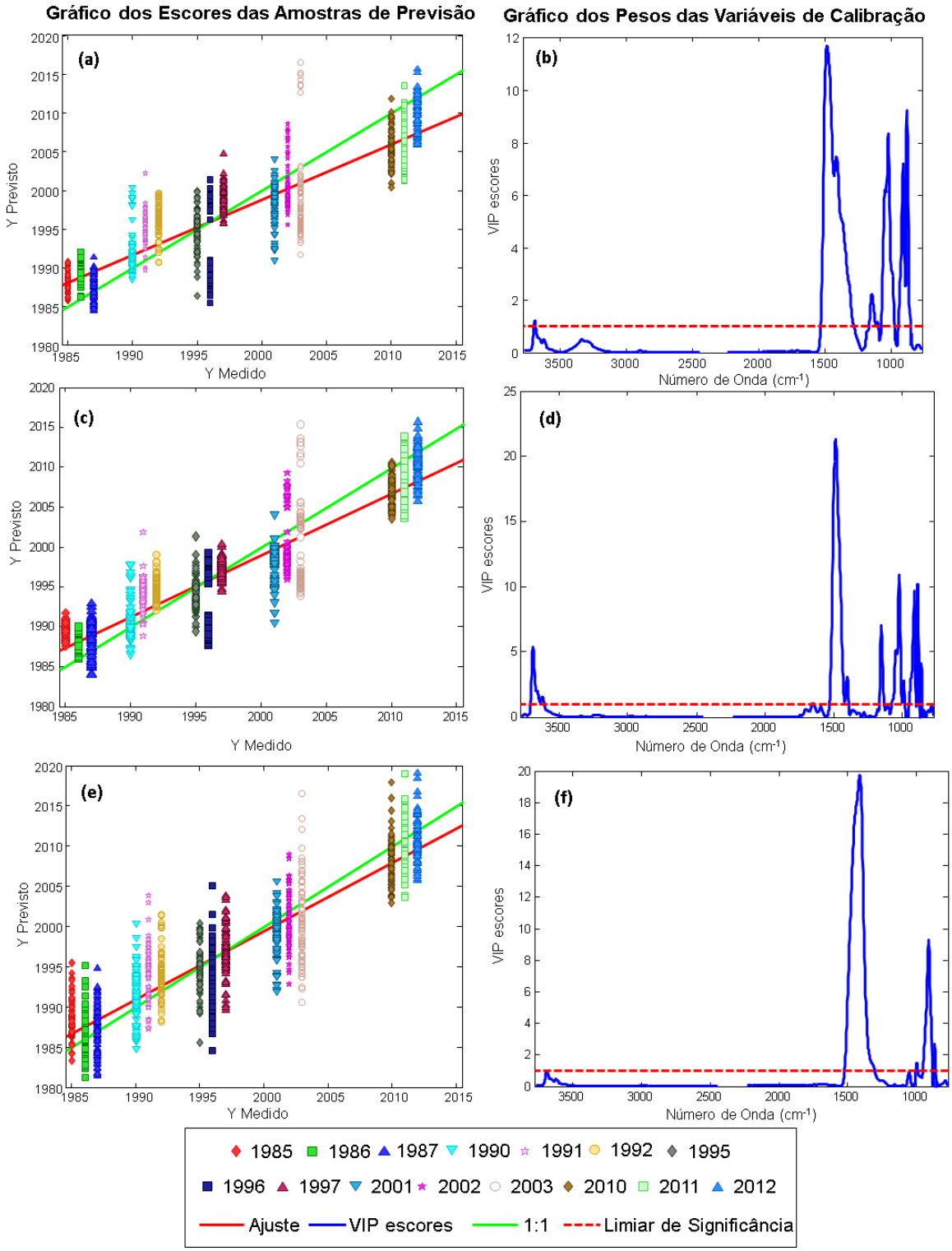

Figura 35 Gráficos de regressão (esquerda) e dos escores VIPs (direita) para os modelos pré-processados (a-b) SNV, suavização e centragem na média, (c-d) SNV, suavização, GLSW (α = 0,48) e centragem na média e (e-f) SNV, suavização, OSC (1 componente) e centragem na média.

Seleção de Variáveis

Modelos sPLS foram construídos com os espectros pré-processados com SNV, suavização (filtro Savitzky-Golay com polinômio de 2ª ordem e janela de 21 pontos) e centragem na média. As superfícies de resposta para o RMSEP e o R^2 dos modelos construídos podem ser observadas na Figura 36.

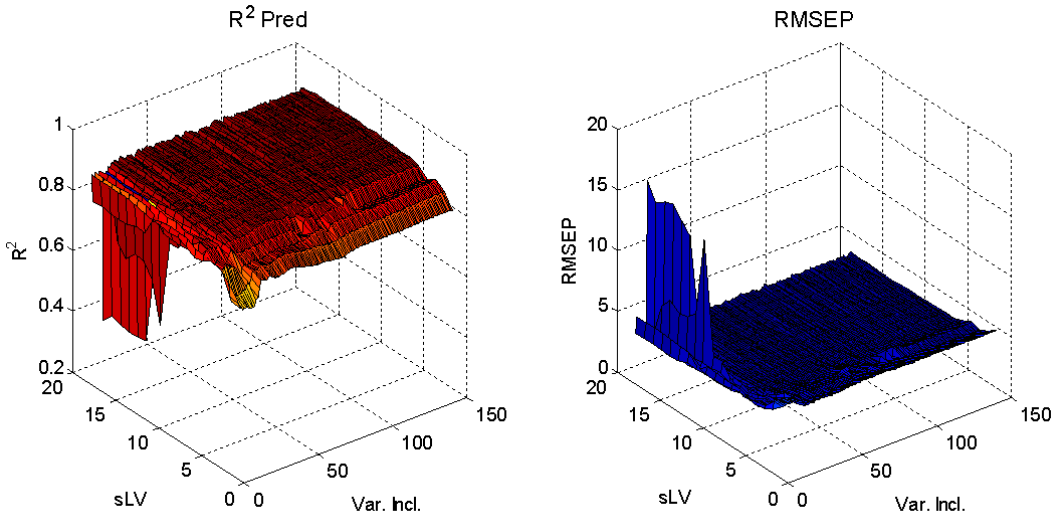

Figura 36 Superfícies de resposta dos modelos sPLS para os documentos.

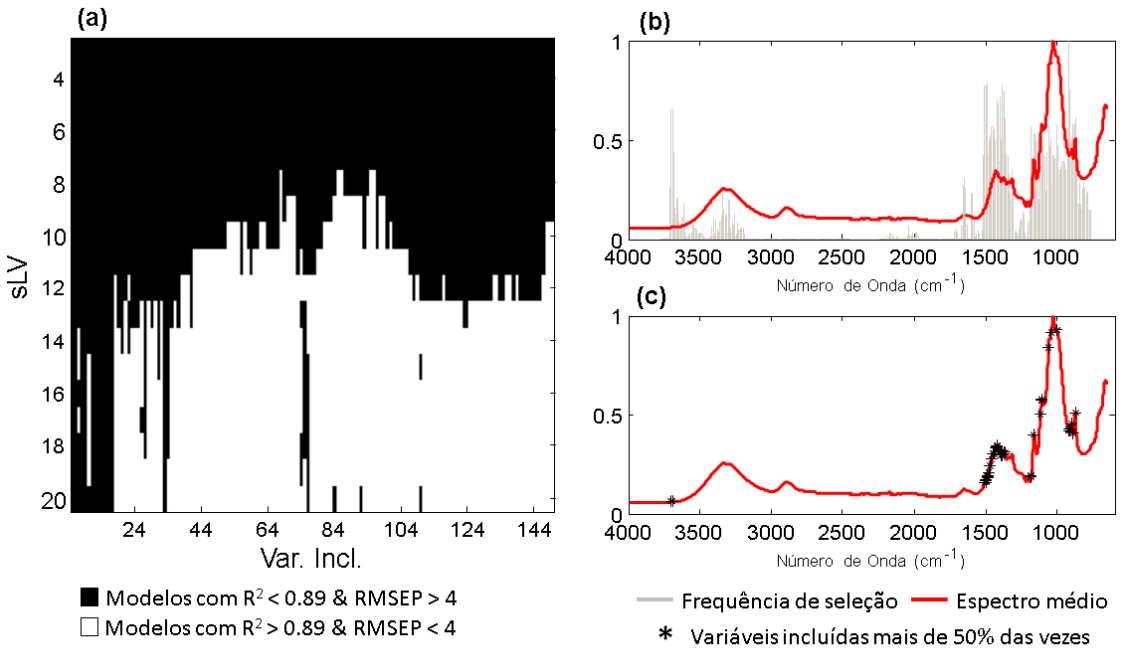

Figura 37 (a) Destaque em branco para as combinações de sLV e Var. Incl. que geram modelos sPLS com R^2 > 0,89 e RMSEP < 4; (b) o espectro médio dos documentos e a frequência de seleção das variáveis em todos os modelos construídos; (c) destaque para as variáveis selecionadas em mais de 50% dos modelos.

É possível observar que os modelos sPLS apresentam resultados estáveis (Figura 36), que não diferem muito entre si. Para otimizar os parâmetros, foram selecionados os modelos cujos valores de $R^2 > 0,89$ e RMSEP < 4 (Figura 37a). É possível observar que os modelos que obedecem esses critérios possuem uma grande quantidade de variáveis latentes, em geral 8 ou mais. Portanto, não é possível determinar um modelo otimizado, mas uma região de estabilidade contendo possíveis modelos capazes de fornecer resultados igualmente aceitáveis.

O gráfico da Figura 37b e Figura 37c mostram a frequência com que as variáveis são selecionadas nos modelos sPLS e as variáveis que foram selecionadas em mais de 50% dos modelos, respectivamente. Portanto, é possível admitir que os modelos mais adequados são aqueles que incluem as variáveis destacadas pela Figura 37c. E, portanto, é possível encontrar não apenas um modelo, mas uma região de combinações de sLV e variáveis incluídas que contém as variáveis em destaque.

A Figura 38 mostra os resultados de dois modelos escolhidos na região de estabilidade. É possível observar uma melhora em relação ao R2 e RMSEP quando comparados com os modelos PLS pré-processados com SNV e Suavização. Com relação aos modelos pré-processados utilizando os filtros GLSW e OSC, é possível observar também uma leve melhora em relação ao R^2, embora o teste estatístico para o RMSEP não acuse diferença significativa. Ainda assim, algumas considerações devem ser feitas.

É possível perceber que, quando 45 variáveis ou mais são incluídas, a maioria dos modelos sPLS seleciona a região da caulinita como importante. Isso ocorre porque este é um composto comum nos documentos antigos, deste conjunto de dados, mas não está necessariamente relacionado com o envelhecimento dos documentos. Uma vez pré-processados com OSC, a informação da caulinita é considerada como um interferente que pode variar entre documentos de um mesmo ano, assim o OSC enxerga que, na verdade, essa região não está relacionada com o envelhecimento do papel e é, portanto, eliminada. Assim, apenas regiões relacionadas com a celulose são consideradas relevantes para a criação do modelo PLS. Por isso, a regressão do modelo pré-processado com OSC apresentou resultados mais adequados para a o estudo da datação dos documentos.

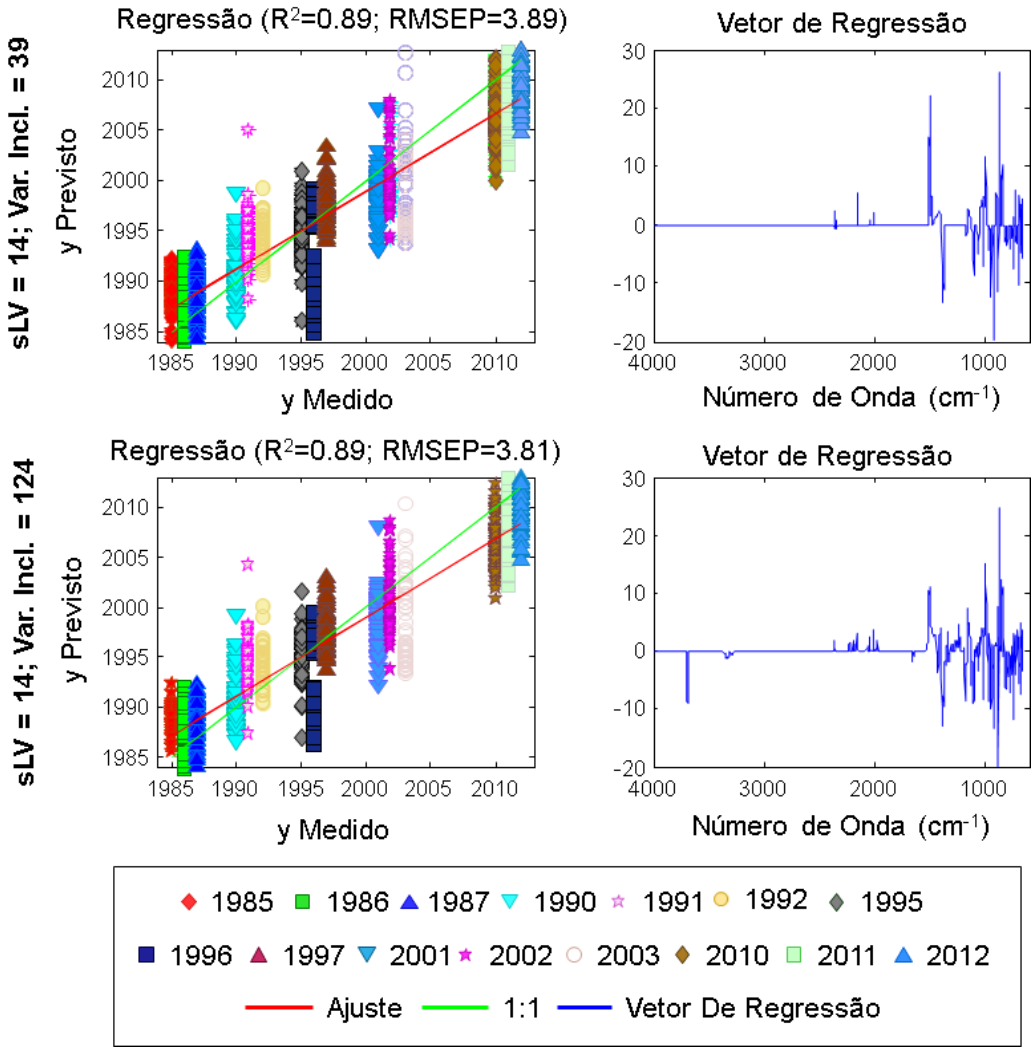

Figura 38 Comparação de dois modelos sPLS construídos com 14 sLV incluindo 39 variáveis (acima) e 124 variáveis (abaixo). Gráfico da regressão (esquerda) e os vetores de regressão (direita).

3.5 CONCLUSÃO

Para o problema de datação de documentos, a avaliação dos efeitos das técnicas de pré-processamento apresentou-se como a etapa mais importante para o estudo, pois a grande variabilidade existente entre os documentos de um mesmo ano foi atenuada com o uso de filtros como GLSW e OSC. O modelo PCA do conjunto de dados evidenciou a grande variabilidade existente em determinados documentos e evidenciou dois agrupamentos importantes relacionados com os documentos mais antigos e os mais recentes.

Para a construção dos modelos PLS foi possível observar que o filtro GLSW não foi capaz de atenuar de forma adequada as diferenças entre documentos de um

mesmo ano. Em contrapartida, o pré-processamento OSC conseguiu reduzir essas diferenças, fornecendo um modelo PLS com R^2 = 0,81 e RMSEP = 3,8 anos, valores adequados considerando a complexidade do conjunto de dados estudado e para as possíveis aplicações. Além disso, as absorções em 1420 e 910 cm^{-1} mostraram grande influência na construção do modelo que, de acordo com a literatura, são essas regiões de absorção relacionadas com a perda da cristalinidade da celulose durante o processo de envelhecimento.

A extensão do modelo PLS associado à uma técnica de seleção de variáveis, sPLS, também mostrou resultados promissores, fornecendo uma grande combinação de modelos possíveis que geram resultados adequados. Em compensação, a maioria dos modelos mostraram grande importância de variáveis associadas com compostos inorgânicos, como a caulinita. Como esses compostos não estão necessariamente relacionados com o envelhecimento da celulose do papel, o modelo PLS com o filtro OSC se mostra mais atrativo para aplicação proposta.

4 IMAGENS HIPERESPECTRAIS PARA A IDENTIFICAÇÃO DE SÊMEN EM TECIDOS

4.1 INTRODUÇÃO

A identificação de fluidos biológicos (como sangue, sêmen e saliva) é uma das etapas mais cruciais numa análise de cenas de crime. É importante não só detectar e confirmar a presença e a natureza desses fluidos numa cena de crime, mas também realizar isso de uma forma não destrutiva, pois estes vestígios podem ser empregados em análises futuras, visando a busca de padrões de DNA que possam levar à identificação da vítima ou do agressor. No que diz respeito à seletividade dos testes, a identificação desses fluidos pode ocorrer de duas formas: (i) presuntiva e (ii) confirmatória. As técnicas presuntivas são aquelas que, em geral, são sensíveis, mas não específicas. Ou seja, geralmente detectam vários falsos positivos. Em contrapartida, os testes confirmatórios são específicos, e identificam apenas o analito de interesse ou apresentam poucas interferências. Consequentemente, uma técnica confirmatória é geralmente necessária posteriormente para determinar a natureza do fluido e a coleta adequada da evidência (ZAPATA; FERNÁNDEZ DE LA OSSA; GARCÍA-RUIZ, 2015). A identificação de sêmen, por exemplo, é de grande importância na inspeção de cenas de crime de caráter sexual.

Um dos maiores problemas para a identificação e coleta adequada de fluidos seminais é a necessidade de imediata identificação das amostras para evitar degradação e perda de informação genética. Portanto, técnicas que sejam não destrutivas, rápidas, confiáveis e que possam ser utilizadas *in situ* são preferíveis, auxiliando nas decisões do investigador forense em campo na procura de evidências. Além disso, problemas advindos de má conduta de analistas e análises equivocadas, como comentado anteriormente, podem ser evitados por meio de metodologias objetivas que não exijam apenas a opinião de um perito. No caso de identificação de resíduos em cenas de crime, as fontes de luz alternativas (*Alternate Light Sources*) e equipamentos portáteis são essenciais para o reconhecimento imediato de resíduos de disparos, explosivos e fluidos biológicos, incluindo sêmen. Entretanto, algumas dessas técnicas, além de exibirem falsos positivos em alguns casos, podem ser demoradas devido às pequenas áreas que podem ser analisadas por vez (KUULA et al., 2012). Para contornar as desvantagens dos métodos atuais, técnicas analíticas

como as espectroscopias vibracionais vêm sendo cada vez mais empregadas como alternativa na busca por resíduos biológicos, conforme já discutido (MURO et al., 2015).

Virkler e Lednev utilizaram um espectrômetro Raman portátil com o objetivo de discriminar diferentes tipos de fluidos corporais, incluindo sêmen, fluido vaginal, saliva, sangue e suor (VIRKLER; LEDNEV, 2008). De acordo com os autores, a metodologia foi capaz de diferenciar os fluidos estudados por simples comparação espectral e atribuição de bandas. Eles também reportaram a capacidade de diferenciar sêmen canino e humano. Lednev e seu grupo de pesquisa têm pesquisado o potencial da espectroscopia Raman na identificação de fluidos biológicos em diferentes trabalhos (SIKIRZHYTSKI; SIKIRZHYTSKAYA; LEDNEV, 2012; SIKIRZHYTSKI; VIRKLER; LEDNEV, 2010). Entretanto, eles também reportaram a dificuldade de trabalhar com amostras de sêmen, afirmando que as manchas de sêmen podem ser difíceis de enxergar quando iluminadas por luz branca e, além disso, é difícil identifica-las por um único perfil espectral em Raman devido à sua natureza heterogênea (MCLAUGHLIN; LEDNEV, 2015).

Como mencionado anteriormente, os mapeamentos de grandes áreas por espectroscopia Raman podem ser extremamente demorados, fazendo com que a espectroscopia no infravermelho se apresente como uma alternativa na busca por manchas de fluidos. Trabalhos anteriores já provaram a capacidade da técnica MIR em discriminar diferentes fluidos corporais (ORPHANOU, 2015), porém a espectroscopia NIR ainda apresenta maior versatilidade, como já mencionado. Especificamente para aplicações utilizando imagens hiperespectrais, a espectroscopia NIR já apresenta sistemas de imagem capazes de varrer grandes áreas (EDELMAN; VAN LEEUWEN; AALDERS, 2012). Além disso, quando associada a métodos quimiométricos, ela é capaz de oferecer resultados confiáveis e objetivos, atendendo a esta demanda dos laboratórios periciais.

Geralmente, a grande quantidade de informação fornecida requer um tratamento especial dos dados que pode ser suprido por meio de técnicas de análise multivariada. Para fins de identificação e diferenciação, as abordagens de reconhecimento de padrões supervisionadas e não supervisionadas podem ser empregadas. Os métodos não supervisionados, como PCA e Análise de

Agrupamentos Hierárquicos (HCA: *Hierarchical Cluster Analysis*) possuem a vantagem de realizar análises de amostras sem nenhum conhecimento prévio podendo ser aplicadas como técnicas presuntivas. Em contrapartida, os métodos supervisionados consistem em técnicas de classificação propriamente ditas, que são capazes de prever a classe de uma determinada amostra (pixel) desconhecida após a validação do modelo, podendo ser empregados para abordagens confirmatórias.

Já técnicas de resolução de curvas como MCR-ALS possuem a flexibilidade de estimar a contribuição de um analito conhecido em uma mistura, mostrando-se adequadas para aplicações forenses, em que o analito de interesse é geralmente conhecido, mas pode estar presente em diferentes contextos químicos. Em recente publicação, a técnica CLS associada à HSI-NIR foi utilizada para visualizar sêmen, fluido vaginal e ureia em tecidos de algodão (ZAPATA; ORTEGA-OJEDA; GARCÍA-RUIZ, 2017). Uma limitação da metodologia proposta é que, quando técnicas como CLS são aplicadas, é de fundamental importância conhecer todos os componentes da mistura, isto é, não só o analista deve conhecer o espectro do sêmen, mas também de todos os compostos presentes na mistura, como substrato e contaminantes. Essa limitação pode ser evitada pelo uso de MCR-ALS. Adicionalmente, a natureza do substrato utilizado (algodão) apresenta uma alta influência no espectro dos compostos, o que faz com que esse problema em particular seja muito mais complexo e diferentes substratos devem ser considerados e avaliados.

Essa possibilidade da grande variedade de substratos disponíveis também pode ser abordada por PCA. A flexibilidade da PCA do ponto de vista forense reside no fato de que esse tipo de análise não exige nenhuma informação prévia sobre a composição química de nenhum composto presente na amostra; neste sentido, a PCA pode ser empregada como um método de triagem, buscando por diferentes características espectrais. Em contraste com essas técnicas, métodos supervisionados como a análise discriminante não apresentam a mesma flexibilidade, mas fornecem confiança estatística ao diferenciar o analito de outras substâncias, como falsos positivos.

4.2 OBJETIVOS

O objetivo do presente trabalho é avaliar o potencial da HSI-NIR juntamente com modelos multivariados como uma metodologia rápida, não destrutiva e confiável para (i) identificar manchas de fluidos de interesse forense em diferentes tecidos (abordagem presuntiva) e (ii) diferenciar as manchas de sêmen de outros compostos utilizando modelos de classificação (abordagem confirmatória).

- Avaliar o potencial das técnicas PCA e MCR-ALS para identificação de manchas em diferentes tecidos
- Avaliar o potencial classificatório dos modelos PLS-DA, sPLS-DA e SVM-DA para diferenciar sêmen de outros compostos.

4.3 METODOLOGIA

4.3.1 Amostras

Amostras de sêmen, lubrificante e outros fluidos biológicos (sêmen animal e leite materno) foram colocados sobre substratos (pedaços de tecido) criando uma mancha. Dois conjuntos de dados diferentes foram utilizados: (i) tecidos 100% algodão de diferentes cores (branco, vermelho, verde, amarelo e preto) e (ii) tecidos beges e brancos de diferentes composições (algodão, cetim e malha). É importante enfatizar que os tecidos coloridos de algodão possuem uma textura diferente quando comparados com os tecidos do conjunto brancos/beges. Todas as amostras de sêmen utilizadas foram doadas pelo Laboratório de Andrologia (ANDROLAB) da Universidade Federal Rural de Pernambuco. Cada amostra de sêmen foi obtida de diferentes doadores.

O projeto foi aprovado pelo Comitê de Ética da UFPE, por se tratar de uma pesquisa envolvendo manipulação de material biológico. O Certificado de Apresentação para Apreciação Ética do prjeto é 42279815.9.0000.5208 e o número do parecer 1.059.225.

4.3.2 Amostras sobre Tecidos Coloridos

Cinco tipos de tecido diferentes de algodão (branco, vermelho, verde, amarelo e preto) foram utilizados como substrato onde amostras de sêmen foram depositadas (humano, bode e cavalo), bem como substâncias conhecidamente comuns por serem

falsos positivos ou possivelmente encontradas em cenas de crime. Quatro marcas de lubrificantes (L1-L4), uma amostra de leite materno (BM1), quatro amostras de sêmen humano (S1-S4), uma de sêmen de cavalo (HS) e uma de sêmen de bode (GS) foram colocadas nos tecidos criando manchas. Uma mistura de lubrificante e sêmen também foi produzida. Para o conjunto de dados de Tecidos Coloridos, as amostras foram analisadas em 6 classes: lubrificantes, leite materno, sêmen humano, sêmen de cavalo, sêmen de bode e tecido puro.

4.3.3 Amostras sobre Tecidos Brancos/Beges

Cinco diferentes tecidos beges e brancos (algodão branco, algodão bege, cetim branco, malha branca e malha bege) foram utilizados como substratos para realizar o depósito dos fluidos. Quatro marcas de lubrificantes (L1-L4, as mesmas utilizadas para o conjunto anterior), oito amostras de sêmen humano (S1-S7 e SX, proveniente de um doador vasectomizado) e duas amostras de leite materno (BM1 e BM2) foram utilizadas para criar manchas. Três manchas foram criadas para cada lubrificante e amostras de leite materno em cada tecido. A quantidade de amostra de sêmen doada é diferente para cada doador e, por esse motivo, diferentes tamanhos e quantidades de manchas foram criadas por doador e por tecido.

Para o conjunto de dados Branco/Bege, os compostos foram divididos em 4 classes: lubrificante, sêmen humano, leite materno e tecido puro. Cada amostra de sêmen e leite materno foi proveniente de diferentes doadores.

Uma amostra de sêmen puro, de um doador diferente, numa placa de teflon foi utilizada para a aquisição de espectros, a ser utilizado como espectro de referência para os modelos MCR-ALS. Amostras dos dois conjuntos de dados foram deixadas para secar em condições ambiente por, pelo menos, 1 semana antes das análises.

4.3.4 Aquisição Espectral

Todas as Imagens foram adquiridas utilizando o modelo SWIR (Short Wave Infrared) do sistema de imagem SisuCHEMA da Specim (Oulu, Finlândia). O tamanho das imagens variou de acordo com cada mancha. A aquisição espectral das imagens foi realizada na faixa de 900-2500 nm com resolução espectral 6,3 nm e resolução espectral FWHM (Full Width Half Maximum) igual a 10 nm. As lentes na aquisição das imagens foram de 50 mm e o tamanho dos pixels de 156 μm^2.

4.3.5 Pré-processamento de dados

Antes de qualquer tratamento multivariado dos dados, técnicas de pré-processamento foram avaliadas para atenuar as variações não relacionadas com o analito. Foram avaliadas, sobretudo, técnicas como SNV, MSC e derivadas (1ª e 2ª ordem). Após o pré-processamento, diferentes técnicas foram empregadas com o objetivo de: (i) identificar as amostras de sêmen em tecido e (ii) identificar a natureza da mancha, diferenciando o sêmen de outros compostos.

4.3.6 Conjunto de Treinamento e Previsão

O conjunto de dados de Tecidos Brancos/Beges foram divididos em Conjunto de Treinamento e de Previsão para os modelos de classificação. Uma vez que os Tecidos Coloridos não possuíam a mesma variabilidade dos Tecidos Brancos/Bege, apenas o Conjunto de Treinamento foi utilizado nessas análises.

O Conjunto de Treinamento para os Tecidos Brancos/Beges foi construído com uma mancha de cada um dos doadores S1-S4, uma mancha de cada marca de lubrificante e uma mancha de cada doador de leite materno. O conjunto de Previsão foi composto pelas manchas remanescentes dos doadores S1-S4, BM1-BM2, lubrificantes L1-L4 e também as amostras dos doadores S5-S7 e SX, que não foram utilizadas na fase de construção do modelo. O esquema representativo da Figura 39 mostra como essa divisão foi realizada.

Uma Região de Interesse (ROI) das manchas do Conjunto de Treinamento foi selecionada de cada imagem, um quadrado de 61x61 pixels foi desdobrado e utilizado para a construção dos modelos. O conjunto de Previsão foi utilizado de duas formas para a avaliação dos modelos: (i) uma ROI do mesmo tamanho foi selecionada das imagens do Conjunto de Previsão, desdobradas e utilizadas para a obtenção dos valores de taxa de erro (ER), sensibilidade (Sn) e Especificidade (Sp); e também (ii) os modelos foram utilizados para a construção das imagens de previsão. Entretanto, apenas as imagens dos lubrificantes, leite materno e sêmen dos doadores S5-S7 e SX foram utilizadas nas imagens de previsão para simplificar os resultados.

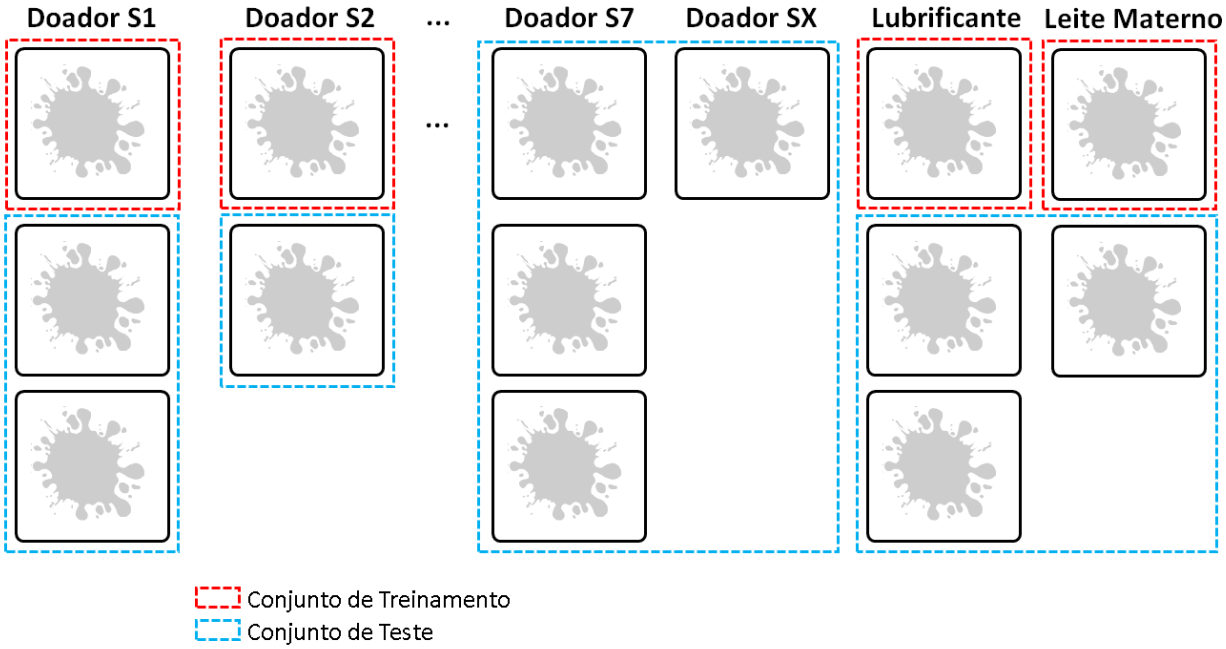

Figura 39 Esquema representativo da divisão das amostras do conjunto de tecidos Brancos/Beges nos conjuntos de Treinamento e Previsão.

Por motivos de desempenho computacional, o número de amostras de treinamento e previsão foi reduzido para a construção dos modelos sPLS-DA e SVM-DA utilizando uma compressão espacial na ROI utilizada anteriormente.

Todo o tratamento quimiométrico foi realizado utilizando o software Matlab, o PLS_Toolbox (Eigenvector Research Inc., EUA), Hypertools Toolbox gratuitamente disponível em (www.models.life.ku.dk), a interface MCR-ALS GUI 2.0 também gratuitamente disponível em (http://www.mcrals.info/). O algoritmo sPLS-DA foi utilizado como descrito na referência (CALVINI; ULRICI; AMIGO, 2015).

4.4 RESULTADOS E DISCUSSÃO

4.4.1 Características Espectrais e Pré-processamento

A etapa de pré-processamento provou-se uma etapa essencial da análise. Não só tecidos de algodão, mas também as malhas (misturas de poliéster) mostram grandes contribuições do tecido nos espectros das amostras. A textura dos tecidos também mostrou alta influência nos perfis espectrais. A Figura 40 mostra os espectros médios dos compostos sobre os diferentes tecidos.

Os espectros dos compostos sobre algodão branco obtidos mostram a grande influência da celulose (Figura 40), sendo possível identificar as bandas de absorção

mais importantes. O 1º e 2º sobretons do estiramento C-H em 1780 e 1220 nm, respectivamente; 1º sobretom do estiramento O-H e sua deformação em 1490 e 1940 nm, respectivamente; a importante absorção em 2110 nm originada do estiramento O-H, deformação e estiramento C-O e em 2276 nm absorção de estiramentos O-H, C-C e C-H e deformação C-H (ALI et al., 2001). Para os tecidos coloridos, é importante enfatizar a alta absorção do tecido preto na região de 900-1900 nm (Figura 40c), que pode estar relacionada com o pigmento utilizado na produção do tecido. Em geral, é possível visualizar que diferentes tecidos de algodão, exceto o preto, apresentam apenas pequenas diferenças entre si como variações de linha de base. Entretanto, devido à complexidade dos espectros NIR, é difícil atribuir as bandas para os espectros dos tecidos que contêm poliéster (malha e cetim), embora absorções características possam ser observadas e coerentes com as descritas na literatura (GHOSH; RODGERS, 2008).

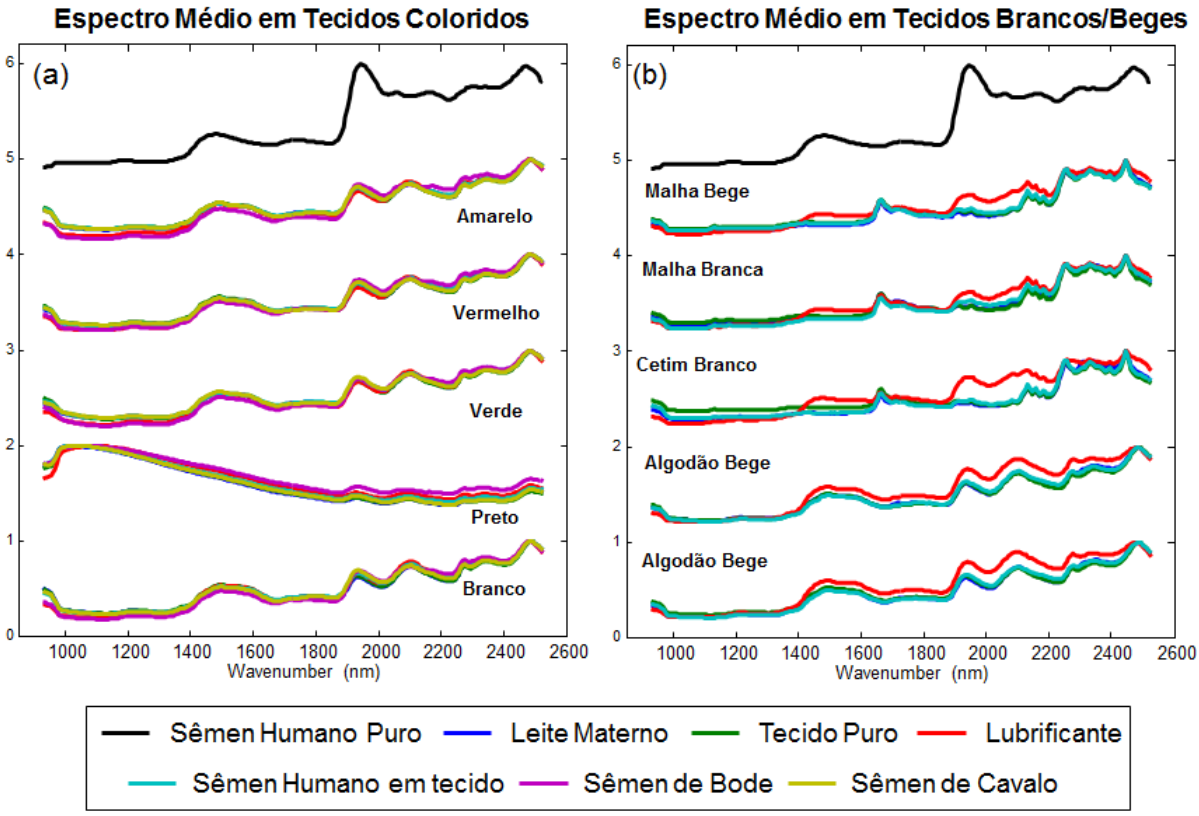

Figura 40 Espectro médio bruto NIR dos compostos sobre (a) tecidos Coloridos e (b) tecidos Brancos/Beges.

Sêmen de bode mostra uma banda de absorção entre 2170-2180 nm que, de acordo com a literatura, é uma importante região associada a proteínas que seguem

os modos vibracionais: (i) N-H 2º sobretom; (ii) estiramento C=O; (iii) N-H dobramento no plano; (iv) combinações de estiramentos C-N (WORKMAN; WEYER, 2007).

Os espectros dos tecidos sintéticos, cetim e malha, não exibem as absorções O-H que são características da celulose (tecido de algodão). Porém ainda apresentam grande influência nos espectros dos compostos, como será mostrado mais adiante.

Técnicas de pré-processamento como SNV e MSC foram avaliadas, porém apenas o uso de derivadas forneceu resultados adequados utilizando PCA como critério de análise, mostrando de forma mais evidente a diferença entre manchas e tecidos. Por essa razão, 1ª derivada com filtro de suavização Savitzky-Golay foi empregada. Para os tecidos Brancos/Beges, uma janela de 11 pontos foi utilizada e para os tecidos Coloridos, 15 pontos, exceto o tecido preto para o qual foi necessário empregar uma janela de 21 pontos para obter um espectro mais suavizado. Todas as suavizações foram realizadas utilizando um polinômio de 2ª ordem para o ajuste.

4.4.2 Identificando Manchas de Sêmen em Diferentes Tecidos

Após o pré-processamento dos dados, modelos de PCA das imagens foram construídos para identificar a presença de manchas de sêmen. A Figura 41 mostra a imagem dos escores e os respectivos gráficos dos pesos das PCs 1 e 3 para amostras de sêmen sobre os tecidos Coloridos. É possível perceber a influência da textura do tecido nas imagens dos escores que estão majoritariamente representadas pela 1ª componente. Isso significa que a maior fonte de variabilidade dos dados, cerca de 40 a 50%, está associada a efeitos físicos, que não foram corrigidos mesmo fazendo uso de técnicas de pré-processamento para eliminá-los. Da mesma forma que em PC1, a PC2 também apresentou apenas variabilidade associada à textura dos tecidos. De fato, a identificação das manchas só ocorre na 3ª PC e, mesmo assim, ainda fornece informações relacionadas à textura. De qualquer forma, ainda é possível identificar as manchas de sêmen, especialmente no tecido preto, que é de difícil identificação pelos métodos usuais.

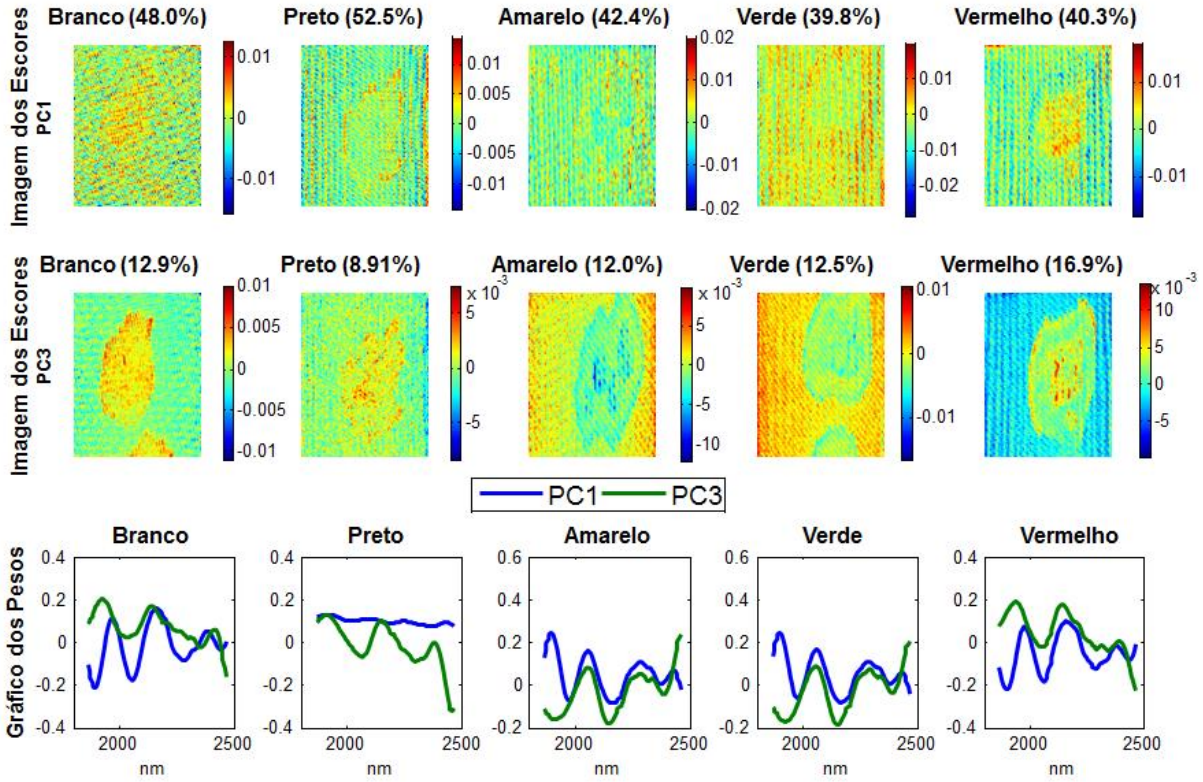

Figura 41 Imagem dos escores das PCs 1 e 3 para os tecidos Coloridos e gráfico dos pesos para as respectivas PCs (abaixo).

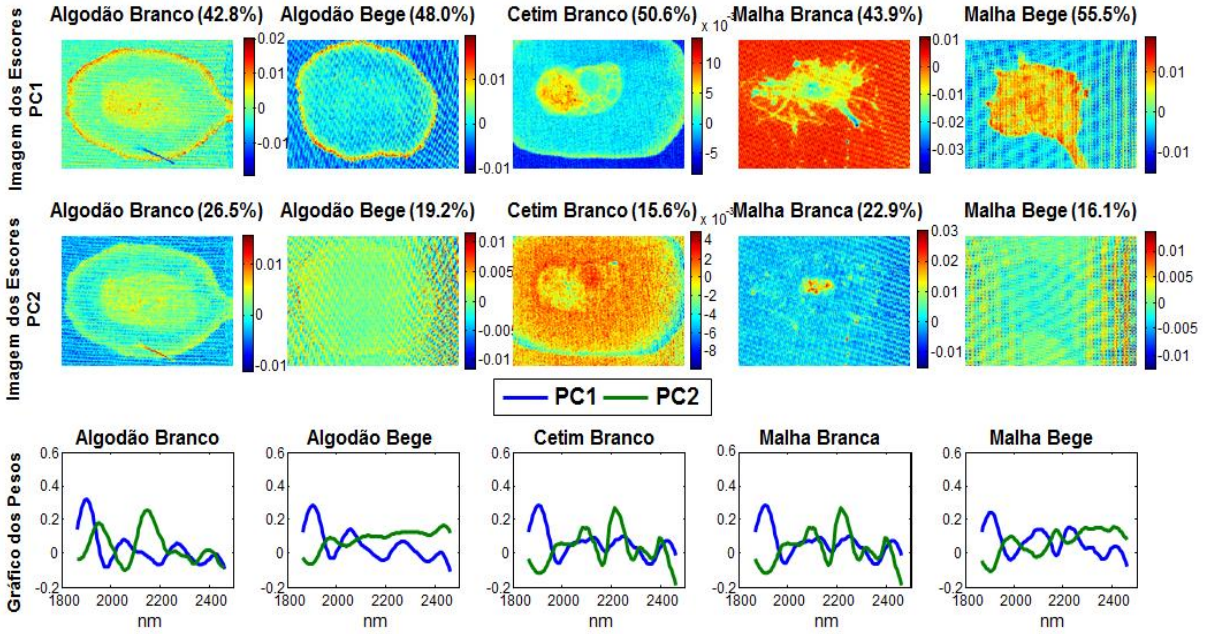

Figura 42 Imagem dos escores das PCs 1 e 2 para os tecidos Brancos/Beges e gráfico dos pesos para as respectivas PCs (abaixo).

Para o conjunto de dados Brancos/Beges (Figura 42) as manchas de sêmen podem ser facilmente identificadas nas imagens dos escores da PC1. Os tecidos do conjunto Brancos/Beges apresentam uma textura mais homogênea quando comparados com o conjunto dos tecidos coloridos e, por isso, a variabilidade associada a esta característica não é tão evidente. Portanto, a maior variabilidade dos dados está associada às diferenças entre o tecido e a mancha de sêmen, dessa forma fica claro que o número de PCs para observar as manchas é significativamente menor em tecidos mais lisos.

A técnica de MCR-ALS também foi empregada nas imagens para os diferentes tecidos. Um modelo de dois componentes foi construído utilizando, como estimativas iniciais, um espectro puro de sêmen e um do tecido utilizado. Espectros de referência do sêmen e do tecido foram utilizados para aumentar a matriz e restrições de igualdade e não negatividade nas concentrações foram impostas. Os mapas de concentração foram obtidos para as manchas em cada tecido e os espectros otimizados (Figura 43, a Figura 45) foram posteriormente comparados com as estimativas iniciais.

Os mapas de distribuição da Figura 43 e da Figura 44 mostram claramente uma diferença de concentração do sêmen ao longo das manchas, mais claramente observado nos tecidos Brancos/Beges (Figura 44). Isso pode ser observado pelos modelos porque as amostras de sêmen possuem uma composição complexa e a distribuição dos seus compostos ao longo do tecido pode se dar de forma bastante heterogênea ao longo do tempo. Além disso, as características do próprio tecido também influenciam no processo de dispersão.

É importante notar que as amostras de sêmen sobre os tecidos Coloridos (Figura 43) são mais difíceis de identificar devido às características de textura desses tecidos. Ainda assim, elas apresentam uma melhor visualização das manchas quando comparadas com as imagens dos escores dos modelos PCA.

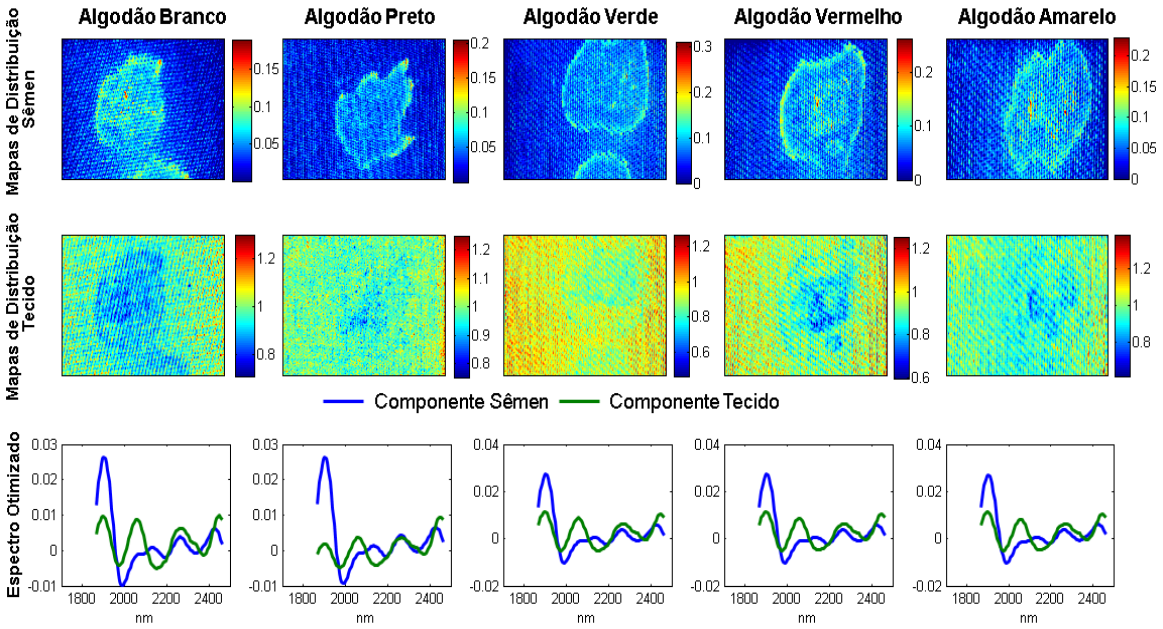

Figura 43 Mapas de distribuição e espectros otimizados para amostras de sêmen obtido com o modelo MCR-ALS para os Tecidos Coloridos.

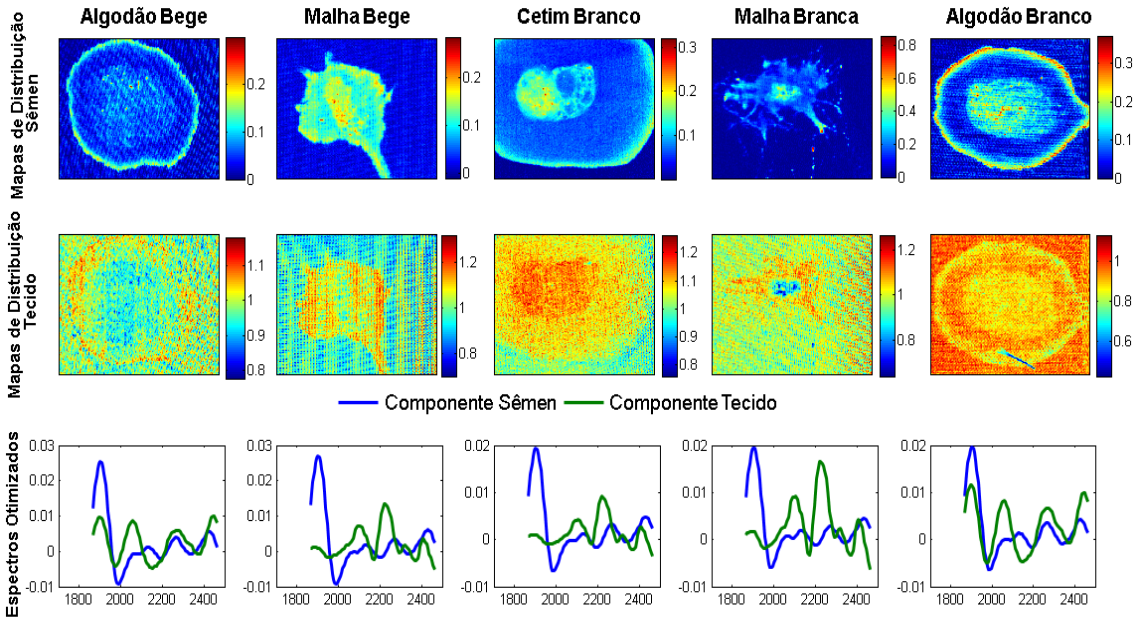

Figura 44 Mapas de distribuição e espectros otimizados para amostras de sêmen obtido com o modelo MCR-ALS modelo MCR-ALS para os Tecidos Brancos/Beges.

O componente de sêmen do espectro otimizado obtido a partir dos modelos MCR-ALS, para todos os tecidos, mostrou alta correlação com o espectro puro do sêmen com valores entre 0,993 e 0,999 (Figura 45).

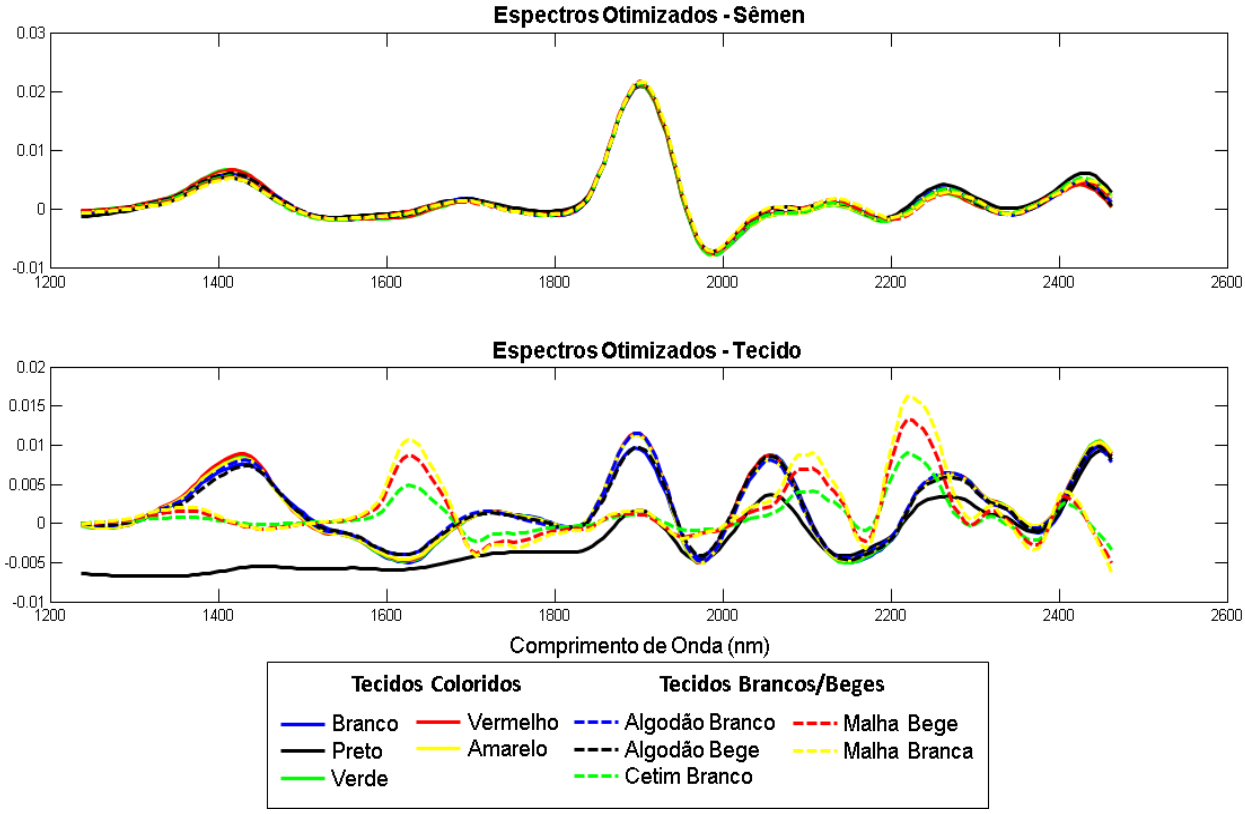

Figura 45 Espectros otimizados dos componentes sêmen (cima) e tecido (baixo) obtidos a partir das imagens de manchas de sêmen.

Quando o mesmo modelo é aplicado para imagens cujas manchas não são de sêmen do conjunto de Tecidos Brancos/Beges, os coeficientes de correlação entre os espectros otimizados dos compostos (que não são sêmen) e o espectro puro do sêmen é, em geral menor (em torno de 0.66-0.99); exemplo do tecido de algodão branco na Figura 46, porém alguns componentes (L2, L4, BM1 e BM2) ainda apresentam alta correlação. Embora altos valores de correlação tenham sido obtidos, a maioria dos mapas de distribuição associados ao componente sêmen não representa a mancha, mas o tecido.

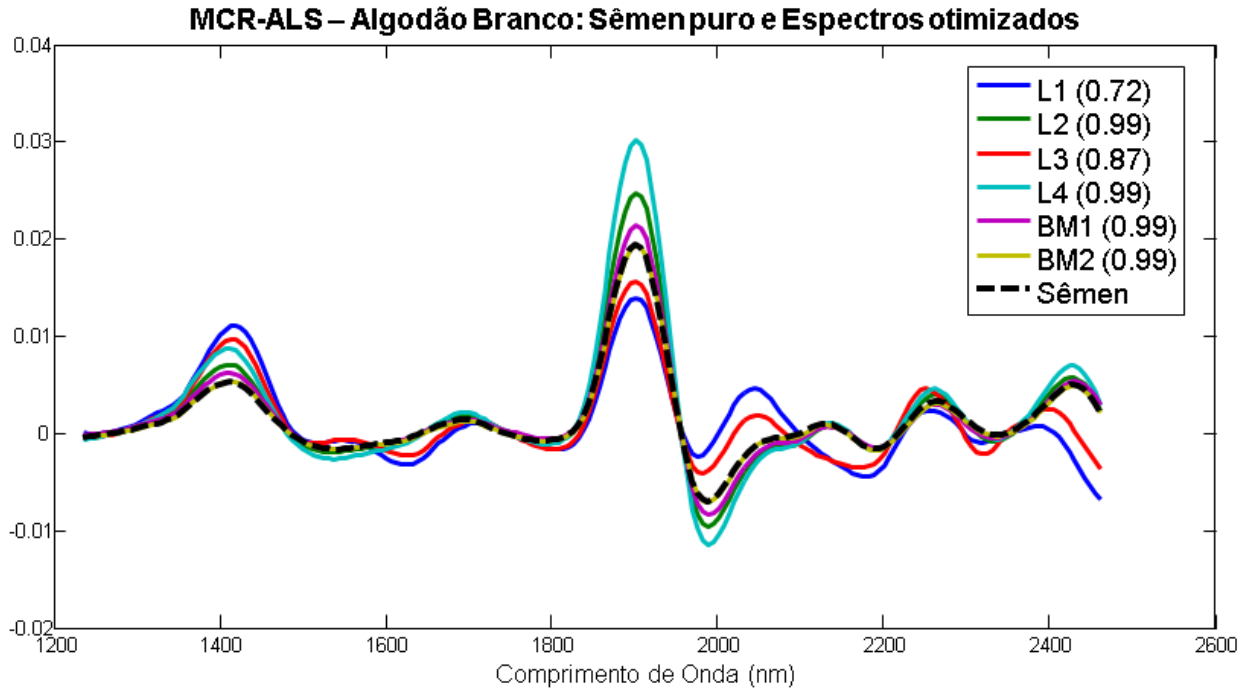

Figura 46 Espectros de sêmen puro e espectros otimizados do componente sêmen obtidos do modelo MCR-ALS para as imagens das manchas de outros compostos. Tecido de algodão Branco do conjunto de Tecidos Brancos/Beges.

No caso do lubrificante L2 (Figura 47), o mapa de distribuição e o espectro otimizado se mostraram consistentes com as amostras de sêmen, apresentando uma alta correlação com o espectro puro do sêmen e um mapa de distribuição consistente com a mancha. Já o caso do lubrificante L1 mostra que o espectro otimizado apresenta uma correlação menor com o espectro do sêmen puro (0.72). Para os tecidos coloridos, todos os componentes de sêmen obtidos a partir da otimização do modelo MCR-ALS com manchas de outros compostos mostraram alta correlação com o sêmen, mostrando que técnicas de classificação ainda são necessárias para avaliar o potencial das HSI-NIR em diferenciar os diferentes compostos estudados.

Também é possível observar que os componentes do tecido para os diferentes tecidos (Figura 45) mostram consistência entre si; os tecidos de cetim, malha branca e malha bege possuem poliéster em sua composição, enquanto os demais são compostos somente por algodão. O tecido preto é o que apresenta o espectro mais diferente, especialmente na região entre 100-1800 nm.

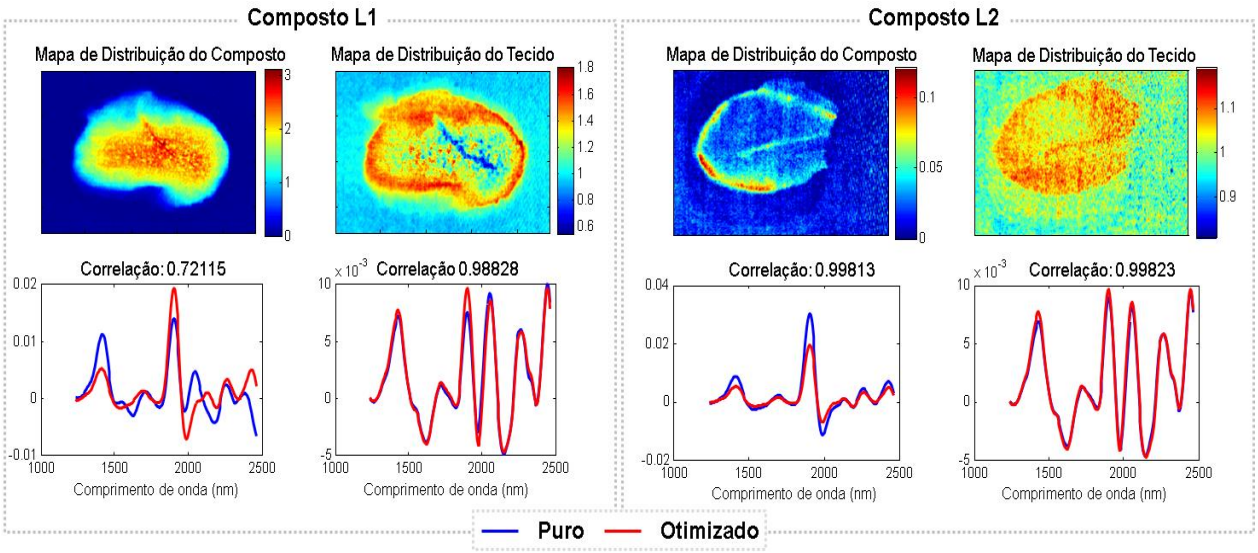

Figura 47 Modelo MCR-ALS aplicado à Imagem da mancha dos lubrificantes L1 e L2. Mapas de distribuição acima e espectros otimizados abaixo. Tecido de Algodão Branco do conjunto de Tecidos Brancos/Beges.

Comparando os resultados obtidos por PCA e MCR-ALS é possível atestar vantagens e desvantagens de cada metodologia. Enquanto a PCA possui a grande vantagem de ser uma técnica não supervisionada e, portanto, não necessita de nenhuma referência inicial para a construção dos modelos, MCR-ALS possui a vantagem de fornecer resultados mais fáceis de interpretar. De fato, MCR-ALS apresentou modelos mais bem-sucedidos na identificação de manchas nos diferentes tecidos, pois a variabilidade relacionada com a textura não interfere, aparentemente, nos resultados dos modelos MCR-ALS, como nos de PCA.

Na verdade, para aplicações forenses, um dos maiores desafios é a identificação de um composto conhecido em diferentes contextos de cenas de crime; e o caso dos resultados de MCR-ALS se mostraram bem-sucedidos, pois foram capazes de resolver os sinais dos conjuntos de dados mesmo que os espectros dos compostos se apresentassem sobrepostos pela alta absorção dos tecidos, especialmente os tecidos de algodão. Entretanto, a metodologia proposta não deve ser empregada como uma técnica confirmatória, pois não é capaz de diferenciar sêmen de outros compostos.

4.4.3 Diferenciar, Classificar e Discriminar Sêmen de Outros Compostos

Os espectros dos tecidos coloridos apresentaram espectros muito similares e, como mencionado anteriormente, mostraram apenas pequenas mudanças na linha de

base. Portanto, um modelo foi construído utilizando os dados em tecido branco e os remanescentes coloridos como conjunto de previsão. Para esse modelo inicial, a derivada de Savitzky-Golay foi empregada como descrito anteriormente para minimizar essa variabilidade entre os tecidos de cores diferentes. Porém, os modelos não apresentaram resultados satisfatórios para Sn e Sp e, portanto, os modelos foram construídos separadamente para cada tecido.

PLS-DA

Modelos PLS-DA foram construídos para os dois conjuntos de dados (Tecidos Coloridos e Brancos/Beges), como mencionado na seção de metodologia do presente capítulo. Embora modelos para as classes de cada composto tenham sido construídos e mostrados nos Apêndices (A e B), a classe sêmen é a mais crítica e de maior importância para esta abordagem e, consequentemente, apenas as figuras de mérito relativas aos modelos para a classe de sêmen foram levadas em consideração, como mostradas na Tabela 4.

É possível perceber que os valores de Sn e Sp estão, majoritariamente, entre 0,8 e 0,9, com baixas taxas de erro (ER). Para os tecidos Coloridos, o mesmo resultado é observado, com altos valores de sensibilidade e especificidade e baixas taxas de erros. Os modelos para tecido preto e malha bege forneceram valores de especificidade igual a 0,65 e 0,77, respectivamente, ocasionados pela presença de falsos positivos, isto é, espectros de outros compostos que são classificados como sêmen. As curvas ROC da Figura 48 mostram o comportamento dos modelos de classificação para sêmen em quatro situações: (i) algodão branco; (ii) malha bege; (iii) cetim branco; e (iv) algodão preto.

Tabela 4 Resumo dos resultados dos modelos PLS-DA para a classe de sêmen.

PLS-DA										
	Brancos/Beges (CV)			**Brancos/Beges (pred)**				**Coloridos (CV)**		
	Sn	**Sp**	**ER**	**Sn**	**Sp**	**ER**		**Sn**	**Sp**	**ER**
Algodão Branco	0.98	0.99	0.02	0.92	0.99	0.05	Branco	0.96	0.96	0.04
Algodão Bege	0.97	0.94	0.04	0.95	0.92	0.06	Preto	0.96	**0.65**	0.2
Cetim Branco	0.92	0.96	0.06	0.89	0.94	0.09	Verde	0.95	0.89	0.08
Malha Branca	0.85	0.89	0.1	0.86	0.95	0.1	Vermelho	0.98	0.96	0.03
Malha Bege	0.98	0.92	0.05	0.83	**0.77**	0.2	Amarelo	0.95	0.89	0.08

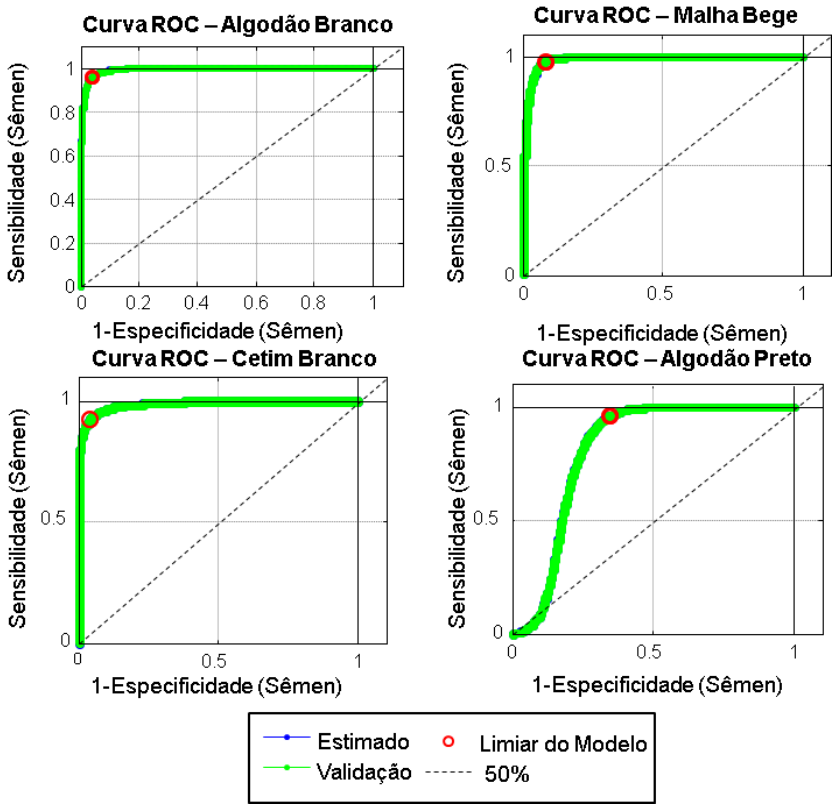

Figura 48 Curvas ROC para a classe de sêmen nos tecidos: (i) algodão branco; (ii) malha bege; (iii) cetim branco; e (iv) algodão preto.

As imagens de previsão para os modelos PLS-DA para os tecidos coloridos podem ser observadas na Figura 49. É possível notar que a previsão para a Malha Bege, que mostrou um baixo valor de Sp para o sêmen, apresenta muitos pixels em vermelho relacionados com tecido que foram classificados como sêmen (Figura 49e). Além disso, é curioso notar que, embora a malha bege não tenha apresentado valores muito destoantes de Sn (0,83), uma mancha de sêmen de um mesmo doador (S6) não foi classificada como sêmen, apresentando um caso de falso negativo, e refletido numa queda do valor de Sn quando comparado com os outros tecidos. Esse caso representa um grande problema no presente contexto, pois um caso de falso negativo representa uma amostra de sêmen que deixaria de ser coletada.

As imagens de previsão das manchas em algodão branco e bege mostram dois erros característicos de classificação: (i) pixels classificados como sêmen nas bordas dos lubrificantes L1 e L3; e (ii) pixels classificados como leite materno na mancha do lubrificante L1, Figura 49b e Figura 49c. Pixels do lubrificante L1 classificados como leite materno também podem ser visualizados na imagem de previsão para a Malha Bege (Figura 49e). No caso das amostras de algodão branco e bege, as classificações

equivocadas podem estar associadas com a forma como as manchas de lubrificante e leite materno secam e se difundem ao longo do tecido. Possivelmente, substâncias dos compostos que se depositam nas bordas das manchas de leite se assemelham com substâncias presentes no sêmen.

No caso das imagens de previsão para os tecidos Coloridos (Figura 50), o desempenho da classificação é pior, revelando a grande influência da textura na construção desses modelos. Não só isso, mas particularmente a classificação do tecido preto (Figura 50c) não fornece uma imagem das machas de forma bem definida, em compensação não apresenta nenhum caso de falso negativo, o que é de extrema importância, como visto anteriormente. Para eliminar o ruído e a variabilidade associada à textura dos tecidos, modelos esparsos foram avaliados com o objetivo de melhorar as previsões e identificar os canais espectrais que melhor discriminam as diferentes classes.

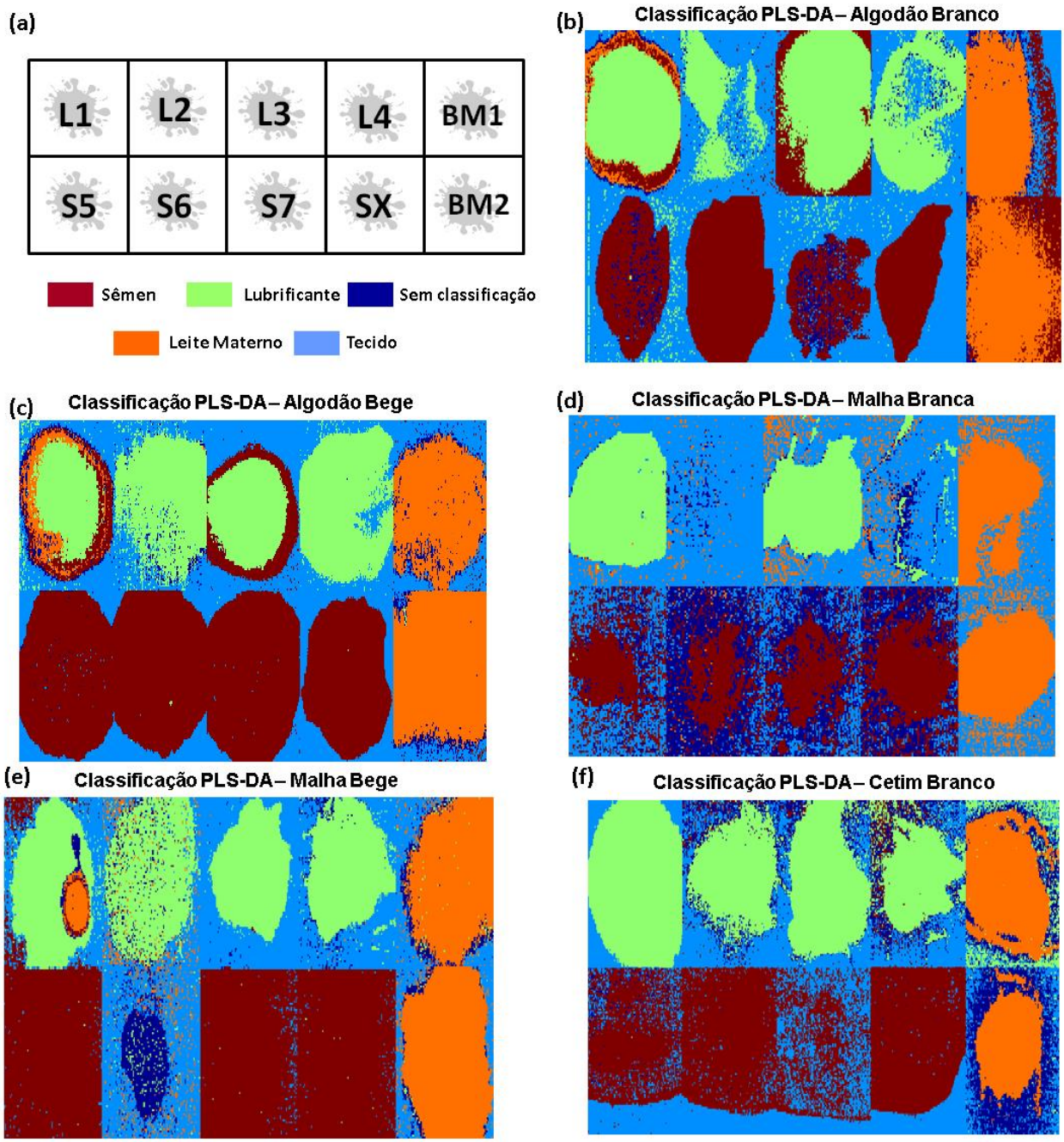

Figura 49 (a) Esquema de manchas para os tecidos Beges/Brancos e legenda; Imagens de previsão dos modelos PLS-DA para os tecidos (b) algodão branco, (c) algodão bege, (d) malha branca, (e) malha bege e (f) cetim branco.

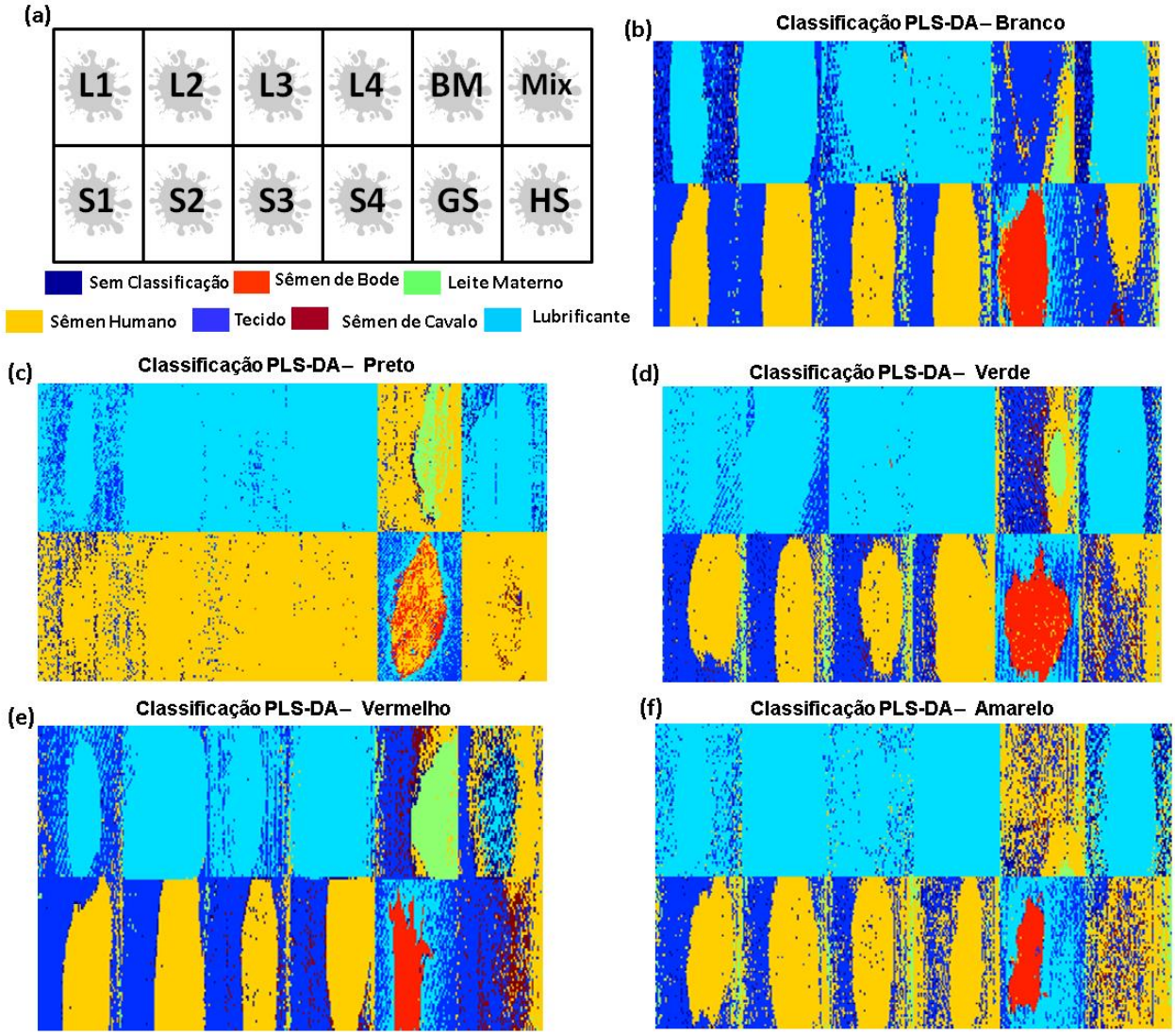

Figura 50 (a) Esquema de manchas para os tecidos Coloridos e legenda; Imagens de previsão dos modelos PLS-DA para os tecidos de algodão (b) branco, (c) preto, (d) verde, (e) vermelho e (f) amarelo.

sPLS-DA

Modelos sPLS-DA foram construídos para identificar o número ótimo de sLV e o número de variáveis originais a serem incluídas para a construção de cada sLV; foi realizada uma combinação de 3 até 10 sLV incluindo, para cada uma, de 5 até 100 variáveis originais. Com base na eficiência dos modelos (EFF), mostradas na Figura 51, uma combinação diferente desses dois parâmetros foi selecionada para cada classe. Os resultados para todos os modelos sPLS-DA podem ser encontrados nos Apêndices C e D, porém apenas os modelos para as classes de sêmen foram levados

em consideração para a avaliação da metodologia proposta. Os resultados para os modelos da classe sêmen podem ser encontrados na Tabela 5.

Tabela 5 Resumo dos resultados dos modelos sPLS-DA para a classe de sêmen.

sPLS-DA										
	Brancos/Beges (CV)			Brancos/Beges (pred)				Coloridos (CV)		
	Sn	Sp	ER	Sn	Sp	ER		Sn	Sp	ER
Algodão Branco	0.99	0.99	0.01	0.98	0.91	0.05	Branco	0.96	0.96	0.04
Algodão Bege	0.96	0.97	0.04	0.94	0.94	0.06	Preto	**0.77**	0.88	0.2
Cetim Branco	0.96	0.96	0.04	**0.59**	0.96	0.3	Verde	0.91	0.95	0.08
Malha Branca	0.89	0.89	0.1	0.88	0.95	0.1	Vermelho	0.97	0.98	0.03
Malha Bege	0.97	0.97	0.03	0.92	**0.68**	0.2	Amarelo	0.92	0.94	0.08

Os gráficos de Eficiência da Figura 51 mostram como a estabilidade dos modelos varia na medida em que mais sLV vão sendo adicionadas e quantas variáveis originais são incorporadas. É possível perceber uma determinada estabilidade do modelo, para o tecido de Algodão Branco, a partir de 3 sLV, em que há ainda um aumento de eficiência quando 8 sLV são utilizadas. O número de variáveis originais incluídas apresenta estabilidade, aumentando a eficiência do modelo em torno de 60 variáveis. Por isso, a escolha desses parâmetros para modelar a classe sêmen no tecido Algodão Branco (como pode ser visto no Apêndice C). Analogamente ao tecido Algodão Branco, é possível realizar a seleção dos parâmetros para os demais tecidos. Em particular, o tecido Algodão Preto apresenta grande instabilidade do modelo, mostrando os melhores resultados com 9 sLV e 71 variáveis originais incluídas.

Novamente, os valores para Sn e Sp são aproximadamente 0.9, com exceção dos tecidos Malha Bege e Algodão Preto, que apresentam valores iguais a 0,68 e 0,77 para especificidade, respectivamente. O tecido Cetim Branco também apresenta um decréscimo em seu valor de sensibilidade, igual a 0.59.

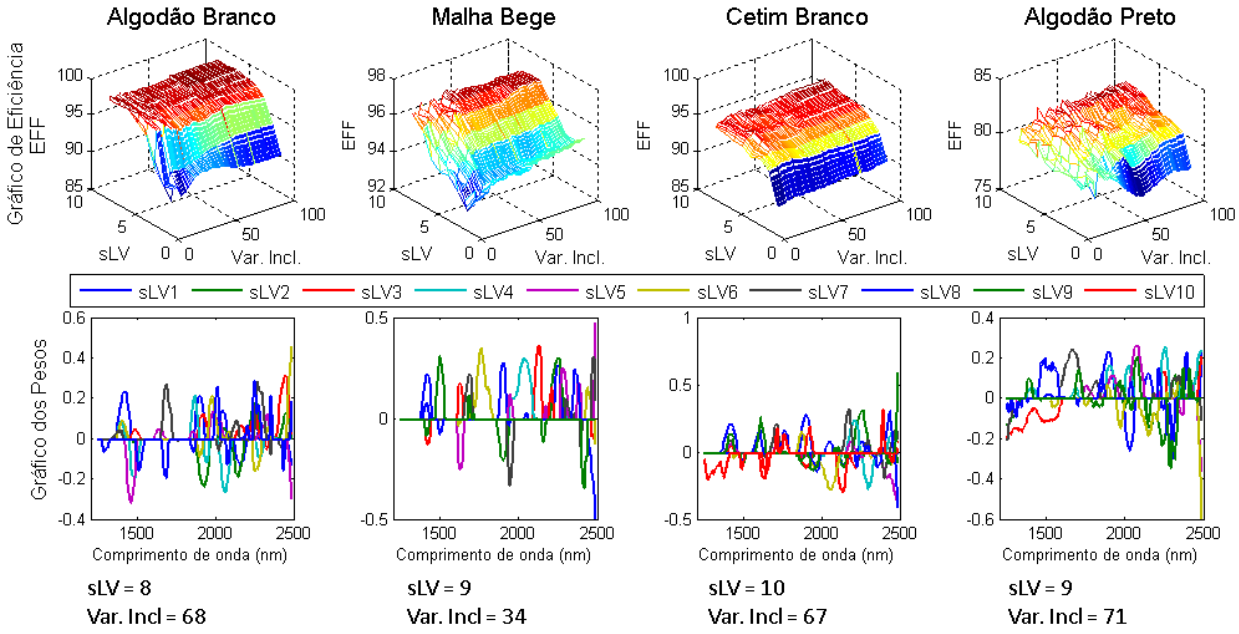

Figura 51 Gráfico de Eficiência para os modelos sPLS-DA da classe sêmen (acima) e os gráficos dos pesos (abaixo) para os tecidos de algodão branco, malha bege, cetim branco e algodão preto.

A Figura 52 e a Figura 53 mostram as imagens de previsão para os modelos construídos para os conjuntos Brancos/Beges e Coloridos, respectivamente. É possível observar, a partir das imagens que, em geral, os modelos sPLS-DA perdem um pouco em especificidade quando comparados com os modelos PLS-DA. Por exemplo, muitos pixels que pertencem à classe tecido no Cetim Branco são classificados como sêmen e no tecido Malha Bege, pixels pertencentes à classe de lubrificante foram classificados como sêmen (Figura 52e e Figura 52f).

Já na Figura 53 é possível observar que há uma manutenção da sensibilidade dos modelos, mesmo no tecido preto, os pixels relativos às manchas de sêmen são geralmente classificados corretamente. Em termos de especificidade, assim como para os tecidos Brancos/Beges, há uma piora, representada pela presença de pixels, principalmente de leite materno e sêmen de cavalo (BM e HS) classificados como sêmen humano.

Embora a imposição da esparsidade resulte em perda de especificidade, é importante notar nas imagens que há uma melhora na sensibilidade dos modelos. Por exemplo, a mancha de sêmen do doador S6, que não é identificada no modelo PLS-DA, pode ser identificada como uma mistura de sêmen e lubrificante quando modelo

sPLS-DA é empregado. Neste sentido, os modelos sPLS-DA mostram uma grande vantagem frente aos modelos PLS-DA, pois não geram falsos negativos para a classe sêmen.

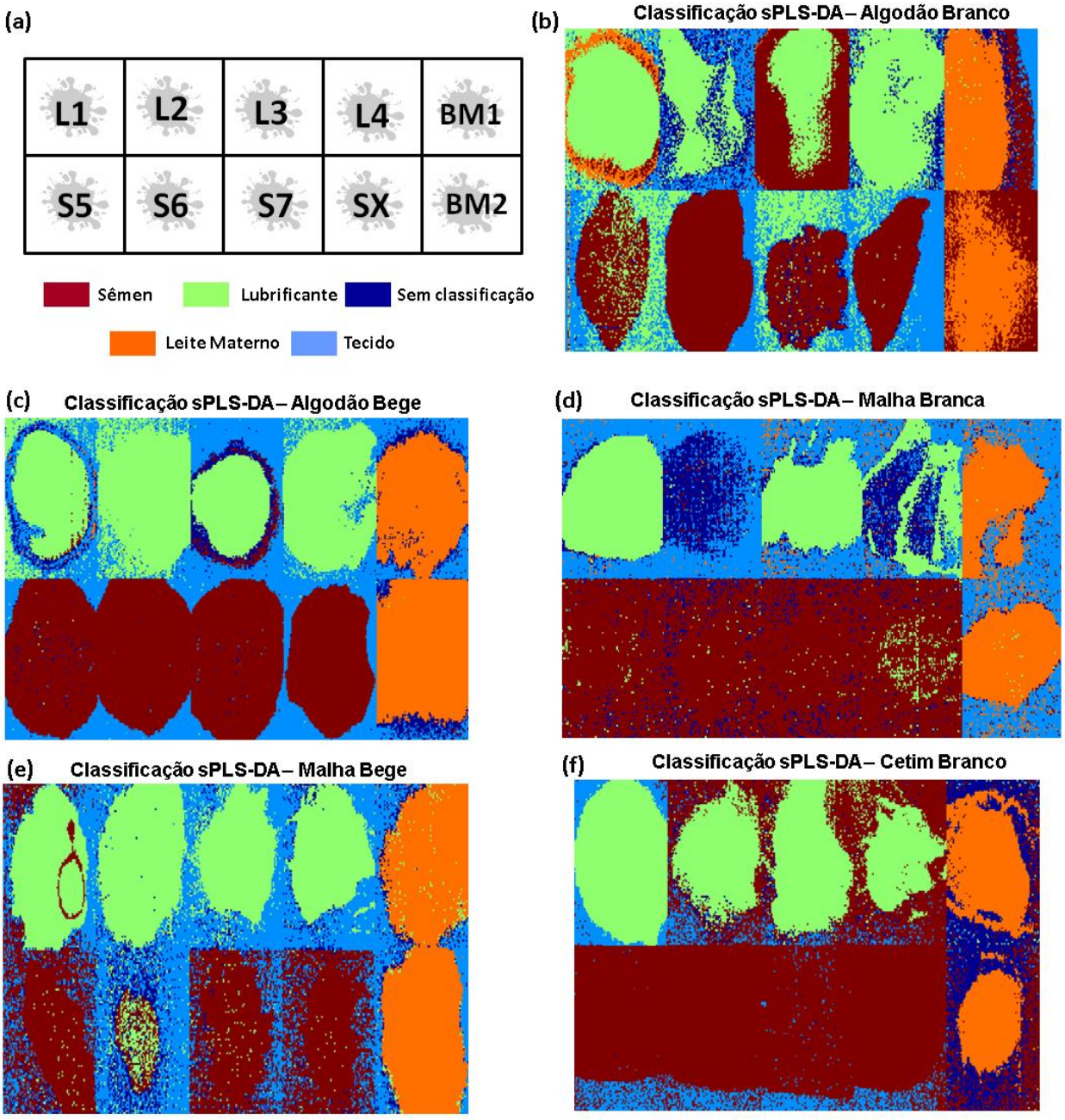

Figura 52 (a) Esquema de manchas para os tecidos Beges/Brancos e legenda; Imagens de previsão dos modelos sPLS-DA para os tecidos (b) algodão branco, (c) algodão bege, (d) malha branca, (e) malha bege e (f) cetim branco.

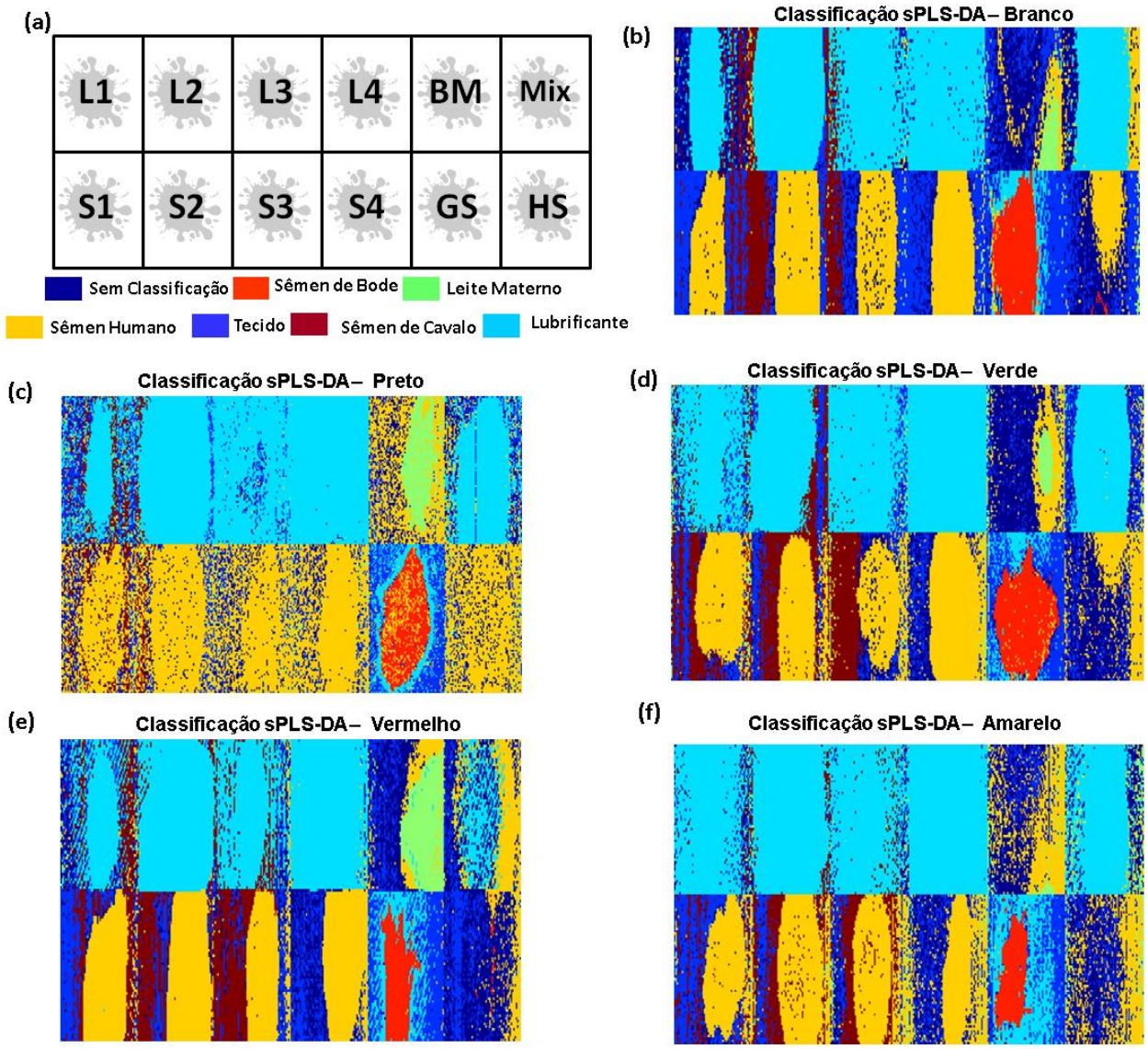

Figura 53 (a) Esquema de manchas para os tecidos Coloridos e legenda; Imagens de previsão dos modelos sPLS-DA para os tecidos de algodão (b) branco, (c) preto, (d) verde, (e) vermelho e (f) amarelo.

Os escores dos modelos sPLS-DA e PLS-DA mostram uma grande sobreposição de classes, ocasionando uma significativa confusão na previsão dos resultados. Isso se dá, provavelmente, pela já mencionada alta influência do substrato nos espectros dos componentes. Por essa razão, uma abordagem não linear foi avaliada com objetivo de melhorar os modelos.

SVM-DA

A técnica SVM para classificação foi empregada utilizando uma compressão por PLS e o algoritmo para Função de Base Radial. Dois parâmetros são

automaticamente otimizados e estão relacionados com as bordas de cada classe e um termo de penalidade para erros de classificação das amostras de validação. Os resultados para todos os modelos SVM-DA podem ser encontrados nos Apêndices E e F; e os resultados para as classes de sêmen estão expostas na Tabela 6.

Tabela 6 Resumo dos resultados dos modelos SVM-DA para a classe de sêmen.

SVM-DA										
	Brancos/Beges (CV)			Brancos/Beges (pred)				Coloridos (CV)		
	Sn	Sp	ER	Sn	Sp	ER		Sn	Sp	ER
Algodão Branco	0.97	1.0	0.01	0.95	0.99	0.03	Branco	0.97	0.99	0.02
Algodão Bege	0.93	0.95	0.06	0.84	0.96	0.1	Preto	0.88	0.93	0.09
Cetim Branco	0.97	1.0	0.02	0.91	1.00	0.08	Verde	0.95	0.98	0.03
Malha Branca	0.90	0.97	0.06	0.90	0.96	0.07	Vermelho	0.97	0.99	0.02
Malha Bege	0.93	0.98	0.05	0.81	0.94	0.1	Amarelo	0.94	0.98	0.04

Na verdade, os valores para Sn e Sp para as três abordagens não parecem diferir muito entre si. Porém, para os modelos SVM-DA todos os valores de Sn e Sp são maiores que 0,90, havendo uma queda apenas no valor de Sn para Malha Bege, Algodão bege e tecido Preto. Além disso, as imagens de previsão dos modelos SVM-DA apresentam um melhor desempenho geral quando comparados com os modelos PLS-DA e sPLS-DA (Figura 54 e Figura 55).

As imagens de previsão dos modelos SVM-DA mostram mais claramente as manchas em todos os casos. De fato, em termos gerais, os modelos mostram valores maiores de Sp, fazendo com que a presença de falsos positivos seja menor quando comparada com as outras abordagens quimiométricas. Entretanto, a mancha de sêmen do doador S6 no tecido Malha Bege não é observada (Figura 54e), o que leva novamente ao caso de falso negativo para a classe de sêmen, que não é permitido para a presente aplicação.

Para o tecido algodão Preto do conjunto de Tecidos Coloridos (Figura 55c) fica clara a diferença do desempenho da abordagem não linear. De fato, SVM-DA é a única abordagem capaz de diferenciar sêmen de cavalo de sêmen humano não só no tecido Preto, mas em todos os coloridos. Porém, para o presente contexto, o resultado mais importante é que nenhuma amostra de sêmen humano seja classificada em outra classe (ou não seja classificada). Isto é, os valores de Sn para a classe sêmen são

mais importantes do que os valores de Sp para a classe sêmen ou Sn e Sp para qualquer outra classe.

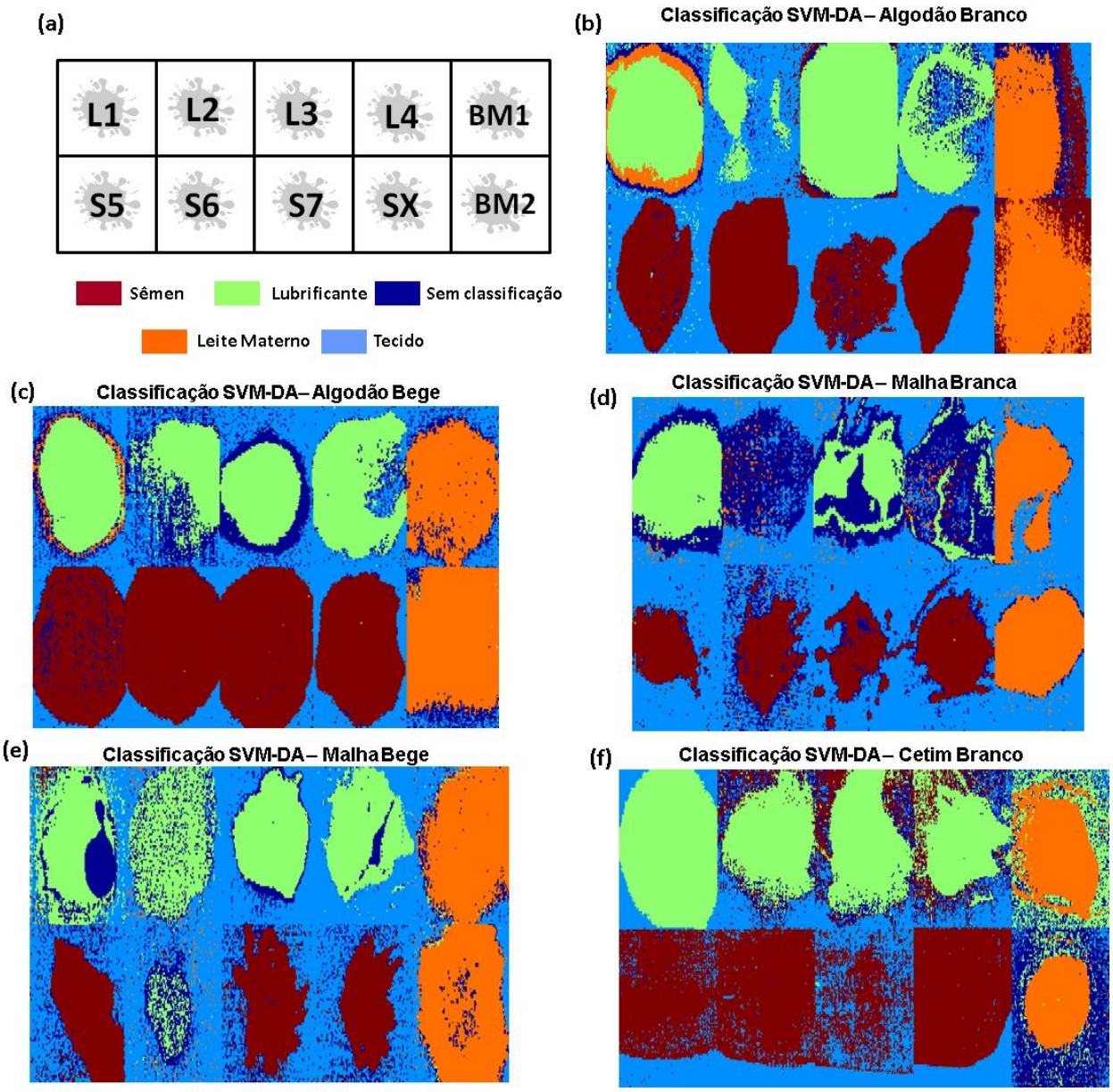

Figura 54 (a) Esquema de manchas para os tecidos Beges/Brancos e legenda; Imagens de previsão dos modelos SVM-DA para os tecidos (b) algodão branco, (c) algodão bege, (d) malha branca, (e) malha bege e (f) cetim branco.

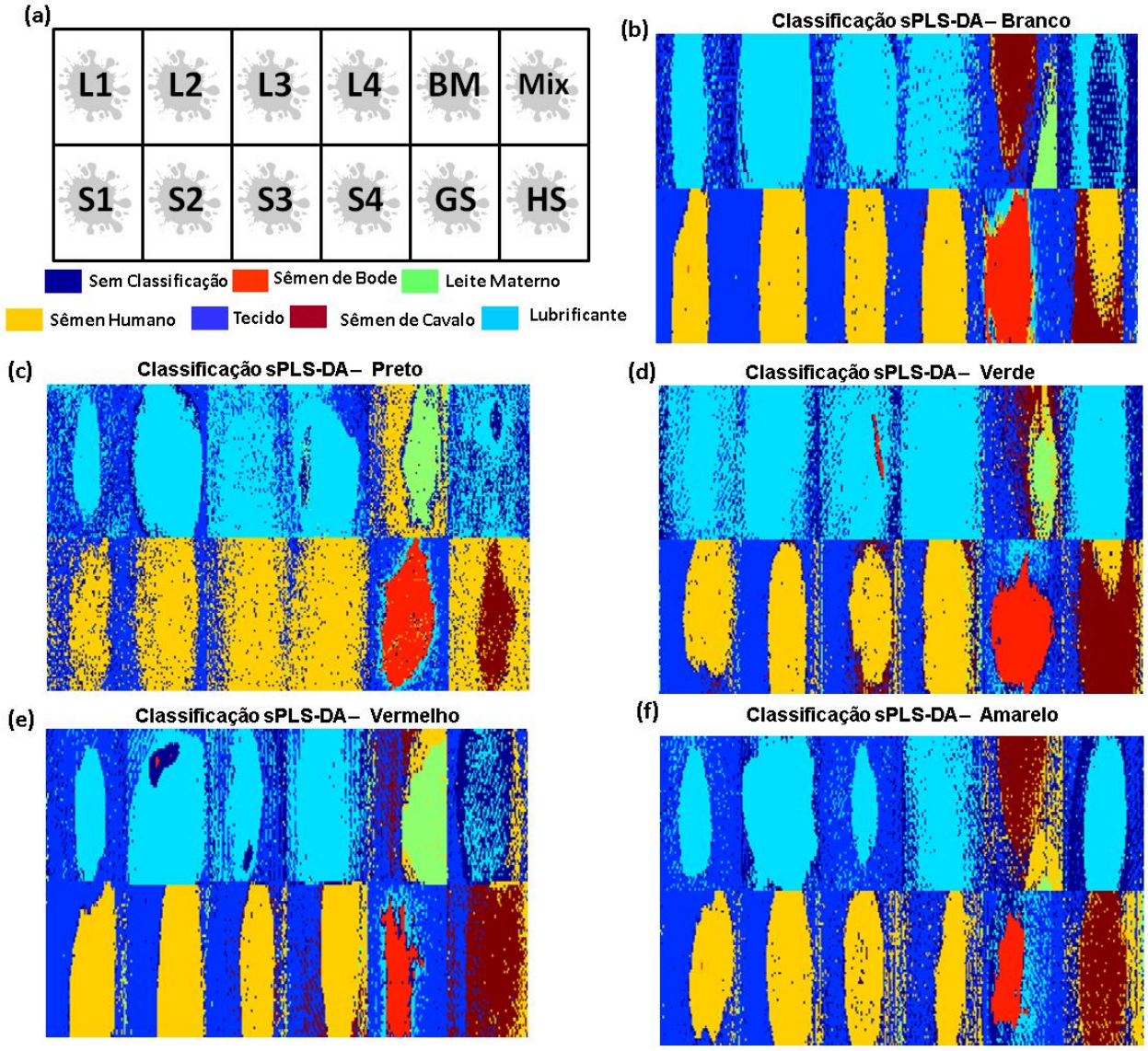

Figura 55 (a) Esquema de manchas para os tecidos Coloridos e legenda; Imagens de previsão dos modelos SVM-DA para os tecidos de algodão (b) branco, (c) preto, (d) verde, (e) vermelho e (f) amarelo.

Na detecção de fluidos corporais, um dos maiores problemas é a identificação de manchas em tecidos escuros, como o preto. Sob iluminação de luz visível ou das fontes de luz alternadas as manchas de sêmen são difíceis de ser identificadas. A abordagem utilizando SVM-DA foi capaz de revelar essas manchas e ainda diferenciar sêmen humano de sêmen animal e dos outros compostos utilizados. Para as imagens de previsão em tecido Preto, é importante notar que, apesar da baixa especificidade, nenhum caso de falso negativo foi observado. sPLS-DA não foi capaz de diferenciar sêmen humano do sêmen de cavalo, em contrapartida foi a única técnica que não mostrou falso negativo para a classe sêmen humano e, portanto pode ser considerada

a abordagem confirmatória mais apropriada, considerando todo o conjunto de tecidos estudados.

4.5 CONCLUSÃO

No presente trabalho, uma metodologia rápida e não destrutiva para a identificação e diferenciação de manchas de sêmen foi proposta combinando HSI-NIR e diferentes métodos de análise multivariada. As técnicas de PCA e MCR-ALS foram empregadas com o objetivo de realizar a identificação de manchas de fluidos sobre tecidos de diferentes cores e composições. A textura dos tecidos mostrou uma alta influência nos resultados, particularmente para PCA, fazendo com que essa técnica seja mais adequada para a busca de manchas em substratos mais lisos e uniformes. Em contrapartida, MCR-ALS foi capaz de identificar a presença de manchas em todos os tecidos e o componente de sêmen otimizado a partir das imagens de manchas de sêmen apresentarou uma alta correlação com o espectro puro do sêmen, fazendo do MCR-ALS uma técnica apropriada como abordagem presuntiva, mas não confirmatória. Não obstante, é importante avaliar metodologias *screening* para uma análise não destrutiva e MCR-ALS mostrou potencial para realizar esse tipo de análise independente do tecido utilizado como substrato.

Para análises confirmatórias, as técnicas de classificação PLS-DA, sPLS-DA e SVM-DA foram posteriormente avaliadas. A técnica SVM-DA mostrou um melhor desempenho no geral, fornecendo modelos mais específicos para todas as classes. Entretanto, os modelos sPLS-DA foram os únicos que não forneceram falsos negativos para a classe de sêmen humano. Portanto, a abordagem utilizando sPLS-DA foi considerada a mais apropriada como técnica de análise confirmatória para a aplicação sugerida, considerando o conjunto de tecidos testados.

Nas ciências forenses, um dos maiores inconvenientes para implementar metodologias analíticas é lidar com a grande variabilidade de situações em que os analitos podem ser encontrados. Por esse motivo, a alta influência dos tecidos nos perfis espectrais mostra o quão importante o estudo do substrato é para a implementação dessas técnicas no contexto das ciências forenses. Ainda assim, foi possível explorar o potencial da HSI-NIR como uma metodologia *screening* presuntiva e confirmatória para identificar manchas de sêmen de forma não destrutiva e levar em consideração a variabilidade dos substratos. Uma vez que o potencial da

espectroscopia NIR se mostrou útil para a detecção de sêmen, equipamentos portáteis podem ser utilizados como instrumentos para análise confirmatória dessas amostras.

5 PERSPECTIVAS FUTURAS

5.1 DIFERENCIAÇÃO DE TINTAS

Um dos problemas encontrados em um dos estudos realizados por nosso grupo de pesquisa e que ainda carece de maior aprofundamento é a diferenciação de tintas em documentos cujo papel apresenta selos de segurança, marcas d'água, entre outros padrões que possam interferir no espectro tinta/papel. Esse tipo de estudo ainda não foi encontrado na literatura.

5.2 DATAÇÃO DE DOCUMENTOS

Para o problema de datação de documentos, um estudo comparativo já está em andamento. As mesmas amostras avaliadas por Infravermelho serão analisadas por espectroscopia Raman a comparação do potencial de cada técnica.

Outro estudo importante que será desenvolvido é a influência de determinados fatores de armazenamento no processo de envelhecimento do papel. Luminosidade, umidade e temperatura das condições experimentais de envelhecimento serão avaliadas por espectroscopia IR e Raman. É possível acompanhar as mudanças espectrais de amostras envelhecidas artificialmente e identificar se essas mudanças ocorrem de forma similar à que foi observada nos documentos naturalmente envelhecidos.

Outra técnica espectroscópica que pode ser utilizada para datação de documentos antigos é a espectroscopia de Terahertz. Esse método fornece espectros cuja intensidade de sinal possui regiões que estão diretamente associadas com o grau de cristalinidade da celulose, como mostrado por Vieira e Pasquini (VIEIRA; PASQUINI, 2014). Uma vez que equipamentos robustos estejam disponíveis no mercado, essa técnica pode ser investigada para datação de documentos antigos. Além disso, a espectroscopia de infravermelho também pode ser combinada a outras técnicas espectroscópicas por fusão de dados, podendo fornecer mais informações e resultados adequados à aplicação forense.

5.3 IDENTIFICAÇÃO DE FLUIDOS BIOLÓGICOS

Particularmente para o estudo de sêmen, ainda é necessário um estudo mais amplo envolvendo diferentes substratos. Além de tecidos de alta absorção, é extremamente importante realizar estudos em tecidos estampados, de forma que seja possível identificar a influência da estampa nas imagens hiperespectrais.

Além disso, avaliar a variabilidade entre doadores também é importante. Como uma das amostras de sêmen apresentou falso negativo em um dos tecidos (não foi classificada em nenhuma classe), é possível que essa amostra contenha informações que não estejam contidas no conjunto de treinamento. Essa particularidade só ocorre para o tecido de Malha Bege, evidenciando, mais uma vez, a importância da influência dos substratos.

Estudos de datação envolvendo manchas de fluido seminal também podem ser realizados. Não só a estimativa da idade absoluta como a idade relativa de uma determinada mancha, mas também a comparação da idade desta mancha com outros possíveis fluidos corporais que possam ser encontrados numa cena de crime.

Outros trabalhos que já estão em desenvolvimento pelo nosso grupo de pesquisa envolvem a análise do sangue. A identificação e diferenciação de sangue por falsos-positivos em diferentes substratos está sendo realizada, além da datação de manchas que também já está em andamento.

REFERÊNCIAS

ADAM, C. D.; SHERRATT, S. L.; ZHOLOBENKO, V. L. Classification and individualization of black ballpoint pen inks using principal component analysis of UV-vis absorption spectra. **Forensic science international**, v. 174, n. 1, p. 16–25, jan. 2008.

ALI, M. et al. Spectroscopic studies of the ageing of cellulose paper. **Polymer**, v. 42, n. 2001, p. 2893–2900, 2001.

AMIGO, J. M.; BABAMORADI, H.; ELCOROARISTIZABAL, S. Hyperspectral image analysis. A tutorial. **Analytica Chimica Acta**, v. 896, p. 34–51, 2015.

ANDRADE, J. M. et al. Procrustes rotation in analytical chemistry, a tutorial. **Chemometrics and Intelligent Laboratory Systems**, v. 72, n. 2, p. 123–132, 2004.

AREA, M. C.; CHERADAME, H. Paper aging and degradation: Recent findings and research methods. **BioResources**, v. 6, n. 4, p. 5307–5337, 2011.

BANAS, K. et al. Multivariate analysis techniques in the forensics investigation of the postblast residues by means of Fourier transform-infrared spectroscopy. **Analytical chemistry**, v. 82, n. 7, p. 3038–44, 1 abr. 2010.

BARKER, M.; RAYENS, W. Partial least squares for discrimination. **Journal of Chemometrics**, v. 17, n. 3, p. 166–173, 2003.

BORBA, F. S. L.; HONORATO, R. S.; DE JUAN, A. Use of Raman spectroscopy and chemometrics to distinguish blue ballpoint pen inks. **Forensic Science International**, v. 249, p. 73–82, 2015.

BRERETON, R. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant**. West Sussex: John Wiley & Sons Ltd, 2003.

BRERETON, R. Support Vector Machines for Classification and Regression. **The Analyst**, v. 135, n. 2, p. 230–267, 2009.

BRERETON, R. G.; LLOYD, G. R. Partial least squares discriminant analysis: Taking the magic away. **Journal of Chemometrics**, v. 28, n. 4, p. 213–225, 2014.

BRO, R.; SMILDE, A. Principal component analysis. **Analytical Methods**, n. 6, p. 2812–2831, 2014.

BRUNELLE, R. L.; CRAWFORD, K. R. **Advances in the forensic analysis and dating of writing ink**. [s.l: s.n.].

BULFINCH, T. **O Livro de Ouro da Mitologia: Histórias de Deuses e Heróis**. 26ª ed. Rio de Janeiro: Ediouro Publicações S.A., 2002.

BURNS, D. A.; CIURCZAK, E. W. **Handbook of near-infrared analysis, 3rd ed.** 3. ed. Boca Raton: Taylor & Francis Group, 2009. v. 393

CALCERRADA, M.; GARCÍA-RUIZ, C. Analysis of questioned documents: A review. **Analytica Chimica Acta**, v. 853, n. 1, p. 143–166, 2015.

CALVINI, R.; ULRICI, A.; AMIGO, J. M. Practical comparison of sparse methods for classification of Arabica and Robusta coffee species using near infrared hyperspectral imaging. **Chemometrics and Intelligent Laboratory Systems**, v. 146, p. 503–511, 2015.

CAO, K.-A. L. et al. A Sparse PLS for Variable Selection when Integrating Omics Data. **Statistical Applications in Genetics and Molecular Biology**, v. 7, n. 1, 2008.

CAO, K. A. L.; BOITARD, S.; BESSE, P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. **Bmc Bioinformatics**, v. 12, p. 253, 2011.

CASADIO, F. et al. Erratum: Identification of organic colorants in fibers, paints, and glazes by surface enhanced raman spectroscopy. **Accounts of Chemical Research**, v. 43, n. 6, p. 782–791, 2010.

COATES, J. P. Infrared Spectroscopy for Process Analytical Applications. In: BAKEEV, K. A. (Ed.). . **Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries**. 2. ed. Oxford: Blackwell Publishing Ltd, 2010. p. 91–132.

CRANE, N. J. et al. Infrared spectroscopic imaging for noninvasive detection of latent fingerprints. **Journal of Forensic Sciences**, v. 52, n. 1, p. 48–53, 2007.

DA SILVA, V. A. G. et al. Discrimination of black pen inks on writing documents using visible reflectance spectroscopy and PLS-DA. **Journal of the Brazilian Chemical Society**, v. 25, n. 9, p. 1552–1564, 2014.

DE JUAN, A. et al. Chemometrics tools for image analysis. In: SALZER, H. W.; SIESLER, R. . (Eds.). . **Infrared and Raman Spectroscopic Imaging**. Weinheim: WILEY-VCH, 2009. p. 65–106.

DE JUAN, A.; JAUMOT, J.; TAULER, R. Multivariate Curve Resolution (MCR). Solving the mixture analysis problem. **Analytical Methods**, v. 6, n. 14, p. 4964, 2014.

DE JUAN, A.; TAULER, R. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. **Analytica Chimica Acta**, v. 500, n. 1–2, p. 195–210, 2003.

DE LA OSSA, M. Á. F.; AMIGO, J. M.; GARCÍA-RUIZ, C. Detection of residues from explosive manipulation by near infrared hyperspectral imaging: A promising forensic tool. **Forensic Science International**, v. 242, p. 228–235, 2014.

DEVOS, O.; DUPONCHEL, L. Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. **Chemometrics and Intelligent Laboratory Systems**, v. 107, n. 1, p. 50–58, 2011.

EDELMAN, G. J. et al. Infrared imaging of the crime scene: Possibilities and pitfalls. **Journal of Forensic Sciences**, v. 58, n. 5, p. 1156–1162, 2013.

EDELMAN, G.; VAN LEEUWEN, T. G.; AALDERS, M. C. G. Hyperspectral imaging for the age estimation of blood stains at the crime scene. **Forensic Science International**, v. 223, n. 1–3, p. 72–77, 2012.

ESBENSEN, K.; GELADI, P. Strategy of multivariate image analysis (MIA). **Chemometrics and Intelligent Laboratory Systems**, v. 7, n. 1–2, p. 67–86, dez. 1989.

ESKILDSEN, C. E. A. **Prediction of milk quality parameters using vibrational spectroscopy and chemometrics – opportunities and challenges in milk phenotyping**. [s.l.] University of Copenhagen, 2016.

EZCURRA, M. et al. Analytical methods for dating modern writing instrument inks on paper. **Forensic Science International**, v. 197, n. 1–3, p. 1–20, 2010.

FEARN, T. On orthogonal signal correction. **Chemometrics and Intelligent Laboratory Systems**, v. 50, n. 1, p. 47–52, 2000.

FEARN, T. et al. On the geometry of SNV and MSC. **Chemometrics and Intelligent Laboratory Systems**, v. 96, n. 1, p. 22–26, mar. 2009.

FERREIRA, M. M. C. **Quimiometria: conceitos, métodos e aplicações**. Campinas: Editora da Unicamp, 2015.

FILZMOSER, P.; GSCHWANDTNER, M.; TODOROV, V. Review of sparse methods in regression and classification with application to chemometrics. **Journal of Chemometrics**, v. 26, n. 3–4, p. 42–51, 2012.

FRIEDMAN, J. H.; TUCKEY, J. W. A Projection Pursuit Algorithm for Exploratory Data Analysis. **IEEE Transactions on Computers**, v. c-23, n. 9, p. 881–890, 1974.

GASTEIGER, J.; ENGEL, T. **Chemoinformatics**. Weinheim: WILEY-VCH Verlag GmbH & Co, 2003.

GEMPERLINE, P. J.; KALIVAS, J. H. Sampling Theory, Distribution Functions, and the Multivariate Normal Distribution. In: GEMPERLINE, P. J. (Ed.). . **Practical Guide to Chemometrics**. 2. ed. Boca Raton: Taylor & Francis Group, 2006. p. 41–67.

GHOSH, S.; RODGERS, J. NIR Analysis of Textiles. In: BURNS, D. A.; CIURCZAK, E. W. (Eds.). . **Handbook of near-infrared analysis**. 3. ed. Boca Raton: Taylor & Francis Group, 2008. p. 485–520.

GOODALL, C. Procrustes methods in the statistical analysis of shape. **Journal of the Royal Statistical Society**, v. 53, n. 2, p. 285–339, 1991.

GRAHN, H. F.; GELADI, P. **Techniques and Applications of Hyperspectral Image Analysis**. 1. ed. West Sussex: John Wiley & Sons, 2007. v. 22

GRIFFITHS, P. R. Infrared and Raman Instrumentation for Mapping and Imaging. In: SALZER, R.; SIESLER, H. W. (Eds.). . **Infrared and Raman Spectroscopic Imaging**.

Wetnhetm: WILEY-VCH Verlag GmbH & Co, 2009. p. 3–64.

GUO, Q. et al. Feature selection in sequential projection pursuit. **Analytica Chimica Acta**, v. 446, n. 1–2, p. 85–96, 2001.

HAJJI, L. et al. Artificial aging paper to assess long-term effects of conservative treatment. Monitoring by infrared spectroscopy (ATR-FTIR), X-ray diffraction (XRD), and energy dispersive X-ray fluorescence (EDXRF). **Microchemical Journal**, v. 124, p. 646–656, 2016.

HALL, P. On Polynomial-based Projection Indices for Exploratory Projection Pursuit. **Statistics**, v. 17, n. 2, p. 589–605, 1989.

HANS, K. M.-C. et al. Infrared detection of cocaine and street cocaine in saliva with a one-step extraction. **Analytical Methods**, v. 6, n. 3, p. 666, 2014.

HOU, S.; WENTZELL, P. D. Fast and simple methods for the optimization of kurtosis used as a projection pursuit index. **Analytica Chimica Acta**, v. 704, n. 1–2, p. 1–15, 2011.

HOU, S.; WENTZELL, P. D. Regularized projection pursuit for data with a small sample-to-variable ratio. **Metabolomics**, v. 10, n. 4, p. 589–606, 2014.

HUBER, P. Projection Pursuit. **The Annals of Statistics**, v. 13, n. 2, p. 435–475, 1985.

HURLEY, J. R.; CATTELL, R. B. The procrustes program: Producing direct rotation to test a hypothesized factor structure. **Behavioral Science**, v. 7, n. 2, p. 258–262, 1962.

ISHIKAWA, D. et al. Application of a newly developed portable NIR imaging device to monitor the dissolution process of tablets. **Analytical and Bioanalytical Chemistry**, v. 405, n. 29, p. 9401–9409, 2013.

JARVIS, R. M.; GOODACRE, R. Genetic algorithm optimization for pre-processing and variable selection of spectroscopic data. **Bioinformatics**, v. 21, n. 7, p. 860–868, 2005.

JAUMOT, J.; DE JUAN, A.; TAULER, R. MCR-ALS GUI 2.0: New features and applications. **Chemometrics and Intelligent Laboratory Systems**, v. 140, p. 1–12, 2015.

JONES, M. C.; SIBSON, R. What is Projection Pursuit? **Journal of the Royal Society. Series A (General)**, v. 150, n. 1, p. 1–36, 1987.

KAČÍK, F. et al. Cellulose degradation in newsprint paper ageing. **Polymer Degradation and Stability**, v. 94, n. 9, p. 1509–1514, 2009.

KHER, A. et al. Forensic classification of ballpoint pen inks using high performance liquid chromatography and infrared spectroscopy with principal components analysis and linear discriminant analysis. **Vibrational Spectroscopy**, v. 40, n. 2, p. 270–277, mar. 2006.

KJELDAHL, K.; BRO, R. Some common misunderstandings in chemometrics. **Journal of Chemometrics**, v. 24, n. 7–8, p. 558–564, 2010.

KOWALSKI, B. R.; BEEBE, K. R. An Introduction lo Multivariate calibration and Analysis. **Analytical Chemistry**, v. 59, p. 1007–1017, 1987.

KRZANOWSKI, W. J. **Principles of multivariate analysis : a user's perspective**. Revised ed. New York: Oxford University Press, 2000.

KUMAR, R.; KUMAR, V.; SHARMA, V. Fourier transform infrared spectroscopy and chemometrics for the characterization and discrimination of writing/photocopier paper types: Application in forensic document examinations. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 170, p. 19–28, 2017.

KUULA, J. et al. Using VIS/NIR and IR spectral cameras for detecting and separating crime scene details. **SPIE Defense, Security, and Sensing**, v. 8359, p. 83590P–83590P, 2012.

MARTÍNEZ, J. R. et al. Monitoring the natural aging degradation of paper by fluorescence. **Journal of Cultural Heritage**, p. 1–6, 2017.

MCLAUGHLIN, G.; LEDNEV, I. K. In situ identification of semen stains on common substrates via Raman spectroscopy. **Journal of Forensic Sciences**, v. 60, n. 3, p. 595–604, 2015.

MITCHELL, T. M. **Machine Learning**. Redmond: McGraw-Hill Science/Engineering/Math, 1997.

MURO, C. K. et al. Vibrational Spectroscopy : Recent Developments to Revolutionize Forensic Science. **Analytical chemistry**, v. 87, p. 306–327, 2015.

ORPHANOU, C.-M. The detection and discrimination of human body fluids using ATR FT-IR spectroscopy. **Forensic Science International**, v. 252, p. e10–e16, 2015.

PASQUINI, C. Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. **JOURNAL OF THE BRAZILIAN CHEMICAL SOCIETY**, v. 14, n. 2, p. 198–219, 2003.

PAYNE, G. et al. Visible and near-infrared chemical imaging methods for the analysis of selected forensic samples. **Talanta**, v. 67, n. 2, p. 334–344, 2005.

PEÑA, D.; PRIETO, F. Multivariate Outlier Detection and Robust Covariance Matrix Estimation. **Technometrics**, v. 43, n. 3, p. 286–310, 2001a.

PEÑA, D.; PRIETO, F. J. Cluster Identification Using Projections Cluster Identification Using Projections. **Journal of the American Statistical Association**, v. 96, p. 1433–1445, 2001b.

PRATS-MONTALBÁN, J. M.; DE JUAN, A.; FERRER, A. Multivariate image analysis: A review with applications. **Chemometrics and Intelligent Laboratory Systems**, v. 107, n. 1, p. 1–23, maio 2011.

RASMUSSEN, M. A.; BRO, R. A tutorial on the Lasso approach to sparse modeling. **Chemometrics and Intelligent Laboratory Systems**, v. 119, p. 21–31, 2012.

RINNAN, Å.; BERG, F. V. D.; ENGELSEN, S. B. Review of the most common pre-processing techniques for near-infrared spectra. **TrAC Trends in Analytical Chemistry**, v. 28, n. 10, p. 1201–1222, nov. 2009.

SANCHEZ, E.; KOWALSKI, B. R. Tensorial calibration: I. First-order calibration. **Journal of Chemometrics**, v. 2, n. 4, p. 247–263, 1988.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627–1639, 1964.

SCHEDL, A. et al. Aging of paper – Ultra-fast quantification of 2,5-

dihydroxyacetophenone, as a key chromophore in cellulosics, by reactive paper spray-mass spectrometry. **Talanta**, v. 167, p. 672–680, 2017.

SEMOLINI, R. **Support Vector Machines, Inferência Transdutiva e o Problema de Classificação**. [s.l.] Unicamp, 2002.

SIKIRZHYTSKI, V.; SIKIRZHYTSKAYA, A.; LEDNEV, I. K. Advanced statistical analysis of Raman spectroscopic data for the identification of body fluid traces: Semen and blood mixtures. **Forensic Science International**, v. 222, n. 1–3, p. 259–265, 2012.

SIKIRZHYTSKI, V.; VIRKLER, K.; LEDNEV, I. K. Discriminant analysis of Raman spectra for body fluid identification for forensic purposes. **Sensors**, v. 10, n. 4, p. 2869–2884, 2010.

SILVA, C. S. et al. Classification of blue pen ink using infrared spectroscopy and linear discriminant analysis. **Microchemical Journal**, p. 8–13, mar. 2012.

SJÖBLOM, J. et al. An evaluation of orthogonal signal correction applied to calibration transfer of near infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 44, n. 1–2, p. 229–244, 1998.

SKOOG, D. A.; HOLLER, F. J.; CROUCH, S. R. **Princípios de Análise Instrumental**. 6. ed. Porto Alegre: Bookman, 2009.

SMILDE, A; BRO, R.; GELADI, P. **Multi-way Analysis with Applications in the Chemical Sciences**. West Sussex: John Wiley & Sons Ltd, 2004.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199–222, 2004.

TAULER, R. Multivariate curve resolution applied to second order data. **Chemometrics and Intelligent Laboratory Systems**, v. 30, n. 1, p. 133–146, 1995.

TAULER, R.; SMILDE, A.; KOWALSKI, B. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. **Journal of Chemometrics**, v. 9, n. 1, p. 31–58, 1995.

THANASOULIAS, N. C.; PARISIS, N. A.; EVMIRIDIS, N. P. Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their Vis spectra. **Forensic Science International**, v. 138, p. 75–84, 2003.

THOMAS, J. et al. Raman spectroscopy and the forensic analysis of black/grey and blue cotton fibres: Part 1. Investigation of the effects of varying laser wavelength. **Forensic Science International**, v. 152, n. 2–3, p. 189–197, 2005.

TRAFELA, T. et al. Nondestructive analysis and dating of historical paper based on IR spectroscopy and chemometric data evaluation. **Analytical Chemistry**, v. 79, n. 16, p. 6319–6323, 2007.

UDRIŞTIOIU, F. M. et al. Application of Micro-Raman and FT - IR Spectroscopy in Forensic Analysis of Questioned Documents. **GU J Sci**, v. 25, n. 6, p. 371–375, 2012.

VIEIRA, F. S.; PASQUINI, C. Determination of cellulose crystallinity by terahertz-time domain spectroscopy. **Analytical Chemistry**, v. 86, n. 8, p. 3780–3786, 2014.

VIRKLER, K.; LEDNEV, I. K. Raman spectroscopy offers great potential for the nondestructive confirmatory identification of body fluids. **Forensic Science International**, v. 181, n. 1–3, p. 1–5, 2008.

WENTZELL, P. D. et al. Procrustes rotation as a diagnostic tool for projection pursuit analysis. **Analytica Chimica Acta**, v. 877, p. 51–63, 2015.

WIDENER, A.; DRAHL, C. Forcing Change in Forensic Science. **Chemical & Engeneering News**, v. 92, p. 10–14, 2014.

WILLIAMS, J. C. A Review of Paper Quality and Paper Chemistry. **Lybrary Trends**, v. Paper Qual, p. 203–224, 1981.

WINDIG, W.; STEPHENSON, D. A. Self-Modeling Mixture Analysis of 2Nd-Derivative Near-Infrared Spectral Data Using the Simplisma Approach. **Analytical Chemistry**, v. 64, n. 22, p. 2735–2742, 1992.

WISE, B. M. et al. **Chemometrics Tutorial for PLS _ Toolbox and Solo**. Wenatchee: Eigenvector Research, Inc., 2006.

WOLD, S. et al. Orthogonal signal correction of near-infrared spectra. **Chemometrics and Intelligent Laboratory Systems**, v. 44, n. 1–2, p. 175–185, 1998.

WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics Intelligent Laboratory Systems**, v. 2, p. 37–52, 1987.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109–130, 28 out. 2001.

WORKMAN, J.; WEYER, L. **Practical Guide to Interpretive Near-Infrared Spectroscopy**. [s.l: s.n.].

ZAPATA, F.; FERNÁNDEZ DE LA OSSA, M. Á.; GARCÍA-RUIZ, C. Emerging spectrometric techniques for the forensic analysis of body fluids. **TrAC - Trends in Analytical Chemistry**, v. 64, p. 53–63, 2015.

ZAPATA, F.; ORTEGA-OJEDA, F. E.; GARCÍA-RUIZ, C. Revealing the location of semen, vaginal fluid and urine in stained evidence through near infrared chemical imaging. **Talanta**, v. 166, p. 292–299, 2017.

ZIEBA-PALUS, J. et al. Analysis of degraded papers by infrared and raman spectroscopy for forensic purposes. **Journal of molecular structure**, p. 1–9, 2016.

ZOU, X.; UESAKA, T.; GURNAGUL, N. Prediction of paper permanence by accelerated aging I. Kinetic analysis of the aging process. **Cellulose**, v. 3, n. 1, p. 243–267, 1996a.

ZOU, X.; UESAKA, T.; GURNAGUL, N. Prediction of paper permanence by accelerated aging II. Comparison of the predictions with natural aging results. **Cellulose**, v. 3, n. 1, p. 269–279, 1996b.

APÊNDICES

APÊNDICE A – Resultados dos modelos PLS-DA para o conjunto dos Tecidos Brancos/Beges, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER).

	Tecidos Brancos/Beges						
	Algodão Branco						
	Validação				Previsão		
	LV	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	3	1.00	0.97	0.02	1.00	1.00	0.00
Tecido	5	0.97	0.98	0.03	1.00	0.80	0.10
Lubrificante	4	0.99	0.99	0.01	0.85	0.99	0.08
Sêmen	**5**	**0.98**	**0.99**	**0.02**	**0.92**	**0.99**	**0.05**
	Algodão Bege						
	Validação				Previsão		
	LV	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	3	0.99	0.99	0.01	0.95	0.99	0.03
Tecido	4	0.96	0.91	0.06	0.96	0.85	0.10
Lubrificante	3	1.00	0.99	0.01	1.00	0.97	0.01
Sêmen	**4**	**0.97**	**0.94**	**0.04**	**0.95**	**0.92**	**0.06**
	Cetim Branco						
	Validação				Previsão		
	LV	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	4	1.00	1.00	1.22e^-5	1.00	0.93	0.04
Tecido	4	0.99	0.97	0.02	0.74	0.89	0.18
Lubrificante	4	0.99	0.99	0.01	1.00	0.96	0.02
Sêmen	**4**	**0.92**	**0.96**	**0.06**	**0.89**	**0.94**	**0.09**
	Malha Branca						
	Validação				Previsão		
	LV	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	6	0.99	0.98	0.02	1.00	0.94	0.03
Tecido	6	0.93	0.93	0.07	0.92	0.83	0.12
Lubrificante	3	0.97	1.00	0.02	**0.51**	0.99	0.25
Sêmen	**5**	**0.85**	**0.89**	**0.13**	**0.86**	**0.95**	**0.10**

	Malha Bege						
	Validação				Previsão		
	LV	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	4	0.99	0.98	0.01	0.99	0.93	0.04
Tecido	4	0.92	0.91	0.09	0.76	0.94	0.15
Lubrificante	6	0.91	0.97	0.06	0.79	0.94	0.13
Sêmen	**3**	**0.98**	**0.92**	**0.05**	**0.83**	**0.77**	**0.20**

APÊNDICE B – Resultados dos modelos PLS-DA para o conjunto dos Tecidos Coloridos, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER).

	Tecidos Coloridos								
	Branco					Vermelho			
	Validação					Validação			
	LV	Sn	Sp	ER		LV	Sn	Sp	ER
Leite Materno	4	0.93	0.99	0.04	Leite Materno	5	0.94	0.99	0.04
Tecido	4	0.96	0.87	0.08	Tecido	7	0.97	0.92	0.06
Sêmen de Bode	3	1.00	0.99	0.01	Sêmen de Bode	4	1.00	1.00	0.00
Sêmen de Cavalo	3	0.98	0.99	0.01	Sêmen de Cavalo	3	0.98	0.95	0.04
Lubrificante	3	0.99	0.96	0.03	Lubrificante	4	0.99	0.96	0.02
Sêmen Humano	**3**	**0.96**	**0.96**	**0.04**	**Sêmen Humano**	**4**	**0.98**	**0.96**	**0.03**
	Preto					Amarelo			
	Validação					Validação			
	LV	Sn	Sp	ER		LV	Sn	Sp	ER
Leite Materno	5	0.96	0.96	0.04	Leite Materno	3	0.92	0.96	0.06
Tecido	5	0.82	0.84	0.17	Tecido	5	0.90	0.82	0.14
Sêmen de Bode	5	1.00	0.98	0.01	Sêmen de Bode	5	0.99	1.00	0.01
Sêmen de Cavalo	4	0.97	0.94	0.04	Sêmen de Cavalo	3	0.99	0.98	0.02
Lubrificante	6	0.98	0.95	0.04	Lubrificante	4	0.99	0.95	0.03
Sêmen Humano	**3**	**0.96**	**0.65**	**0.19**	**Sêmen Humano**	**5**	**0.95**	**0.89**	**0.08**
	Verde								
	Validação								
	LV	Sn	Sp	ER					
Leite Materno	6	0.87	0.98	0.07					
Tecido	4	0.91	0.89	0.10					
Sêmen de Bode	4	1.00	1.00	0.00					
Sêmen de Cavalo	3	0.98	0.99	0.02					
Lubrificante	3	0.99	0.94	0.04					
Sêmen Humano	**4**	**0.95**	**0.89**	**0.08**					

APÊNDICE C – Resultados dos modelos sPLS-DA para o conjunto dos Tecidos Brancos/Beges, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER).

	Tecidos Brancos/Beges							
	Algodão Branco							
	Validação					Previsão		
	sLV	Var. Incl.	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	7	34	0.99487	1	0.00434	0.9487	1.00	0.0452
Tecido	8	33	0.979	0.974	0.0214	0.912	0.9131	0.0877
Lubrificante	8	28	0.997	0.9939	0.00392	0.9305	0.9294	0.06971
Sêmen	**8**	**68**	**0.9892**	**0.9919**	**0.00992**	**0.9829**	**0.9106**	**0.051**
	Algodão Bege							
	Validação					Previsão		
	sLV	Var. Incl.	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	9	20	1.00	1	0	0.997	0.974	0.006
Tecido	9	71	0.9459	0.9722	0.0479	0.954	0.941	0.049
Lubrificante	9	52	0.9973	0.9979	0.0024	0.966	0.995	0.027
Sêmen	**8**	**40**	**0.96**	**0.97**	**0.04**	**0.937**	**0.940**	**0.062**
	Cetim Branco							
	Validação					Previsão		
	sLV	Var. Incl.	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	4	29	1.000	1.000	0.0000	0.998	1.000	0.002
Tecido	10	42	0.983	0.997	0.0136	0.898	**0.599**	0.158
Lubrificante	9	33	0.995	0.991	0.0064	0.988	0.978	0.014
Sêmen	**10**	**67**	**0.963**	**0.958**	**0.038**	**0.589**	**0.957**	**0.250**
	Malha Branca							
	Validação					Previsão		
	sLV	Var. Incl.	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	9	30	0.98	0.99	0.018	0.96	1.00	0.03
Tecido	10	46	0.93	0.95	0.062	0.88	0.87	0.12
Lubrificante	9	51	0.99	0.99	0.007	0.98	**0.56**	0.14
Sêmen	**10**	**39**	**0.89**	**0.89**	**0.108**	**0.88**	**0.95**	**0.10**

	Malha Bege							
	Validação					Previsão		
	sLV	Var. Incl.	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	9	33	0.99	1.00	0.0062	0.96	0.99	0.04
Tecido	9	35	0.93	0.95	0.0618	0.86	0.93	0.13
Lubrificante	10	29	0.97	0.94	0.0378	0.73	0.97	0.21
Sêmen	**9**	**34**	**0.97**	**0.97**	**0.0296**	**0.92**	**0.68**	**0.18**

APÊNDICE D – Resultados dos modelos sPLS-DA para o conjunto dos Tecidos Coloridos, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER).

	Tecidos Coloridos										
	Branco						Vermelho				
	Validação						Validação				
	sLV	Var. Incl.	Sn	Sp	ER		sLV	Var. Incl.	Sn	Sp	ER
Leite Materno	9	66	1.00	0.95	0.01	Leite Materno	7	20	0.98	0.95	0.02
Tecido	9	35	0.91	0.96	0.08	Tecido	8	25	0.91	0.96	0.08
Sêmen de Bode	9	26	1.00	1.00	0.00	Sêmen de Bode	7	56	0.99	1.00	0.00
Sêmen de Cavalo	6	29	0.99	0.98	0.01	Sêmen de Cavalo	6	38	0.95	0.97	0.05
Lubrificante	5	15	0.96	0.99	0.03	Lubrificante	5	24	0.98	0.98	0.02
Sêmen Humano	**9**	**44**	**0.96**	**0.96**	**0.04**	**Sêmen Humano**	**6**	**26**	**0.97**	**0.98**	**0.03**
	Preto						Amarelo				
	Validação						Validação				
	sLV	Var. Incl.	Sn	Sp	ER		sLV	Var. Incl.	Sn	Sp	ER
Leite Materno	9	62	0.97	0.96	0.03	Leite Materno	7	30	0.97	0.95	0.03
Tecido	6	78	0.81	0.80	0.19	Tecido	8	42	0.83	0.91	0.15
Sêmen de Bode	8	73	0.98	1.00	0.01	Sêmen de Bode	8	40	1.00	0.99	0.00
Sêmen de Cavalo	8	45	0.96	0.97	0.04	Sêmen de Cavalo	7	27	0.96	0.98	0.03
Lubrificante	9	41	0.95	0.97	0.05	Lubrificante	6	27	0.96	1.00	0.03
Sêmen Humano	**9**	**71**	**0.77**	**0.88**	**0.20**	**Sêmen Humano**	**8**	**48**	**0.92**	**0.94**	**0.08**
	Verde										
	Validação										
	sLV	Var. Incl.	Sn	Sp	ER						
Leite Materno	9	32	0.98	0.88	0.03						
Tecido	8	43	0.89	0.94	0.10						
Sêmen de Bode	7	41	1.00	1.00	0.00						
Sêmen de Cavalo	8	48	0.97	0.97	0.03						
Lubrificante	7	44	0.96	0.99	0.03						
Sêmen Humano	**8**	**53**	**0.91**	**0.95**	**0.08**						

APÊNDICE E – Resultados dos modelos SVM-DA para o conjunto dos Tecidos Brancos/Beges, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER), número de vetores de suporte (SV).

	Tecidos Brancos/Beges									
	Algodão Branco									
	Validação							Previsão		
	LV	SV	nu	gamma	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	3	495	0.076	3.16	0.99	1.00	0.00	0.99	1.00	0.01
Tecido	5	1159	0.114	3.16	0.96	1.00	0.02	0.96	0.94	0.05
Lubrificante	4	935	0.152	3.16	0.99	1.00	0.01	0.81	0.99	0.10
Sêmen	**5**	**852**	**0.152**	**1.00**	**0.97**	**1.00**	**0.01**	**0.95**	**0.99**	**0.03**
	Algodão Bege									
	Validação							Previsão		
	LV	SV	nu	gamma	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	3	382	0.076	0.31623	1.00	1.00	0.00	0.95	1.00	0.03
Tecido	4	771	0.114	1.00	0.86	0.98	0.08	0.85	0.97	0.09
Lubrificante	3	795	0.152	3.162	0.99	1.00	0.00	1.00	0.99	0.01
Sêmen	**4**	**1326**	**0.152**	**10.0**	**0.93**	**0.95**	**0.06**	**0.84**	**0.96**	**0.10**
	Cetim Branco									
	Validação							Previsão		
	LV	SV	nu	gamma	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	4	922	0.190	1.00E-06	1.0	1.0	0.0	1.0	1.0	0.0
Tecido	4	556	0.114	0.001	1.0	1.0	0.0	**0.6**	0.9	0.2
Lubrificante	4	760	0.152	1	1.0	1.0	0.0	1.0	1.0	0.0
Sêmen	**4**	**827**	**0.152**	**3.1623**	**1.0**	**1.0**	**0.0**	**0.9**	**1.0**	**0.1**
	Malha Branca									
	Validação							Previsão		
	LV	SV	nu	gamma	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	6	419	0.076	0.31623	0.95	1.00	0.02	0.99	0.98	0.02
Tecido	6	573	0.114	0.01	0.98	0.97	0.02	0.96	0.95	0.05
Lubrificante	6	738	0.152	3.16E-06	0.98	1.00	0.01	**0.52**	1.00	0.24
Sêmen	**5**	**1124**	**0.152**	**3.1623**	**0.90**	**0.97**	**0.06**	**0.90**	**0.96**	**0.07**

	Malha Bege									
	Validação							Previsão		
	LV	SV	nu	gamma	Sn	Sp	ER	Sn	Sp	ER
Leite Materno	5	467	0.076	1	0.97	1.00	0.02	0.83	0.99	0.09
Tecido	5	573	0.114	0.001	0.95	0.98	0.04	0.74	0.99	0.14
Lubrificante	7	1305	0.152	1	0.94	0.99	0.03	0.70	0.92	0.19
Sêmen	**4**	**1429**	**0.152**	**10**	**0.93**	**0.98**	**0.05**	**0.81**	**0.94**	**0.13**

APÊNDICE F – Resultados dos modelos SVM-DA para o conjunto dos Tecidos Coloridos, mostrando o número de variáveis Latentes (LV), especificidade (Sp), sensibilidade (Sn) e taxa de erros (ER), número de vetores de suporte (SV).

	Tecidos Coloridos						
	Branco						
	Validação						
	LV	SV	nu	gamma	Sn	Sp	ER
Leite Materno	5	573	0.035344	1	0.906	0.999	0.0474
Tecido	5	860	0.10603	1	0.937	0.987	0.0379
Sêmen de Bode	4	209	0.035344	0.31623	0.992	1	0.00423
Sêmen de Cavalo	4	346	0.035344	1	0.97	1	0.0147
Lubrificante	4	1071	0.14147	31,623	0.974	0.987	0.0192
Sêmen Humano	**4**	**818**	**0.14147**	**1**	**0.972**	**0.991**	**0.0186**
	Preto						
	Validação						
	LV	SV	nu	gamma	Sn	Sp	ER
Leite Materno	6	239	0.035	0.31623	0.91	1.00	0.05
Tecido	6	1984	0.265	1.00	0.74	0.96	0.15
Sêmen de Bode	6	203	0.035	0.10	0.96	1.00	0.02
Sêmen de Cavalo	5	190	0.035	0.031623	0.91	1.00	0.04
Lubrificante	7	895	0.141	0.31623	0.97	0.99	0.02
Sêmen Humano	**4**	**937**	**0.141**	**1.00**	**0.88**	**0.93**	**0.09**
	Verde						
	Validação						
	LV	SV	nu	gamma	Sn	Sp	ER
Leite Materno	7	219	0.035	0.01	0.86	1.00	0.07
Tecido	5	836	0.106	1	0.92	0.99	0.04
Sêmen de Bode	5	188	0.035	0.031623	1.00	1.00	0.00
Sêmen de Cavalo	4	196	0.035	0.1	0.97	1.00	0.02
Lubrificante	5	770	0.141	0.31623	0.98	0.98	0.02
Sêmen Humano	**5**	**1891**	**0.141**	**31,623**	**0.95**	**0.98**	**0.03**

	Vermelho						
	Validação						
	LV	SV	nu	gamma	Sn	Sp	ER
Leite Materno	6	186	0.035	3.16E-02	0.93	1.00	0.04
Tecido	8	650	0.106	0.31623	0.95	0.99	0.03
Sêmen de Bode	5	462	0.088	1.00E-06	0.99	1.00	0.00
Sêmen de Cavalo	4	188	0.035	0.031623	0.98	1.00	0.01
Lubrificante	5	918	0.141	1	0.97	0.99	0.02
Sêmen Humano	**5**	**988**	**0.141**	**1**	**0.97**	**0.99**	**0.02**
	Amarelo						
	Validação						
	LV	SV	nu	gamma	Sn	Sp	ER
Leite Materno	4	499	0.035	1	0.86	1.00	0.07
Tecido	6	593	0.106	0.031623	0.93	0.98	0.05
Sêmen de Bode	6	288	0.035	0.31623	0.96	1.00	0.02
Sêmen de Cavalo	4	190	0.035	0.003162	0.98	1.00	0.01
Lubrificante	5	791	0.141	0.31623	0.96	0.99	0.02
Sêmen Humano	**6**	**1298**	**0.141**	**1**	**0.94**	**0.98**	**0.04**