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

ALINE COELHO SILVA FONSECA

IDENTIFICAÇÃO DE VESTÍGIOS DE SANGUE EM DIFERENTES SUBSTRATOS EMPREGANDO EQUIPAMENTO NIR PORTÁTIL E MODELO HIERÁRQUICO

Recife
2020

ALINE COELHO SILVA FONSECA

IDENTIFICAÇÃO DE VESTÍGIOS DE SANGUE EM DIFERENTES SUBSTRATOS EMPREGANDO EQUIPAMENTO NIR PORTÁTIL E MODELO HIERÁRQUICO

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Química Fundamental, PPGQ, da Universidade Federal de Pernambuco como parte dos requisitos necessários à obtenção do grau de Mestre em Química

Área de concentração: Química Analítica

Orientadora: Profª Drª Maria Fernanda Pimentel Avelar

Coorientador: Dr. José Francielson Q. Pereira

Recife
2020

Catalogação na fonte
Bibliotecária Mariana de Souza Alves CRB4-2105

F676i Fonseca, Aline Coelho Silva
Identificação de vestígios de sangue em diferentes substratos empregando equipamento NIR portátil e modelo hierárquico / Aline Coelho Silva Fonseca. – 2020.
96f.: il., fig., tab.
Orientadora: Maria Fernanda Pimentel Avelar.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CCEN, Química, Recife, 2020.
Inclui referências e apêndices.

1. Química Analítica. 2. Sangue Humano. 3. Modelagem Hierárquica. 4. Equipamento Portátil. I. Avelar, Maria Fernanda Pimentel. (orientadora) II. Título.

543 CDD (22. ed.) UFPE-CCEN 2020-175

ALINE COELHO SILVA FONSECA

IDENTIFICAÇÃO DE VESTÍGIOS DE SANGUE EM DIFERENTES SUBSTRATOS EMPREGANDO EQUIPAMENTO NIR PORTÁTIL E MODELO HIERÁRQUICO.

Dissertação apresentada ao Programa de Pós-Graduação no Departamento de Química Fundamental da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Química.

Aprovada em: 21/09/2020

BANCA EXAMINADORA

_____

Profa. Maria Fernanda Pimentel Avelar (Orientadora)
Universidade Federal de Pernambuco
Departamento de Engenharia Química

_____

Profa. Carolina Santos Silva
Universidade Federal de Pernambuco
Departamento de Engenharia Química

_____

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

_____

Profa. Fernanda Araújo Honorato
Universidade Federal de Pernambuco
Departamento de Engenharia Química

AGRADECIMENTO

Agradeço a Deus, ao Universo e a todos os seres aos quais estamos conectados, pois *"somos ondas do mesmo mar, folhas da mesma árvore e flores do mesmo jardim"*.

À minha professora e orientadora Maria Fernanda Pimentel Avelar, pela orientação, disponibilidade, confiança, paciência, sugestões, dedicação e conselhos. A professora Fernanda é um exemplo de profissional e uma inspiração para todos nós que estamos iniciando a vida acadêmica e temos a alegria de compartilhar seus ensinamentos.

Ao meu coorientador José Francielson Queiroz Pereira, pela amizade, pelos ensinamentos, paciência, dedicação, disponibilidade, sugestões, pelos comentários e dicas fundamentais para ultrapassar as dificuldades deste trabalho e pela disponibilização dos dados para a construção desse projeto.

À Maria Júlia Leal Vieira por fornecer os dados utilizados para o desenvolvimento desse trabalho. E a todos os voluntários que aceitaram participar do estudo e autorizaram a coleta das amostras de sangue.

À minha família, em especial aos meus pais, Arlete e Américo, e ao meu irmão Arthur, por todo o carinho, apoio, compreensão, conselhos e amor oferecidos, independente do momento.

A João Roberto pelo companheirismo, paciência, carinho, conselhos, sugestões e amor. Aos meus sogros e cunhado, que se tornaram minha segunda família.

Aos amigos Bruno, Erklaylle, Fernando, Flávia, Flávio e aos demais colegas do LAC pela parceria, pelas horas de estudo em grupo, pelos bolos, pelas figurinhas, pelos desabafos, pelas muitas risadas, e por todas as experiências compartilhadas.

A Carolina que sempre esteve com as portas abertas para tirar dúvidas, oferecer conselhos, ouvir meus desabafos, pelo suporte emocional e pelas aulas maravilhosas nas reuniões de grupo.

Ao Laboratório de Combustíveis (LAC) e a todos os seus integrantes que possibilitaram o desenvolvimento desse trabalho.

A cada integrante da Pós-Graduação e do Departamento de Química Fundamental - UFPE, especialmente aos professores, a Patrícia, e ao Sr. Ademias, por toda dedicação empregada ao programa e ao departamento.

Ao CNPq, pela bolsa de mestrado concedidas. Ao NUQAAPE, NEQUIFOR, INCTAA, FACEPE e a CAPES pelo incentivo ao projeto, à UFPE pelo suporte institucional.

"The truth, however ugly in itself, is always curious and beautiful to seekers after it."
(CHRISTIE, 2013, p.144)

RESUMO

Um dos tipos de evidência mais importantes em uma investigação criminal são vestígios de sangue humano. Para que essas evidências de sangue sejam investigadas, é preciso que elas sejam apropriadamente identificadas e coletadas no local do crime. Os métodos atualmente utilizados para identificá-las podem apresentar falsos positivos devido a contaminações ambientais por diversas substâncias, sejam industrializados ou naturais. Sabendo a necessidade da aplicação de metodologias abrangentes para identificação de sangue, o presente trabalho teve o objetivo de desenvolver modelos hierárquicos para classificação, baseados na espectroscopia de infravermelho próximo (NIR), capaz de identificar vestígios de sangue humano (SH), sangue animal (SA) e de falsos positivos comuns (FPC) em pisos com diferentes composições, cinco tipos de cerâmicas e quatro tipos de porcelanatos (com colorações e rugosidades diversas). Os espectros foram obtidos utilizando um espectrômetro portátil MicroNIR (Viavi) após seis dias de secagem. O conjunto de dados foi dividido em três conjuntos de treinamento (CT) para avaliar a robustez dos modelos. Foram criados três modelos hierárquicos para cada conjunto de treinamento através sucessão de técnicas de Análise de Componentes Principais (PCA), Modelagem Independente e Flexível de Analogia de Classe (SIMCA) ou Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA). Os modelos utilizaram duas regras de decisão. A primeira regra, baseada no critério Q-estatístico, funcionou como um filtro para separar as amostras de sangue das amostras de FPC, e utilizou o critério de Q-Estatístico obtido através de um modelo de Análise de Componentes Principais (PCA). A segunda regra de decisão foi construída para discriminar as amostras de sangue humano e sangue animal, sendo que para esse fim foram construídos modelos PCA, SIMCA e PLS-DA. Também foram construídos modelos para os mesmos conjuntos de treinamento com a faixa reduzida de 1300-1600 nm considerando a mesma metodologia anterior. Por último, desenvolveu-se um protocolo para avaliar a representatividade de um piso externo com relação a variabilidade dos pisos inclusos no conjunto de treinamento. Os modelos hierárquicos construídos usando PLS-DA como segunda regra de decisão obtiveram resultados de sensibilidade e especificidade iguais a 1. Os modelos construídos a partir dos conjuntos de treinamento mais abrangentes e que utilizaram um modelo PCA para discriminar as amostras de SH e SA obtiveram 100% de classificação correta para amostras de SH. Os modelos hierárquicos de classificação construídos para esses mesmos conjuntos de treinamento apresentaram os melhores resultados, onde todos os modelos construídos obtiveram valores de sensibilidade acima de 0,94 e especificidade acima de 0,78, isso ressalta a importância da escolha do conjunto de treinamento para a construção do modelo. O protocolo criado para avaliar a representatividade de um substrato com relação aos demais incluídos no modelo se mostrou bastante eficiente, indicando quando um piso externo não tinha a variabilidade contemplada indicando que o modelo de identificação de sangue poderia não ter a performance esperada.

Palavras-chaves: Sangue Humano. Modelagem Hierárquica. Equipamento Portátil. Infravermelho Próximo.

ABSTRACT

One of the most important evidence in a criminal investigation is the human blood traces. For the investigation of these blood samples to happen, they must be identified and collected at the crime scene. Currently the methods used to identify those samples can lead to false negative results due to environmental contamination by various substances, whether industrialized or natural. Knowing the need to apply comprehensive methods for blood identification, the present study aimed to develop hierarchical models for classification, based on near infrared (NIR) spectroscopy, capable of identifying traces of human blood (HB), animal blood (AB) and common false positives (CFP) on floors with different compositions, five types of ceramics and four types of porcelain tiles (with different colors and roughness). The spectra were obtained using a portable MicroNIR spectrometer (Viavi) after six days of drying. The data set was divided into three training sets (CT) to assess the robustness of the models. Three hierarchical models were created for each training set through succession of Principal Component Analysis (PCA), Independent and Flexible Class Analogy Modeling (SIMCA) or Discriminant Analysis by Partial Least Squares (PLS-DA). The models use two decision rules, where the first rule, based on the Q-statistical criterion, acted as a filter to separate blood samples from CFP, and used the Q-Statistical criterion obtained through a Principal Component Analysis (PCA) model. The second decision rule was built to discriminate human blood from animal blood samples, and for this purpose were built PCA, SIMCA and PLS-DA models. Models were also built for the same training sets with a reduced range of 1300-1600nm, considering the same previous methodology. Finally, a protocol was developed to evaluate the representativeness of a prediction floor in relation to the training sets floors variability. The hierarchical models built using PLS-DA as the second decision obtained the best results of sensitivity and specificity equal to 1. The models built through more embracing training sets and which used a PCA model to discriminate HB from AB samples classified correctly 100% of the HB samples. The hierarchical classification models built for these same training sets show the best results, all of them obtained values of sensitivity above 0.94 and specificity above 0.78, this highlights the importance of choosing the training set for a model construction. The protocol created to evaluate the representativeness of a substrate in relation to the training sets substrates proved to be quite efficient, when the predicted floor did not have his variability contemplated it indicated that the blood identification model could not performed as expected.

Keywords: Human Blood. Hierarchical Modelling. Handheld Equipment. Near Infrared.

LISTA DE ILUSTRAÇÕES

LISTAS DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ALS	Mínimos Quadrados Alternados
ANN	Rede Neural Artificial
ATR	Reflectância Total Atenuada
DA	Análise Discriminante
DM	Modelagem de Densidade
FN	Falso Negativo
FIR	Infravermelho Distante
FP	Falso Positivo
FPC	Falso Positivo Comum
FT	Transformada de Fourier
HCA	Análise de Agrupamentos Hierárquicos
IR	Infravermelho
LDA	Análise Discriminante Linear
MCR	Resolução Multivariada de Curvas
MH	Modelo Hierárquico
MIR	Infravermelho Médio
NIR	Infravermelho Próximo
PCA	Análise de Componentes Principais
PFM	Método de Função de Probabilidade
PLS	Mínimos Quadrados Parciais
PP	*Projection Pursuit*
PRESS	Soma Quadrática dos Resíduos de Previsão
RMSEC	Raiz do Erro Quadrático Médio de Calibração
RMSECV	Raiz do Erro Quadrático Médio de Validação
RMSEP	Raiz do Erro Quadrático Médio de Predição
SA	Sangue Animal
SIMCA	Modelagem Independente e Flexível por Analogia de Classes
SH	Sangue Humano
Sn	Sensibilidade
Sp	Especificidade
SPA	Algoritmo das Projeções Sucessivas

SVM	Máquinas de Vetores de Suporte
VP	Verdadeiro Positivo
VN	Verdadeiro Negativo
UFPE	Universidade Federal de Pernambuco

LISTA DE SÍMBOLOS

$\mathbf{X}$	Matriz do conjunto de dados
$\bar{x}$	Valor médio
$\mathbf{Y}$	Matriz de classes
$\mathbf{T}$	Matriz de escores
$\mathbf{E}$	Matriz de resíduos
T^2	Parâmetro T^2 de Hotteling
$\mathbf{t}$	Vetor de escores
Q	Parâmetro Q-Estatístico
e	Elementos da matriz E
$\mathbf{Q}$	Matriz de *loadings* de **Y**
$\mathbf{P}$	Matriz de *loadings* de **X**
$\widehat{\mathbf{Y}}$	Matriz de valores preditos para classes
$\widehat{\mathbf{B}}$	Matriz dos coeficientes de regressão
$\mathbf{X_{pred}}$	Matriz dos dados de predição

Sobrescrito

T	Transposta

Subscrito

i	Linhas da matriz (amostras)
j	Colunas da matriz (variáveis)

Símbolos gregos

Δ	Diferença
υ	Nível de energia vibracional

SUMÁRIO

APÊNDICE E - GRÁFICOS DOS ESPECTROS ORIGINAIS, MÉDIA DOS ESPECTROS ORIGINAIS, GRÁFICOS DOS ESPECTROS PRÉ-PROCESSADOS E MÉDIA DOS ESPECTROS PRÉ-PROCESSADOS, GRÁFICOS DE ESCORES DE PC1XPC2 E PC1XPC3 E SEUS RESPECTIVOS *LOADINGS*, E GRÁFICO DE INFLUÊNCIA CONSTRUÍDOS A PARTIR DO CONJUNTO DE DADOS COM FAIXA ESPECTRAL REDUZIDA. 95

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O Atlas da Violência de 2019 publicado pelo Instituto de Pesquisa Econômica Aplicada (IPEA), mostrou que em 2017 foi registrado um recorde de casos de homicídio no Brasil, totalizando 65.602 mortes, equivalendo a aproximadamente 32 mortes para cada cem mil habitantes. Contudo, a taxa de elucidação de homicídios no país ainda é desconhecida, apesar de ser estimado que em alguns estados ela seja em torno de 10% a 20% (CERQUEIRA et al., 2019). Para que mais casos sejam solucionados e a taxa de elucidação aumente, é necessário que as polícias científicas realizem investigações bem sucedidas.

O principal objetivo de um inquérito criminal é encontrar a verdade, sendo uma das principais ferramentas para alcançar essa solução: a reconstrução do cenário do crime. Nesse sentido, a análise de evidências é de extrema importância para a averiguação e reconstrução dos fatos ocorridos (BEVEL; M. GARDNER, 2008). Um dos tipos de evidências mais importantes são os fluidos corporais e, dentre eles, o sangue (TAKAMURA et al., 2018). As análises de vestígios de sangue podem responder perguntas como quem estava presente na cena, o que aconteceu, se havia ou não a presença de drogas, entre outras questões (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

Contudo, para que essas evidências de sangue sejam investigadas, é preciso que elas sejam apropriadamente identificadas e coletadas no local do crime. Muitas vezes, esses vestígios até conseguem ser visualmente identificados, mas por possuir uma coloração avermelhada, eles podem ser confundidos com outras substâncias que possuem colorações similares, mas não são sangue, os chamados falsos positivos comuns (EDELMAN et al., 2012). Quando uma evidência chega ao laboratório ela é submetida a testes capazes de determinar se sua natureza é realmente sangue. Esses testes geralmente envolvem reações químicas catalíticas e o uso de reagentes oxidantes e indicadores que mudam de cor no momento em que ocorre a oxidação catalisada pela hemoglobina (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

Para evitar que esses testes sejam realizados em amostras que não sejam realmente de sangue, e que tempo e reagentes sejam desperdiçados, o ideal é que testes presuntivos e confirmativos sejam realizados ainda nos locais de crime (VIRKLER; LEDNEV, 2009a). Esses testes devem ser rápidos, seguros, sensíveis, específicos e, principalmente, não devem contaminar ou destruir as amostras para não inviabilizar posteriores análises como a de DNA (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

Os testes presuntivos mais comumente utilizados baseiam-se em reações colorimétricas como o teste de Kastle-Meyer e Hemastix (CASALI et al., 2020), ou ainda na emissão de luz devido à quimioluminescência ou fluorescência quando reagido com o luminol (HAYASHI et al., 2019). Contudo, esses métodos apresentam algumas desvantagens como a possibilidade de ainda obterem resultados falsos positivos devido a contaminações ambientais por produtos de limpeza ou proteínas animais e vegetais (MISTEK; LEDNEV, 2015). Além disso, o uso do luminol precisa de alguns cuidados para que suas propriedades não sejam alteradas, como a necessidade dele ser mantido sob refrigeração, requerer preparação no momento de sua utilização e o uso em ambiente escuro (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

Outro fator importante que deve ser levado em consideração são os vestígios de sangue animal que podem ser confundidos com sangue humano. A maioria dos testes presuntivos não são específicos para sangue humano, levando novamente a necessidade de que testes confirmativos sejam realizados (ZHANG et al., 2016).

Diferentes estudos mostram que uma alternativa para se obter testes confirmativos na identificação de sangue humano é o uso de técnicas espectroscópicas vibracionais, como Infravermelho e Raman (VIRKLER; LEDNEV, 2009a); (ZOU et al., 2016). Essas técnicas têm caráter não invasivo e não destrutivo, sendo capazes de preservar a integridade dos vestígios analisados, e permitir que análises de contraprova ou complementares sejam realizadas.

Silva e colaboradores (2019) publicaram um trabalho de revisão que reuniu diversas aplicações das técnicas vibracionais combinadas com o uso da quimiometria para solução de problemas forenses, como o uso de espectroscopia Raman e Infravermelho para verificação de adulteração de documentos, identificação de resíduos de disparo de armas de fogo, investigação de drogas ilícitas, reconhecimento e datação de manchas de sangue, entre outras. Além disso, a popularização de equipamentos portáteis torna a aplicação dessas técnicas ainda mais interessantes para fins forenses, visto que eles são leves, apresentam um relativo baixo custo, e podem ser deslocados para realizar análises em campo permitindo que o perito ganhe tempo e não use reagentes desnecessariamente (PEREIRA et al., 2017). Contudo, deve ser levado em consideração que além da complexidade natural envolvendo a interpretação dos espectros obtidos no infravermelho próximo, por exemplo, o uso desses equipamentos portáteis em cenas de crime, torna a matriz dos dados coletados ainda mais complexa, visto que os vestígios podem ser encontrados em diferentes substratos, contextos e condições.

Para extrair informações de interesse de maneira mais precisa, rápida e abrangente faz-se necessário o uso de métodos estatísticos multivariados, também conhecidos como técnicas quimiométricas. Atualmente, existem muitos métodos que podem ser aplicados para auxiliar na

identificação de amostras de sangue, assim como na estimativa da idade de sua deposição (SHARMA; KUMAR, 2018).

Em 2017, nosso grupo de pesquisa publicou um trabalho utilizando um espectrômetro portátil de infravermelho próximo para identificação de manchas de sangue em diferentes substratos, onde diferentes modelos de classificação específicos para cada substrato foram construídos separadamente (PEREIRA et al., 2017).

Dando continuidade a esse estudo, e sabendo que em um cenário criminal é necessário o uso de metodologias abrangentes, o presente trabalho teve o objetivo de criar uma abordagem unificada e robusta capaz de identificar manchas de sangue humano em pisos com diferentes composições, sendo cinco tipos de cerâmicas e quatro tipos de porcelanatos, com colorações e rugosidades diversas.

1.2 OBJETIVO GERAL

Desenvolver um método analítico para identificação de manchas de sangue humano em diferentes substratos (pisos) utilizando um espectrômetro no infravermelho próximo (NIR) ultra portátil e técnicas quimiométricas.

1.2.1 Objetivos específicos

I. Avaliar a eficácia da técnica de Análise de Componentes Principais (PCA) aplicada aos dados de infravermelho próximo, para avaliação de manchas de sangue humano entre outras substâncias com coloração similar ao sangue.
II. Avaliar a eficácia das técnicas de reconhecimento de padrão supervisionados tais como: Modelagem Independente por Analogia de Classes (SIMCA) e Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) aplicadas aos dados de infravermelho próximo, para identificação e classificação de manchas de sangue humano entre outras substâncias de aspecto similar.
III. Aplicar a metodologia desenvolvida na primeira etapa do trabalho para desenvolver modelos hierárquicos capazes de identificar sangue humano depositados em diferentes tipos de pisos.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo está apresentada uma breve fundamentação teórica dos temas necessários à compreensão e desenvolvimento deste trabalho. Inicialmente, ressalta-se a importância da Química Analítica para as aplicações da Química Forense com ênfase na identificação de amostras de fluidos corporais, mais especificamente de sangue humano. Em seguida serão descritos conceitos de Espectroscopia na região do Infravermelho e Quimiometria. Por fim, serão apresentados estudos recentes para identificação de vestígios em possíveis contextos criminais.

2.1 QUÍMICA FORENSE

Em 1832, um químico britânico chamado James Marsh usou pela primeira vez um experimento para demonstrar a presença de arsênico em tecidos humanos diante de um júri. Esse foi o ponto de partida para o uso da química como ferramenta para investigação criminal, fazendo surgir poucos anos depois estudos na área de toxicologia forense que, posteriormente, faria parte de uma disciplina maior chamada Química Forense (ELKINS, 2019).

Ao longo dos anos, com avanço da ciência, foram desenvolvidas novas técnicas e métodos analíticos, que proporcionaram melhorias significativas na aplicação da química como ferramenta auxiliar nas investigações por parte da polícia científica (SIEGEL, 2016). Atualmente, muitos laboratórios de criminalística utilizam no seu dia a dia equipamentos para cromatografia, espectrometria de massas, espectroscopia no infravermelho, espectrofotometria no ultravioleta-visível, voltametria, e inúmeras outras técnicas para analisar diferentes materiais (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

É evidente que a Química Analítica está intimamente ligada ao desenvolvimento da Química Forense, fornecendo metodologias, princípios e técnicas para auxiliar o perito criminal na busca pela verdade. Para alcançar esse objetivo, as evidências são elementos essenciais. Elas podem incluir qualquer tipo de material físico encontrado na cena do crime, como itens do cotidiano: produtos químicos, tecidos, fibras, cabelos, vidros, impressões digitais, documentos, tintas, corantes, drogas, fluidos corporais, entre outros (ELKINS, 2019).

Dentre as evidências citadas, os vestígios de fluidos corporais são um dos mais complexos, pois são de natureza heterogênea, podendo ser descobertos em uma variedade de contextos criminais, assim como em diversas superfícies (MURO et al., 2016). Os fluidos corporais também têm grande importância porque podem ser uma fonte de evidência da

presença de um indivíduo na cena de crime através da identificação de DNA (VIRKLER; LEDNEV, 2009a). Um dos mais importantes fluidos corporais comumente encontrado em investigações, especialmente em se tratando de crimes violentos, é o sangue.

2.1.1 Vestígios de Sangue

O sangue é um dos vestígios mais valiosos que pode ser recuperado em uma cena de crime (BREMMER et al., 2012). Ele é capaz de fornecer diferentes tipos de informações que podem ajudar a elucidar os acontecimentos. O formato, a dispersão, a quantidade e a posição das manchas de sangue encontradas podem indicar a forma como o delito foi cometido, e facilitar a reconstrução da cena do crime (SHARMA; KUMAR, 2018).

Em investigações criminais, quanto mais tempo se leva para desvendar a verdade, significa que mais tempo o autor do delito estará impune. Dessa forma, é imprescindível que as evidências sejam coletadas e identificadas o mais cedo possível (BEVEL; M. GARDNER, 2008). Atualmente, os vestígios que aparentam ser sangue são identificados visualmente ou com o auxílio de testes colorimétricos e em seguida são levados para os laboratórios de criminalística onde testes confirmatórios são realizados.

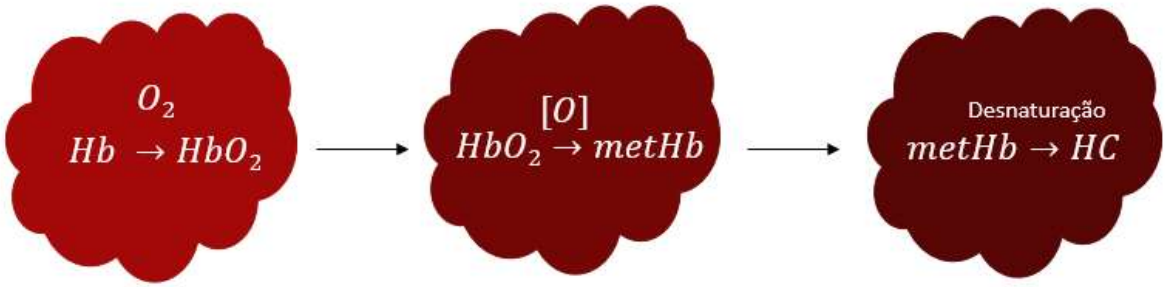
Em geral, esses testes colorimétricos envolvem o uso de reagentes como benzidina, Kastle-Meyer, fenolftaleína e o luminol que irão promover uma reação catalítica (S. DE MARTINIS; F. DE OLIVEIRA, 2016). Apesar de serem testes sensíveis, eles são pouco específicos, sendo alguns deles suscetíveis a resultados falsos positivos com peroxidases de plantas, metais (cobre, ferro e outros), e produtos à base de cloro (MISTEK; LEDNEV, 2015).

Vale ressaltar ainda que os vestígios de sangue são bastante complexos, devido a natureza de sua composição. O sangue é um tecido conjuntivo fluido, em que 55% do seu volume é constituído por plasma sanguíneo onde nele estão suspensas uma mistura complexa de células, proteínas e enzimas. Ele é responsável por diversas funções no corpo como o transporte de oxigênio, nutrientes e hormônios. Apesar de 95% do plasma sanguíneo ser formado por água, ele ainda possui em sua composição eletrólitos e proteínas. Os outros 45% do volume do sangue são ocupados por diferentes tipos de células, como hemácias, leucócitos, plaquetas e eritrócitos (REECE *et al*., 2015).

Quando o sangue deixa de circular no corpo e entra em contato com alguma superfície, ele começa a sofrer um processo de oxidação completa, em que as moléculas de hemoglobina (Hb) presentes nos eritrócitos reagem com o oxigênio formando a oxi-hemoglobina (HbO_2) que se transforma em seguida em meta-hemoglobina (metHb) que sofrerá um processo de

desnaturação ao hemicromo (HC), conforme pode ser exemplificado na Figura 1 (ZADORA; MENZYK, 2018). O ferro presente na Hb encontra-se no estado de oxidação Fe^{2+}, é ele que irá se ligar ao oxigênio nas moléculas de HbO_2 e, após essa ligação, ~~e~~ a molécula de O_2 se torna um ânion superóxido. Quando ocorre a auto-oxidação de HbO_2 em metHb, fora do corpo humano, a ligação com o O_2 não é mais possível e o estado de oxidação do ferro se apresentará na forma de Fe^{3+} (BREMMER et al., 2012).

Figura 1 – Processo de degradação da hemoglobina ao sair do corpo.



Fonte: Adaptado de ZADORA; MENZYK, (2018).

Esse processo de transformação da oxi-hemoglobina em meta-hemoglobina e, posteriormente, em hemicromo, ocasiona a mudança de coloração do sangue de uma tonalidade vermelha para marrom escuro permite que métodos ópticos sejam utilizados para identificação de sangue (CADD et al., 2018). Além disso, outras informações provenientes dos demais componentes presentes no sangue (proteínas, água, e lipídeos) também podem ser visualizadas usando métodos espectroscópicos como a Espectroscopia no Infravermelho (BREMMER et al., 2012).

2.2 ESPECTROSCOPIA NO INFRAVERMELHO (IR)

A espectroscopia de absorção no infravermelho é fundamentada na interação da radiação eletromagnética com a matéria de modo que a energia absorvida pelas moléculas provoca transições vibracionais e rotacionais. Como esses efeitos dependem primeiramente da massa dos átomos e da força da ligação ou ligações químicas que os conectam, pode-se dizer que os espectros vibracionais de uma espécie química dependem da sua natureza intrínseca e constituem uma espécie de impressão digital de uma substância (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

A região espectral da radiação no infravermelho compreende as faixas do espectro eletromagnético com comprimento de onda entre 780 nm e e 10^7 nm. Essa faixa pode ser subdividida em infravermelho próximo (NIR, do inglês *Near Infrared*), médio (MIR, do inglês *Middle Infrared*) e distante (FIR, do inglês *Far Infrared*). A região MIR se estende entre 2500 nm e 50000 nm, enquanto a região do NIR cobre a faixa entre 780 nm e 2500 nm e a do FIR a faixa entre 50.000 nm e 10^7 nm (SKOOG; HOLLER; CROUCH, 2018).

A radiação eletromagnética é composta por um campo elétrico e um campo magnético que oscilam perpendicularmente entre si e em relação à direção de sua propagação. Para que ocorra absorção da radiação na região do infravermelho é necessário que o campo elétrico gerado seja capaz de interagir com uma molécula ocasionando uma alteração no seu momento dipolo e que essa energia da radiação seja igual a diferença de energia entre dois níveis energéticos ($\Delta\upsilon$) da molécula (S. DE MARTINIS; F. DE OLIVEIRA, 2016).

As moléculas podem apresentar diferentes tipos de vibrações e rotações. Sendo que as vibrações moleculares podem ocorrer através do estiramento de ligação, quando há variação na distância entre os átomos no eixo de ligação, ou devido a deformações angulares, quando há variação do ângulo entre duas ligações em um mesmo plano ou fora dele. Além disso, as vibrações de estiramento e deformação podem ser simétricas, quando ocorrem no mesmo sentido, ou assimétricas, quando ocorrem em sentidos opostos (SKOOG; HOLLER; CROUCH, 2018).

Um modelo proposto para auxiliar na compreensão do mecanismo dessas vibrações é o modelo de um oscilador harmônico, muito usado na física para entender um sistema que consiste em duas massas conectadas por uma mola (SKOOG; HOLLER; CROUCH, 2018). Esse modelo é aplicável ao contexto da espectroscopia quando se assume condições em que moléculas vibram no seu estado fundamental (υ=0) e apenas as transições para o primeiro estado excitado (υ = 1) são permitidas. Contudo, esse modelo não se aplica a transições com $\Delta\upsilon > 2$, situação que ocorre quando há sobretons e combinações de vibrações (PASQUINI, 2003). Assim, um outro modelo pode ser utilizado para compreender esse sistema, o modelo do oscilador anarmônico que considera alguns comportamentos não ideais, como a repulsão entre nuvens eletrônicas e a ocorrência das bandas de sobretons e de combinações vibracionais (BURNS; CIURCZAK, 2009). Além disso, esse modelo prevê que a probabilidade dessas transições diminui com o aumento do número quântico vibracional, o que gera espectros de menor intensidade, como é o caso dos espectros de infravermelho próximo (BURNS; CIURCZAK, 2009).

2.2.1 Espectroscopia no Infravermelho Próximo

A região do infravermelho próximo localiza-se logo após o visível, entre os comprimentos de onda de 780 nm e 2500 nm, como já mencionado. Contém quase exclusivamente bandas de absorção que podem ser atribuídas a sobretons e combinações vibracionais, sendo um tipo de radiação com bandas largas, de menor intensidade e muito sobrepostas, o que dificulta as atribuições de banda e diminui fortemente a especificidade da espectroscopia NIR (WORKMAN; WEYER, 2009). Apesar disso, a espectroscopia na região do infravermelho próximo tem diversas vantagens de interesse analítico: a rapidez na resposta espectral e o fato de ser não-destrutiva. Além disso, ela demanda um preparo mínimo ou nenhum de amostra, pois em muitos casos é possível realizar uma análise direta (PASQUINI, 2003). Quando comparados a outros equipamentos que utilizam outras regiões do infravermelho, como o MIR, os equipamentos de NIR são mais robustos e facilmente miniaturizáveis atualmente.

Os espectros são geralmente obtidos por medidas de transmitância, refletância difusa ou transfletância (PASQUINI, 2003). Contudo, deve-se ter atenção aos espectros de amostras sólidas, pois informações relacionadas a espalhamento, reflexão difusa, reflexão especular, brilho da superfície, índice de refração e polarização da luz refletida podem estar sobrepostas à informação vibracional no infravermelho próximo (BURNS; CIURCZAK, 2009).

Devido a esses efeitos os espectros podem apresentar desvios de deslocamento ou de linha de base (efeitos aditivos e multiplicativos) que também podem estar associados à coloração do amostra, tamanhos variáveis de partícula e variabilidade do percurso ou substrato (WORKMAN; WEYER, 2009). Tudo isso torna a interpretação direta dos espectros no NIR um pouco mais complexas.

Uma maneira eficaz para minimizar esses efeitos e extrair informação útil dos dados espectrais é aplicar ferramentas quimiométricas. Além do mais, a disponibilidade de métodos quimiométricos fez com que a percepção de bandas de intensidades mais baixas pudessem ser identificadas e mais bem exploradas (PASQUINI, 2003).

2.3 QUIMIOMETRIA

A Quimiometria é uma área relativamente recente, seu estudo começou por volta da década de 1970 e seu desenvolvimento está diretamente relacionado à popularização do uso de computadores na área da Química. Ela pode ser definida como a aplicação de ferramentas

matemáticas e estatísticas para otimizar experimentos e procedimentos afim de extrair o máximo de informações relevantes em um conjunto de dados de natureza química (OTTO, 2017).

A quimiometria tem muita aplicação em estudos que envolvem uma grande quantidade de informações por meio da coleta de dados, o que é muito comum na aplicação de diversos métodos analíticos. Essa grande quantidade de informações pode dificultar na realização de uma interpretação mais objetiva dos dados, além de levar bastante tempo. Nesse sentido, o uso de técnicas quimiométricas torna mais eficiente a extração e interpretação de informações analíticas (KUMAR; SHARMA, 2018).

Uma vez que os dados sejam coletados e que se tenha o interesse de aplicar técnicas quimiométricas a eles, alguns procedimentos devem ser realizados, como organizá-los no formato de uma matriz de dados **X**, de modo que cada coluna *j* da matriz se refira a uma variável, e cada linha *i* esteja associada a uma amostra. Assim, a matriz de dados **X** será representada com um total de *I* linhas (amostras) e *J* colunas (variáveis) e as ferramentas matemáticas e estatísticas poderão ser aplicadas (FERREIRA, 2015).

É importante salientar que a primeira etapa no tratamento quimiométrico dos dados deve ser realizar um estudo prévio para que informações que não estejam associadas ao analito sejam removidas. Isso implica, muitas vezes, no uso de pré-processamentos nos dados de modo que problemas associados, por exemplo, a ruídos ou dispersões da radiação sejam eliminados (BEEBE; PELL; SEASHOLTZ, 1998).

2.3.1 Técnicas de Pré-Processamento

O objetivo das técnicas de pré-processamento é reduzir a interferência de informações que não são de interesse. Essas interferências podem estar presentes na forma de ruídos ou de deslocamentos aditivos e multiplicativos na linha de base. Os ruídos são variações que ocorrem nas medições e que não estão correlacionadas com qualquer medições referentes aos dados (BRO et al., 2008).

A etapa de pré-processamento é muito importante e deve ser feita de maneira cuidadosa, visto que ela pode ter uma influência positiva (eliminando ruídos, por exemplo) ou negativa (distorcendo informações de interesse) no conjunto de dados (FERREIRA, 2015).

As ferramentas utilizadas nessa etapa, podem ser aplicadas de duas maneiras diferentes: a) nas amostras, quando aplicadas em uma amostra (linha) por vez ao longo de todas as suas

variáveis, ou b) nas variáveis (colunas), quando aplicadas em uma variável por vez em todas as amostras (BEEBE; PELL; SEASHOLTZ, 1998).

As principais técnicas de pré-processamento nas amostras são normalização, suavização, correções de linha de base e derivadas, enquanto os métodos mais usados de pré-processamento nas variáveis são centragem na média e autoescalonamento (BEEBE; PELL; SEASHOLTZ, 1998).

2.3.1.1 Centragem na Média

A centragem na média é uma operação em que, inicialmente, calcula-se o valor médio de cada coluna $\bar{x}_j$ da matriz **X**, conforme a Equação 1, e em seguida, subtrai-se o valor médio de cada um dos valores da coluna x_{ij} obtendo-se o valor centrado na média x_{ij}^{CM}, conforme a Equação 2. Essa técnica muda a origem do modelo, e é muito usada no cálculo de Análise de Componentes Principais (PCA) (GEMPERLINE, 2006) .

$$\bar{x}_j = \frac{1}{I}\sum_{i=1}^{I} x_{ij} \tag{1}$$

$$x_{ij}^{CM} = x_{ij} - \bar{x}_j \tag{2}$$

Após esse pré-processamento a estrutura dos dados não sofre nenhuma alteração, o único resultado que ocorre é a translação do eixo para o valor médio de cada um deles (FERREIRA, 2015).

2.3.1.2 Padronização Normal de Sinal (SNV)

O pré-processamento por Padronização Normal de Sinal (SNV, do inglês *Standard Normal Variate*) corrige efeitos aditivos e multiplicativos que são, geralmente, consequências de interferências por espalhamento de radiação e de tamanho de partícula sólida. Essa correção se dá através de uma normalização, em que o espectro é centralizado na média e em seguida é dividido pelo desvio padrão dos valores de intensidade espectral de cada espectro (FEARN *et al*., 2009).

2.3.1.3 Suavização

A Suavização é uma ferramenta muito utilizada para reduzir componentes aleatórias dos dados, principalmente de espectros e cromatogramas, também chamados "ruídos". Existem diferentes métodos de suavização, mas geralmente eles utilizam uma "janela" de pontos no espectro para calcular, por exemplo, um valor médio dentre esses pontos, o ponto central daquele trecho é então substituído pela média. E, esse processo é repetido ao longo de todo o espectro e, por fim, obtém-se um espectro com uma aparência mais suave (BEEBE; PELL; SEASHOLTZ, 1998). Uma das técnicas mais comuns é a suavização de Savitzky-Golay que funciona ajustando uma função polinomial aos dados nesse intervalo. O aumento da "janela" de pontos implica em um aumento no filtro de ruídos e, consequentemente, em uma maior suavização do espectro (GEMPERLINE, 2006).

2.3.1.4 Derivadas

As técnicas de derivada são muito usadas para corrigir flutuações de linha de base. A primeira derivada é muito útil para corrigir efeitos aditivos, isso se deve ao fato de que a primeira derivada de qualquer constante é zero. Caso a linha de base apresente uma inclinação, usa-se a segunda derivada para corrigir esse efeito (GEMPERLINE, 2006; RINNAN; BERG; ENGELSEN, 2009).

A transformação por derivada é linear e as curvas retém os aspectos quantitativos dos sinais (BRERETON, 2003), contudo os picos sofrem deformações porque a derivada de um ponto qualquer em uma curva é a inclinação da reta tangente a esse ponto. Como a primeira derivada é igual a zero no centro do pico, essa é uma boa maneira de identificar com precisão sua posição. Enquanto na segunda derivada, o ponto de máxima absorbância no espectro original torna-se o ponto de máximo negativo (FERREIRA, 2015). No entanto, deve-se destacar que o uso da derivada acaba amplificando os ruídos e para minimizar essa amplificação dos ruídos, o ideal é que uma etapa de suavização seja realizada anteriormente (BRERETON, 2003).

O método mais utilizado para o cálculo da derivada é o Método de Savitzky-Golay que possui uma etapa de suavização e utiliza um filtro de média móvel onde os ajustes são feitos a partir do cálculo de um polinômio, também é utilizada uma "janela" móvel de pontos e o ponto no centro da janela é substituído pelo valor estimado pelo polinômio (BEEBE; PELL; SEASHOLTZ, 1998; FERREIRA, 2015).

2.3.2 Análise de Componentes Principais (PCA)

Apesar da Quimiometria ter se estabelecido como uma disciplina na década de 1970, o método de Análise de Componentes Principais (PCA) já havia sido introduzido em 1901 e divulgado em 1930 quando foi aplicado por Hotelling na área da psicologia (FERREIRA, 2015). Desde então, a PCA é uma das técnicas mais difundidas na Quimiometria e, provavelmente, umas das mais bem aceitas no meio acadêmico (BRERETON, 2003).

A PCA é uma técnica não supervisionada que pode ser empregada para reconhecimento de padrões que se baseia na variância dos dados para construir um novo espaço de variáveis ortogonais, não-correlacionadas, e que seguem a direção de máxima variabilidade. A intenção é realizar uma compressão dos dados através de combinações lineares das variáveis originais que possuem informações similares (BRO; SMILDE, 2014).

Essa compressão resulta em um novo sistema de coordenadas que é capaz de descrever um maior número de informações em um menor espaço de variáveis. Essas novas variáveis são chamadas de componentes principais (PC) e as novas coordenadas são chamadas de escores. O cosseno dos ângulos entre o eixo das componentes principais e as variáveis originais são chamados de pesos ou *loadings* e indicam a influência da variável original em relação a PC (BEEBE; PELL; SEASHOLTZ, 1998; BRERETON, 2003).

Matematicamente, a PCA pode ser representada como uma decomposição da matriz de dados **X**, conforme pode ser visto na Equação 3. Onde **T** representa a matriz de escores e possui o mesmo número de linhas que a matriz **X**, $\mathbf{P^T}$ é a matriz de *loadings* e possui o mesmo número de colunas da matriz original, e **E** é a matriz de resíduos (BRERETON, 2003; BRO; SMILDE, 2014).

$$\mathbf{X} = \mathbf{TP^T} + \mathbf{E} \tag{3}$$

Essas projeções resultantes da decomposição da matriz **X** são consequência de uma transformação linear. A primeira componente principal é responsável por explicar a maior parte da variabilidade dos dados, enquanto a segunda PC explica a segunda maior variabilidade, e assim sucessivamente, de modo que a última PC irá explicar o menor percentual de variabilidade (FERREIRA, 2015).

A variabilidade explicada por uma componente, é a variância dos escores dessa PC, enquanto a variabilidade total dos dados é a soma das variâncias dos escores de todas as componentes. Pode-se concluir que o número máximo de componentes é o número total de

variáveis do conjunto de dados original, posto que assim tem-se 100% da variabilidade dos dados (BEEBE; PELL; SEASHOLTZ, 1998).

Para realizar a interpretação de um modelo PCA é necessário observar quatro elementos: os dados, os escores, os *loadings* e os resíduos (BRO; SMILDE, 2014). A visualização dos dados originais é de extrema importância para a compreensão e interpretação dos demais elementos da PCA, de modo que essa etapa deve ser realizada logo no início. Ela permite observar se há algum efeito que pode impactar na construção do modelo, como a presença de ruídos, ou efeitos multiplicativos, e se é necessário fazer uso de pré-processamentos, como foi explicado no capítulo anterior.

Os gráficos de escores geralmente são apresentados através de um eixo de uma PC versus o eixo de outra PC. É possível perceber semelhanças e diferenças entre as amostras através da localização dos seus escores. As amostras que são semelhantes têm os escores com coordenadas próximas, enquanto o contrário é observado para amostras distintas (BEEBE; PELL; SEASHOLTZ, 1998).

Esse gráfico deve ser interpretado juntamente com os gráficos de *loadings* que estão diretamente relacionados com as variáveis que exercem maior influência nos escores das componentes principais (BRO; SMILDE, 2014).

Outro elemento importante é a matriz de resíduos **E**. Ela é formada pelas informações que não foram explicadas pelas componentes principais do modelo. O ideal é que seja identificado comportamento aleatório nos resíduos (BRO; SMILDE, 2014).

Uma das dificuldades no uso da Análise de Componentes Principais é determinar o número de componentes principais que são relevantes para interpretar os dados. Isso é o equivalente a definir a variabilidade necessária para descrevê-los (BEEBE; PELL; SEASHOLTZ, 1998). De modo que essa etapa deve ser realizada com cuidado, visto que um modelo construído usando um grande número de PC, ou seja, com grande variabilidade, pode estar sobreajustado e, consequentemente, estar incluindo informações irrelevantes, como resíduos. Enquanto um modelo com um menor número de PC pode estar subajustado e não representar bem todas as informações.

A variância explicada por componente principal diminui à medida que o número de componentes cresce, a maior quantidade de informações relevantes está nas primeiras PC. Dessa forma, faz sentido visualizar inicialmente, as primeiras componentes. Deve-se atentar que ao adicionar mais componentes e incluir mais variações ocorrerá alteração nos resíduos, o que impactará diretamente na identificação de amostras anômalas (BRO; SMILDE, 2014). Assim, é necessário definir o número ideal de PC que irá representar um conjunto de dados

antes de iniciar essa etapa de identificação de amostras com comportamento diferente das demais.

Apesar de algumas anomalias ou *outliers* conseguirem ser, aparentemente, observadas ainda nos dados originais, algumas discrepâncias são difíceis de identificar e só ficam evidentes após a construção do modelo PCA. Observando os escores das componentes principais já é possível realizar uma análise preliminar de amostras que possuem algum comportamento diferente, contudo a importância da PC, ou seja, o percentual de variabilidade que ela explica, deve ser sempre levado em consideração, juntamente com a análise dos resíduos daquela amostra (BRO; SMILDE, 2014).

A partir da matriz de resíduos **E** pode-se calcular alguns parâmetros estatísticos que auxiliam na avaliação do comportamento das amostras. Um desses parâmetros é o Q ou Q-Estatístico, que indica se a amostra está bem ajustada ao modelo construído. O cálculo do Q-Estatístico é realizado através da soma quadrática dos elementos e_{ij} de cada linha da matriz **E** para amostra *i*, conforme mostra a Equação 4 (FERREIRA, 2015).

$$Q_i = \sum_{j=1}^{J} (e_{ij})^2 = e_i^T e_i \tag{4}$$

O Q-Estatístico indica o quão bom é o ajuste de uma amostra ao modelo. Amostras com um valor de Q-Estatístico elevado, indicam que há uma grande diferença entre ela e sua projeção no modelo, o que pode sinalizar uma possível anomalia devido ao seu comportamento diferente das demais (FERREIRA, 2015).

Outro parâmetro estatístico que auxilia na avaliação de amostras anômalas foi estudado por Hotelling em 1931, e consiste em uma generalização estatística do teste *t* de Student, de modo que ele pode ser calculado para um modelo multivariado (HOTELLING, 1992), conforme pode ser visto na Equação 5.

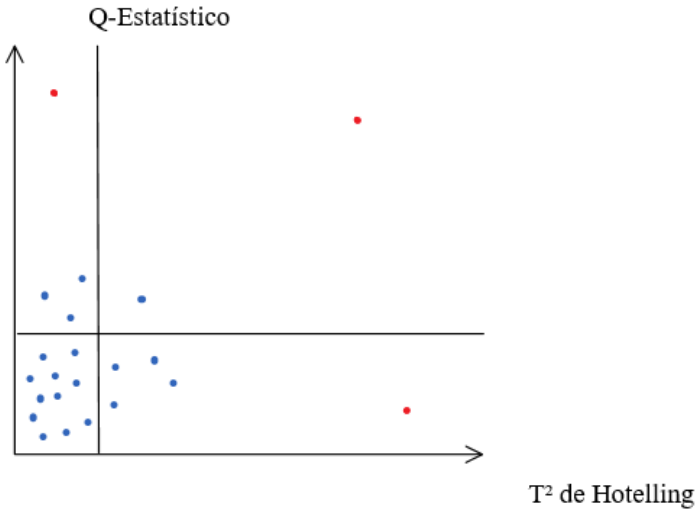
$$T_i^2 = \frac{t_i^T (\boldsymbol{T}^T \boldsymbol{T})^{-1} \boldsymbol{t}_i}{I - 1} \tag{5}$$

Onde T^2 é chamado de T^2 de Hotteling, **T** é a matriz de escores $(I \times R)$ das amostras, $\mathbf{t}_i$ é um vetor $(R \times 1)$ dos escores da i-ésima amostra. A partir da estimativa do T^2 de Hotteling podem ser calculados seus limites para 95% de confiança que servem para identificar uma possível amostra anômala (BRO; SMILDE, 2014).

Comparando-se os dois parâmetros estatísticos, pode-se entender melhor suas diferenças da seguinte maneira: os resíduos Q representam a magnitude da variação restante em cada amostra após a projeção através do modelo. Enquanto os valores de T^2 de Hotelling representam uma medida da variação em cada amostra até o centro do modelo.

Usualmente, ambos os parâmetros são apresentados em um único gráfico, também chamado de Gráfico de Influência, onde o eixo y é o eixo referente ao Q-Estatístico e o eixo x se refere ao valores de T^2 de Hotelling, conforme pode ser visto na Figura 2 (BRO; SMILDE, 2014).

Figura 2 - Exemplo de um Gráfico de Influência.



Fonte: Adaptado de BRO; SMILDE, (2014).

2.3.3 Validação Cruzada

Muitas vezes, os modelos PCA são utilizados para outras finalidades que não envolvem apenas a exploração dos dados, como ocorrem quando se deseja prever amostras externas. Nesses casos, é muito importante que o modelo não esteja sobreajustado às amostras usadas na sua construção, de modo a garantir que ele tenha uma boa eficiência na detecção de amostras fora do conjunto de dados. A alternativa mais utilizada para isso é submeter o modelo aos métodos de validação (BRERETON, 2003). Um dos métodos mais comumente utilizados é o da validação cruzada.

Existem diferentes técnicas de validação cruzada, mas todas elas têm o objetivo de garantir a robustez do modelo. Todas as técnicas se baseiam no mesmo princípio básico: deixar uma parte do conjunto de dados fora do modelo, construir um modelo com os dados restantes

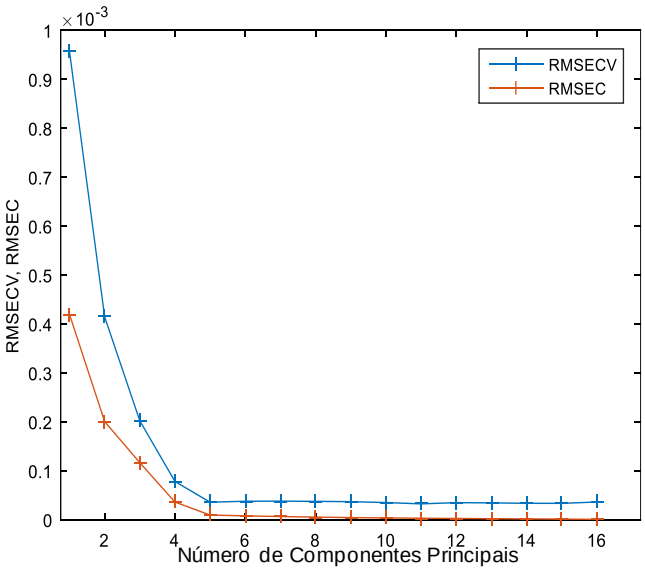
e prever as amostras deixadas de fora (BRO et al., 2008). Esse processo é realizado repetidas vezes, inserindo no conjunto de dados as amostras retiradas, e retirando novas amostras para serem previstas. A partir do conjunto de resíduos, pode-se calcular a Soma Quadrática dos Resíduos de Previsão (PRESS) e a Raiz do Erro Quadrático Médio de Validação Cruzada (RMSECV) conforme as Equações 6 e 7, onde $e_{p_{ij}}$ é o resíduo de predição da amostra *i* e da variável *j* após *r* componentes.

$$PRESS = \sum_{i=1}^{I}\sum_{j=1}^{J}(e_{p}{}_{ij}^{(r)})^2 \tag{6}$$

$$RMSECV = \sqrt{\frac{PRESS}{IJ}} \tag{7}$$

Para auxiliar na escolha do número ótimo de PC, o RMSECV pode ser comparado a Raiz do Erro Quadrático Médio de Calibração (RMSEC) (BRO; SMILDE, 2014). Usualmente, esses parâmetros são visualizados na forma de um gráfico, conforme a Figura 3, em que o eixo x se refere ao número de componentes, e o eixo y aos valores de RMSEC e RMSECV. Idealmente, o número ótimo de PC é indicado quando os valores dos erros são mínimos ou eles não se alteram, além disso também deve se considerar a menor diferença entre os erros de ambos os parâmetros.

Figura 3 - Exemplo de um gráfico de comparação entre RMSECV e RMSEC em função do número de componentes principais usados no modelo.



Fonte: Autora.

2.3.4 Técnicas de Classificação

As técnicas de classificação são também conhecidas como métodos de análise de reconhecimento de padrão supervisionada, elas são chamadas assim porque utilizam um conjunto de amostras com classificação conhecida como forma de treinar o algoritmo (BEEBE; PELL; SEASHOLTZ, 1998). A definição de uma classe pode ser entendida como um grupo de objetos que possui uma ou mais propriedades em comum (OLIVERI, 2017).

As amostras são classificadas de acordo com essas propriedades de interesse usando medições relacionadas a elas, a partir disso uma regra de classificação é determinada. O conjunto de dados usado para realizar essa identificação é chamado de conjunto de treinamento (GEMPERLINE, 2006).

Existem diferentes técnicas que podem ser utilizadas com essa finalidade, elas são geralmente agrupadas em duas categorias de acordo com a abordagem utilizada para realizar esse reconhecimento de padrões: os métodos discriminantes, que se baseiam essencialmente na diferença entre amostras pertencentes a classes diferentes e os métodos de modelagem de classe, que utilizam as similaridades entre as amostras de uma mesma classe para distingui-las das demais (MARINI, 2013).

Essas diferenças entre os métodos fornecem perspectivas diferentes na forma de classificar as amostras. Ao prever uma amostra de teste em um método discriminante, ela será classificada como pertencente a uma das classes modeladas, enquanto uma amostra prevista por um método de modelagem de classes pode ser classificada em uma única classe, em mais de uma classe ou em nenhuma das classes (MARINI, 2013).

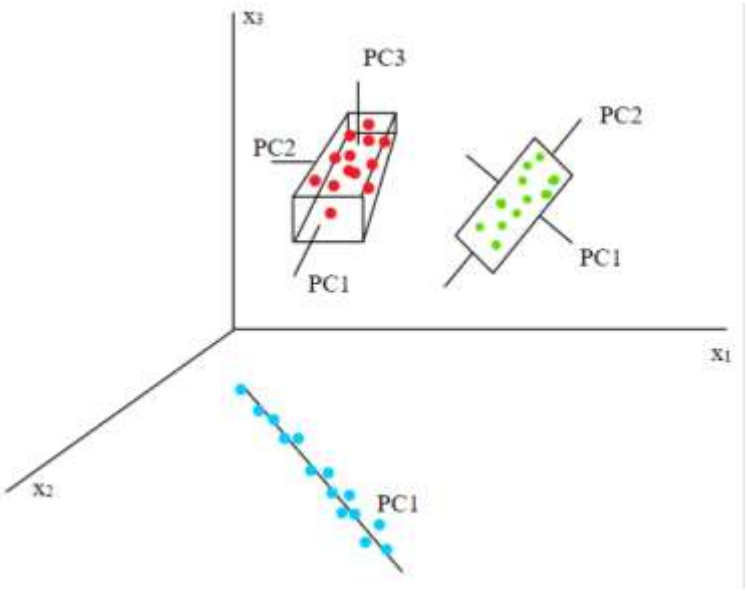
2.3.4.1 Modelagem Independente por Analogia de Classes (SIMCA)

Um dos métodos mais conhecidos de modelagem de classes individuais é a Modelagem Independente por Analogia de Classes (SIMCA). Ele foi um dos primeiros métodos de classificação utilizados, e é considerado uma das técnicas mais simples de classificação. O SIMCA se baseia na comparação entre a variância das amostras e a variância da classe (GEMPERLINE, 2006).

Nesse método utiliza-se modelos de PCA para reduzir a dimensionalidade e quantificar a variabilidade dos dados. Os modelos PCA são ajustados individualmente para cada classe, de modo que o número de componentes principais pode variar para cada uma delas (FERREIRA, 2015). Os modelos SIMCA são definidos pelo intervalo de valor dos escores das componentes

principais selecionadas, e seus limites no espaço podem corresponder a retângulos (2 PC), paralelepípedos (3 PC) ou hiperparalelepípedos (mais de 3 PC) (OLIVERI; DOWNEY, 2013). Os diferentes modelos de acordo com o número de PC podem ser vistos na Figura 4.

Figura 4 - Modelos SIMCA para classe com diferentes números de componentes principais.



Fonte: Adaptado de OTTO, (2017)

As hipercaixas são definidas a partir da matriz de resíduos **E**. Compara-se a variância residual de uma amostra do conjunto de treinamento com a variância residual média da classe em questão. Essa comparação é realizada através de um teste F que também calcula um limite superior para a variância residual daquelas amostras que pertencem à classe, com o resultado sendo um conjunto de probabilidades de associação à classe para cada amostra . Esse método foi proposto por Wold e determina os limites da hipercaixa quantitativamente, usando o desvio-padrão dos escores em cada PC. Mas esses limites também podem ser calculados através da distância de Mahalanobis ou do T^2 de Hotteling (FERREIRA, 2015) .

O SIMCA classifica as amostras externas comparando a variância residual da amostra de teste projetada no modelo com a variância residual média das amostras que compõem a classe. Essa comparação fornece uma medida direta da semelhança de uma amostra com uma classe específica, indicando uma distância ao modelo. É esse o conceito utilizado para verificar se as amostras desconhecidas pertencem ou não a cada uma das classes (MARINI, 2013).

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

O método de Análise Discriminante por Mínimos Quadrados Parciais (PLS-DA) é um método de classificação que utiliza parâmetros obtidos através do algoritmo de regressão por Mínimos Quadrados Parciais (PLS) para discriminar as classes. Para isso, o modelo busca correlacionar a matriz de dados **X**, com uma matriz **Y** formada a partir de *g* colunas (classes), e do número *n* linhas (amostras) (BALLABIO; CONSONNI, 2013).

A matriz **Y** é preenchida numericamente por valores binários: onde 1 é atribuído às amostras da classe e 0 é atribuído às amostras que não pertencem a classe (FERREIRA, 2015). Quando apenas duas classes são avaliadas, tem se um vetor no lugar da matriz **Y**, e a técnica utilizada é a PLS1. O PLS-DA pode ser utilizado para classificar três ou mais classes, nesses casos usa-se a abordagem PLS2 (OLIVERI, 2017).

As variáveis latentes são obtidas através da combinação linear das variáveis originais de modo a maximizar a covariância entre os escores. Matematicamente, a matriz **X** e a matriz **Y** são decompostas em escores e *loadings* de acordo com as Equações 8 e 9, onde **T** e **U** são os escores da matriz **X** e **Y**, **P** e **Q** são os *loadings* da matriz **X** e **Y**, e $\mathbf{E_X}$ e $\mathbf{E_Y}$ são os resíduos da matriz **X** e **Y**, respectivamente (MARINI, 2013).

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E_X} \tag{8}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{E_Y} \tag{9}$$

A dependência linear entre os escores da matriz **X** e **Y** pode ser determinada através da Equação 10, onde **C** é uma matriz diagonal formada pelos pesos.

$$\mathbf{U} = \mathbf{T}\mathbf{C} \tag{10}$$

Relacionando as Equações 9 e 10, pode-se calcular a matriz dos coeficientes de regressão $\widetilde{\boldsymbol{B}}$, através da Equação 11, que permite a predição dos valores da matriz dependente $\widetilde{\boldsymbol{Y}}$ para amostras desconhecidas a partir das variáveis independentes $\mathbf{X_{pred}}$ fornecidas.

$$\widehat{\mathbf{Y}} = \mathbf{X_{pred}}\,\widehat{\mathbf{B}} \tag{11}$$

Os valores previstos de $\widehat{\mathbf{Y}}$ não serão mais números binários ou inteiros, mas assumirão valores próximos a 1 quando a amostra prevista pertencer a classe, e valores próximos a zero quando a amostra não pertencer (FERREIRA, 2015).

2.3.5 Validação e Figuras de Mérito para Modelos de Classificação

A validação de um modelo é um procedimento que assegura a confiabilidade da metodologia proposta para que ela possa ser aplicada garantindo que o modelo atende às expectativas da finalidade para qual ele foi criado (LÓPEZ; CALLAO; RUISÁNCHEZ, 2015). A maneira mais conservadora de realizar essa validação é testar o modelo com amostras externas, que são independentes do conjunto de treinamento, a essas amostras dá-se o nome de conjunto de teste (BALLABIO; GRISONI; TODESCHINI, 2018).

Esse tipo de validação, geralmente ocorre quando há um conjunto de dados relativamente grande, de modo que ele possa ser separado entre os conjuntos de treinamento e de teste sem que haja prejuízo para a construção do modelo (BRERETON, 2003). As amostras de teste são projetadas no modelo de classificação e as classes preditas são então confrontadas com a natureza das classes conhecidas.

Para modelos qualitativos, a resposta fornecida pelo método é, geralmente, binária como positivo ou negativo, o que faz com que a melhor maneira de avalia-lo seja através de parâmetros métricos, também chamados de figuras de mérito (LÓPEZ; CALLAO; RUISÁNCHEZ, 2015). Como a natureza das amostras preditas é conhecida, pode-se confrontá-los com os resultados obtidos pelos modelos e, assim, tem-se os chamados falsos positivos comuns (FPC), falsos negativos (FN), verdadeiros positivos (VP) e verdadeiros negativos (VN).

Os falsos positivos são amostras que não pertencem a classe, mas foram classificadas como se pertencessem a ela, enquanto falsos negativos são amostras que pertencem a classe, mas foram classificadas erroneamente. Os verdadeiros positivos e negativos são as amostras classificadas corretamente (OLIVERI; DOWNEY, 2013). A partir desses resultados, uma matriz de confusão pode ser construída, conforme pode ser visto na Figura 5, que representa um exemplo de uma matriz de confusão para duas classes.

Figura 5 - Representação genérica de uma matriz de confusão para duas classes, em que VP_1, FP_1, FN_1, VN_1 se referem aos resultados obtidos em relação a classe 1.

		Classe Verdadeira	
		1	**2**
Classe Predita	**1**	VP_1	FP_1
	2	FN_1	VN_1

Fonte: Adaptado de BALLABIO; GRISONI; TODESCHINI, (2018)

A partir dos valores de VP, FP, FN e VN pode-se calcular os principais indicadores utilizados para definir a validação de um método, como a Sensibilidade (Sn) que representa a capacidade do modelo de evitar os falsos negativos, identificando corretamente as amostras pertencentes a classe (BALLABIO; GRISONI; TODESCHINI, 2018), ela pode ser calculada conforme a Equação 11.

Outra figura de mérito é a Especificidade (Sp) que estima a habilidade do modelo em rejeitar amostras de outras classes (falsos positivos), identificando corretamente as amostras negativas, conforme a Equação 12.

$$\mathrm{Sn} = \frac{\mathrm{VP}}{\mathrm{VP} + \mathrm{FN}} \tag{11}$$

$$\mathrm{Sp} = \frac{\mathrm{VN}}{\mathrm{VN} + \mathrm{FP}} \tag{12}$$

2.3.6 Modelos Hierárquicos

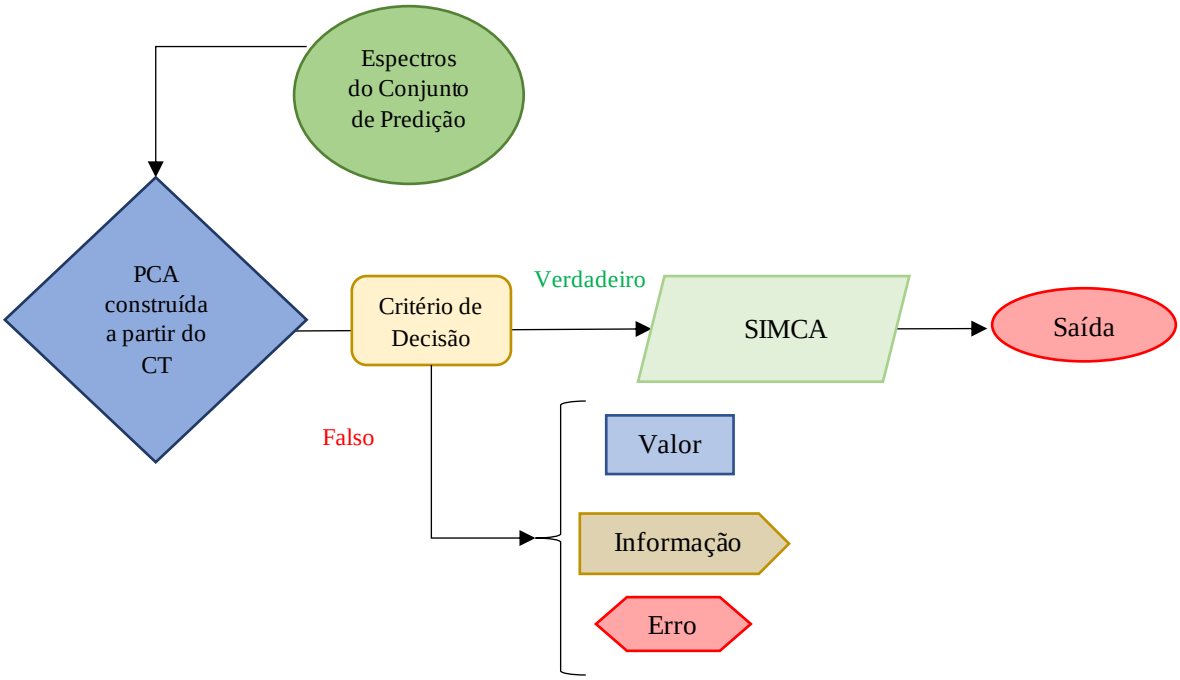
Os Modelos Hierárquicos (MH) são uma forma de executar modelos quimiométricos em sequência através de operações em um formato de árvore de decisão. Quando aplicados com objetivo de classificação, eles conseguem resolver problemas nos quais apenas uma das técnicas de classificação como PLS-DA ou SIMCA não são suficientes ou apropriadas para o problema, de modo que mais de um método pode ser utilizado como critério de classificação (EIGENVECTOR, 2020; PEREIRA, 2019).

Cada etapa do MH funciona como uma espécie de filtro que irá separar os dados que atendem um determinado critério, dos dados que não atendem a esse critério e não deverão passar para etapa seguinte. Essas etapas também podem ser chamadas de nós, e esses nós podem assumir diferentes formas, como um modelo PCA, um modelo PLS-DA, SIMCA, entre outros. Assim, o Modelo Hierárquico funciona através da combinação de diferentes técnicas quimiométricas. Os modelos em cada nó definem qual dos caminhos deve ser selecionado para uma amostra através de uma condição específica que a amostra deve obedecer dentro do modelo utilizado. O nó final é a última etapa de cada ramificação e define qual deve ser a resposta de saída, que funciona como uma "previsão" para qualquer amostra que chegar a essa etapa final (EIGENVECTOR, 2020; PEREIRA, 2019).

A Figura 6 mostra um fluxograma de como funcionaria um modelo hierárquico assumindo um problema de classificação com dois nós de decisão, em que o primeiro deles é uma PCA, cujo critério de decisão é um valor determinado de T^2 de Hotteling, escolhido a partir da avaliação dos dados do conjunto de treinamento. E o segundo é um modelo SIMCA onde

não foi estabelecido um critério, porque ele próprio foi utilizado como regra de decisão e sua resposta final, ou *output* são as probabilidades de os dados do conjunto de teste pertencerem a uma ou mais classes do modelo SIMCA.

Figura 6 - Fluxograma genérico para um Modelo Hierárquico com 2 nós, sendo o primeiro uma PCA e o segundo um modelo SIMCA.



Fonte: Autora.

2.4 MÉTODOS ESPECTROSCÓPICOS E QUIMIOMÉTRICOS PARA IDENTIFICAÇÃO DE SANGUE

Como foi visto ao longo dos tópicos anteriores, as evidências forenses são de extrema complexidade, pois têm naturezas diversas e podem ser encontradas em diferentes condições e superfícies. De modo que identificá-las e classificá-las, apesar de ser de extrema importância, não é uma tarefa trivial. Assim, é imprescindível que avanços tecnológicos na área da Química Analítica possam ser aplicados na Ciência Forense.

Os métodos de espectroscopia vibracional como as técnicas envolvendo Infravermelho e Raman estão cada vez mais sendo utilizados para finalidades forenses, devido a uma série de vantagens envolvendo suas aplicações: como sua capacidade de caracterizar materiais de diferentes composições, o fato deles não serem destrutivos (o que permite análises complementares nas evidências) e de já existirem equipamentos portáteis que podem ser levados ao próprio cenário do crime (SILVA; BRAZ; PIMENTEL, 2019).

Para identificação de vestígios de sangue, as vantagens citadas são de grande interesse, visto que o sangue é um material complexo e que após a identificação da amostra pode ser necessário realizar análises posteriores como a de DNA. A espectroscopia Raman vem sendo bastante estudada com essa finalidade. Diferentes estudos realizados por Vikler e Lednev mostram o potencial da técnica não apenas para identificá-las (VIRKLER; LEDNEV, 2010), como também para distinguir o sangue de diferentes espécies (VIRKLER; LEDNEV, 2009b) quando associadas ao uso de ferramentas quimiométricas como PCA, e ALS.

Um estudo mais recente, realizado por Takamura e colaboradores (2019) desenvolveu modelos de deconvolução (MCR-ALS) para espectros de sangue obtidos através de um equipamento Raman com feixe de excitação na região do NIR (785 nm). Eles coletaram espectros de amostras de sangue depositadas em placas de vidro e envelhecidas durante três a quatro meses sob controle de três temperaturas diferentes (30°C, 24°C e 16°C) e desenvolveram um algoritmo que além de identificar o sangue, permitiu a construção de modelos cinéticos que descreveram o comportamento de auto-oxidação da oxihemoglobina em metahemoglobina e hemicromo para cada uma das temperaturas (TAKAMURA et al., 2019). Esse estudo representa um avanço para a datação de amostras envelhecidas de sangue, pois apresentou a influência da temperatura na degradação de diferentes componentes do sangue. Essa informação é importante em investigações criminais pois mostra que se uma evidência for submetida a diferentes temperaturas isso pode influenciar na estimativa de seu tempo de deposição.

Para avaliar se a espectroscopia Raman era capaz de distinguir amostras de sangue e outros fluidos corporais de substâncias usualmente atribuídas como falsos positivos por testes presuntivos, Rosenblatt e colaboradores (2019) testaram espectros Raman de 24 substâncias em um modelo de classificação de fluidos corporais construído usando a técnica SVM-DA. Eles estabeleceram um critério de 70% de probabilidade como limite de classificação e obtiveram 100% de classificação correta para amostras dos 24 falsos positivos (ROSENBLATT et al., 2019). Esse estudo é bastante interessante, pois utilizou 12 substâncias que são falsos positivos em testes comumente utilizados pela polícia para identificar manchas de sangue como Luminol, Hemastix e Kastle-Meyer. Contudo, esse estudo não foi realizado para diferentes substratos e as amostras foram depositadas em placas de vidro revestidas com folhas de alumínio.

Quando se compara o uso das técnicas de Raman e IR para fins de identificação de vestígios de sangue, é necessário considerar as vantagens e desvantagens de cada uma delas. Por exemplo, é sabido que a utilização do Raman é menos suscetível a influência da presença de água no sangue, o que representa uma vantagem em relação ao IR, contudo os espectros de Raman podem sofrer interferências devido a presença de luz externa e fenômenos de

fluorescências (TAKAMURA et al., 2019), o que dificulta na realização de análises direta em contextos criminais.

Um outro trabalho foi realizado com objetivo de identificar e datar manchas de sangue, dessa vez utilizando Espectroscopia no NIR por Edelman e colaboradores (2012). Eles obtiveram espectros de manchas de sangue e de outras substâncias com coloração avermelhada na região do NIR (800–2778 nm) sob tecidos de algodão e construíram um modelo de regressão (PLS). Os autores obtiveram 100% de sensibilidade e especificidade para classificação de amostras de sangue, e em relação a datação, a raiz do erro quadrático médio de predição foi de 8,9% para manchas com mais de um mês. A Tabela 1 mostra um compilado dos picos de absorção na região do NIR na qual os autores se basearam para caracterizar os componentes sanguíneos (EDELMAN; LEEUWEN; AALDERS, 2012).

Tabela 1 - Lista de picos de absorção no NIR de alguns componentes sanguíneos.

λ (nm)	**Componente**	**Atribuição**
930	Oxihemoglobina	Terceiro sobretom de -CH e vibrações de estiramento de $-CH_2$
970	Água	Combinação de vibrações de alongamento simétricas e assimétricas H-O-H
1454	Água	Combinação de vibrações de alongamento simétricas e assimétricas H-O-H
1690	Hemoglobina, Albumina, Globulina	Primeiro sobretom da vibração de estiramento -CH
1740	Hemoglobina, Albumina, Globulina	Primeiro sobretom da banda em 3477 nm
1940	Água	Combinação de vibrações deformações e de estiramento assimétricas H-O-H
2056	Hemoglobina, Albumina, Globulina	Combinações de Amida A e amida II ou outras combinações
2170	Hemoglobina, Albumina, Globulina	Combinações de Amida B e Amida II ou outros sobretons de Amida II
2290	Hemoglobina, Albumina, Globulina	Combinações de estiramento e deformações de -CH
2350	Hemoglobina, Albumina, Globulina	Combinações de estiramento e deformações de -CH

Fonte: Adaptado de EDELMAN et al., (2012)

É importante destacar que trabalho de Edelman e colaboradores (2012) é um dos estudos pioneiros no uso de espectroscopia NIR para identificação de sangue e apresentou-se adequado para datação de amostras de sangue em curto prazo. Os modelos de regressão usando PLS foram criados individualmente para cada fundo colorido, o que também dificulta seu uso em uma situação real em contexto criminal.

Outros trabalhos foram desenvolvidos utilizando espectroscopia no infravermelho para fins de identificação de sangue, como o estudo de Li e colaboradores (2018) que utilizaram espectros de Reflectância Difusa na região do visível e do infravermelho próximo para classificar o sangue de cinco espécies (cães, cabras, macacos, rato e humanos). As leituras dos espectros foram feitas em 1200 amostras depositadas em tubos de ensaio com 5 ml de sangue na presença de um anticoagulante. Foi utilizado um algoritmo de classificação de Rede Neural Artificial (ANN) para construir os modelos e 20% do conjunto de dados foi utilizado como conjunto de treinamento. Os resultados obtidos mostraram melhores resultados em termos de robustez e precisão para reconhecimento de espécies utilizando uma combinação dos espectros no visível e no NIR na construção do modelo em relação aos modelos construídos utilizando apenas espectros obtidos no vísivel ou no NIR (LI et al., 2018).

Espectros de Reflectância Total Atenuada no Infravermelho (ATR FT-IR) foram utilizados por Lin e Colaboradores (2017) para estimar a idade de manchas de sangue em ambiente interno e externo por até 107 dias. As amostras foram depositadas em placas de vidro e foram definidos dezenove pontos de coleta no tempo (espassados entre 0 e 107 dias). Não foram realizadas análises *in loco*, e cada amostra de mancha de sangue foi coletada em um tubo *Eppendorf* e misturada com 10 µL de solução salina antes de serem analisadas no equipamento ATR, os espectros foram obtidos em triplicatas e usados para a construção de um modelo de regressão usando PLS. O modelo construído mostrou-se capaz de estimar a idade das manchas de sangue, mas a abordagem foi mais útil para a amostras coletadas em longo prazo (7-85 dias), independentemente de elas estarem em um ambiente interno ou externo (LIN et al., 2017).

O estudo desenvolvido por Mistek e colaboradores (2019) também utilizou a técnica de ATR FT-IR para obter espectros de sangue e criar um modelo de classificação usando PLS-DA com a finalidade de determinar o fenótipo de diferentes doadores das amostras sangue. Eles utilizaram um Algoritmo Genético (GA) nos dados pré-processados para selecionar as variáveis mais significativas para a classificação dos grupos (sexo e etnia) e em seguida construíram os dois modelos usando PLS-DA. Eles obtiveram 92% de precisão nas previsões do sexo e na etnia do doador. Apesar de ser um estudo pioneiro ao buscar diferenciar fenótipos de pessoas através dos espectros de sangue, deve-se ressaltar que a validação do modelo foi realizada com um conjunto de teste que representou apenas 13% do conjunto dos dados, totalizando apenas 4 amostras provenientes de 4 doadores (1 homem caucasiano, 1 homem afroamericano, 1 mulher caucasiana e 1 mulher hispânica), o que levanta um questionamento sobre a representatividade dessa validação e torna esse estudo questionável.

Sharma e colaboradores (2020) realizaram um estudo no qual a espectroscopia ATR FT-IR foi usada para discriminar o sangue menstrual de sangue periférico, fluido vaginal, fluido seminal e outras substâncias não biológicas. As amostras foram depositadas em diferentes superfícies (vidro, plástico, frama, madeira, algodão, papel, etc) e secas por 3 dias, após a sua recuperação e obtenção dos espectros, foram construídos modelos de classificação usando PCA-LDA, e também um modelo de regressão usando PLS-R. O modelo de PCA-LDA e PLS-DA apresentaram taxa de classificação correta de 100% de precisão para sangue menstrual e periférico. Esse estudo é interessante por incluir um outro tipo de amostra de sangue que pode estar envolvido em cenas de crime envolvendo violência sexual. Apesar deles terem utilizados diferentes substratos, os espectros não foram obtidos diretamente nos substratos, de modo que essa análise não poderia ser realizada no local do crime como pode acontecer através do uso de equipamentos portáteis.

O desenvolvimento tecnológico de equipamentos portáteis e de imagens hiperespectrais utilizando espectroscopia também despertaram o interesse nessas técnicas para aplicações forenses e diversos estudos foram realizados usando esses equipamentos para identificar vestígios de sangue. Um exemplo é o trabalho de Edelman e colaboradores (2015) que empregaram imagens de câmeras hiperespectrais de reflectância na região do visível (400-720 nm) para identificar manchas de sangue em diferentes tipos de tecido (EDELMAN; VAN LEEUWEN; AALDERS, 2015).

Malegori e colaboradores (2020) usaram uma câmera hiperespectral na região do NIR (HSI-NIR) para identificar diferentes fluidos biológicos: sangue, urina e sêmen. Os fluidos foram depositados em quatro substratos (papel, algodão escuro, algoão branco e jeans) e deixados secar por um dia. As imagens hiperespectrais obtidas foram utilizadas na construção, inicialmente, de uma PCA para fins exploratórios, e a partir dela foi possível destacar os comprimentos de onda mais informativos, fornecendo uma banda espectral característica para cada fluido biológico, o que permitiu propor um processamento de dados direto e minimizar a contribuição de fundo. Eles concluíram que o HSI-NIR conseguiu detectar cada evidência simulada em todos os substratos considerados.

Em 2017, nosso grupo de pesquisa publicou um trabalho desenvolvido por Pereira e colaboradores (2017) utilizando um espectrofotômetro portátil MicroNir 1700 (da Viavi) para obter espectros de manchas de sangue depositadas em diferentes substratos (vidro, cerâmica, porcelanato e metal). Foram avaliadas diversas técnicas quimiométricas para construir os modelos de classificação: SIMCA, GA-LDA, Algoritmo de Projeções Sucessivas (SPA-LDA), e PLS-DA. Esses modelos foram construídos individualmente para cada tipo de substrato do

conjunto de dados, e os melhores resultados foram obtidos com os modelos construídos usando PLS-DA e GA-LDA cujos valores de sensibilidade e especificidade foram iguais a 1 para todos os substratos (PEREIRA et al., 2017). O trabalho tem o mérito de evidenciar o potencial do equipamento portátil para identificação de sangue humano, mas como mencionado, os modelos foram construídos separadamente para cada substrato e, portanto, não são suficientemente robustos para utilização em casos reais.

Todos os estudos descritos acima, associaram o uso de técnicas espectroscópicas a ferramentas quimiométricas para identificar, classificar ou datar as amostras de sangue e outros fluidos corporais. Contudo, as substância não-biológicas, conhecidas como falso-positivos podem ter naturezas diversas o que pode dificultar a construção dos modelos de classificação. Como é o caso das técnicas de Análise Discriminantes, que são métodos rígidos, projetados para atribuir cada amostra a uma classe modelada. De modo que, esses modelos não podem identificar amostras que não pertençam as classes previamente modeladas, o que leva a classificações incorretas. Essa propriedade é indesejável na análise forense, pois uma variedade de materiais suspeitos pode ser encontrada em investigações criminais, incluindo alimentos e detergentes químicos, além de fluidos corporais. Dessa forma, o ideal é excluir o mais rápido possível as amostras de substâncias que podem levar a resultados errôneos e que são muito diferentes de sangue, por isso há um grande interesse na construção de modelos que funcionem por etapas, sendo a primeira delas capaz de separar vestígios de sangue dos demais.

Um estudo desenvolvido por Takamura e colaboradores (2018) descreve um método de classificação diferente dos citados acima para discriminar espectros de amostras de fluidos corporais e de outras substâncias não-biológicas. Apesar deles terem criado um modelo de classificação usando PLS-DA, eles adicionaram uma etapa posterior que eles chamaram de árvore de classificação dicotômica com agrupamento hierárquico. Os espectros dos fluidos corporais (sangue, saliva, sêmen, urina e suor) foram obtidos usando ATR FT-IR, e o modelo PLS-DA foi construído para 5 classes. Os autores então incorporaram a Análise de Agrupamento por Métodos Hierárquicos (HCA) aos escores do PLS-DA, o que resultou em um dendrograma com quatro agrupamentos, e de acordo com ele foi construído um novo esquema de classificação que consistiu em uma árvore de classificação dicotômica com testes de Q-Estatístico no final de cada extremidade o que permitiu a identificação de amostras com comportamento diferente das classes modeladas pelo PLS-DA e tornou a metodologia mais robusta para exclusão de falsos-positivos (TAKAMURA et al., 2018).

O estudo desenvolvido durante o meu período de mestrado aqui descrito visa preencher uma lacuna que não foi preenchida pelos trabalhos anteriormente citados, a criação de um

modelo para identificação de vestígios de sangue um pouco mais compatível com a realidade forense encontrada em cenas de crime. Os trabalhos que identificaram vestígios de sangue em diferentes substratos, construíram modelos individuais para cada um deles. Enquanto o trabalho aqui descrito buscou desenvolver uma modelagem que incluísse a variabilidade dos diferentes substratos para a construção de um modelo mais robusto, capaz de ser usado para identificar amostras de sangue em diferentes pisos.

3 METODOLOGIA

Neste capítulo é detalhada a metodologia empregada para o desenvolvimento deste trabalho. Inicialmente, apresenta-se o preparo das amostras, seguido da forma como os espectros no infravermelho próximo foram obtidos e do método utilizado para a construção dos modelos de classificação, e por último como foi desenvolvido um protocolo de avaliação do substrato. Ressalta-se que os espectros utilizados como conjunto de dados no presente trabalho foram cedidos pelos autores envolvidos no trabalho de Pereira *et al* (2017).

3.1 PREPARO DAS AMOSTRAS

As amostras foram coletadas seguindo o procedimento operacional descrito por Pereira *et al* (2017). O sangue humano (SH) foi coletado por um colaborador qualificado que introduziu agulhas individuais e esterilizadas diretamente nos capilares dos dedos de 22 doadores (12 mulheres e 10 homens). As amostras de sangue animal (SA) foram adquiridas e fornecidas pelo hospital veterinário da Universidade Federal Rural de Pernambuco (UFRPE) sendo essas provenientes de cinco animais (3 carneiros e 2 cavalos). Além disso, foram utilizados sete produtos comerciais de coloração avermelhada que podem ser visualmente confundidas com manchas sangue, foram elas: vinagre balsâmico, batom vermelho, vinho tinto, pimenta, geleia, ketchup e molho shoyo. Essas substâncias serão chamadas nesse trabalho de falsos positivos comuns (FPC) e foram selecionadas dentre os produtos testados por Edelman *et al*., (2012). As amostras foram depositadas diretamente sobre os substratos ou utilizando uma pipeta de Pasteur (2 gotas) em 9 diferentes tipos de substrato: 5 cerâmicas e 4 porcelanatos. A aparência dos substratos pode ser observada na Figura 7.

Figura 7 - Fotografia dos nove substratos utilizados neste trabalho.

Fonte: Autora.

Não houve controle do volume de sangue depositado em cada mancha, e o número de manchas depositadas em cada substrato foi diferente devido a quantidade de material coletada de cada doador.

3.2 AQUISIÇÃO DOS ESPECTROS

Os espectros de infravermelho próximo que compõem a base de dados utilizada foram obtidos através do uso de um espectrômetro MicroNir 1700 fabricado pela empresa VIAVI. Este equipamento tem dimensões de 45 mm de diâmetro e 42 mm de altura e pesa 60g, e possui um filtro linear acoplado a um arranjo de detectores de Índio Gálio e Arsênio (InGaAs), e duas lâmpadas de Tungstênio que são a fonte de radiação infravermelha. A faixa espectral na qual ele opera inicia em 908 nm e vai até 1676 nm, a resolução óptica é de 1,25% do comprimento de onda central, ou seja, em um comprimento de 1000 nm a resolução é de 12,5 nm. Além disso, o dispositivo possui uma interface USB que permite a transferência das medidas obtidas em 64 varreduras e tempo de integração de 5 milissegundos.

O procedimento operacional realizado para a aquisição dos espectros seguiu uma metodologia similar a descrita por (PEREIRA et al., 2017). Antes da coleta dos espectros, foi esperado um tempo de seis dias de secagem a temperatura em torno 25°C. Para contemplar a máxima variabilidade de cada amostra, foram adquiridos quatro espectros em diferentes posições em cada mancha. Também foram coletados 3 espectros diretamente da superfície dos substratos. O conjunto de dados foi constituído pelos 986 espectros coletados, dentre eles 506 espectros de sangue humano, 200 espectros de sangue animal e 280 espectros de falsos positivos comuns. A Tabela 2 detalha a quantidade de espectros de cada uma das classes por substrato contida no banco de dados.

Tabela 2 - Quantidade de espectros coletados em cada substrato.

Substratos	**Espectros**		
	Sangue Humano	**Sangue Animal**	**Falsos Positivos Comuns**
PO1	65	32	-
PO3	60	8	56
PO4	40	32	56
PO5	75	-	-
CE1	69	32	-
CE3	64	32	56
CE4	-	32	-
CE5	67	32	56
CE6	66	-	56
Total	506	200	280

3.3 CONSTRUÇÃO DOS MODELOS

Para realizar a avaliação dos dados espectrais, assim como a construção dos modelos de classificação, utilizou-se o software Matlab (MATLAB® R2010a 7.10.0.499, MathWorks). O PLS Toolbox (Eigenvector Research, Inc) foi utilizado para construir os modelos hierárquicos, assim como os modelos PCA, SIMCA e PLS-DA.

Para facilitar na interpretação dos espectros foram definidos dois conjuntos de classes: o conjunto de doadores, dividido em três classes (SH, SA e FPC) e o conjunto de substratos, dividido em 9 classes (CE1, CE3, CE4, CE5, CE6, PO1, PO3, PO4 e PO5).

3.3.1 Avaliação dos Dados e Pré-Processamento

Os dados originais foram observados inicialmente de acordo com o substrato no qual as manchas foram depositadas. Avaliou-se a necessidade de cortes no início e no final dos espectros, e diferentes pré-processamentos foram utilizados com o objetivo de identificar e minimizar as diferenças significativas entre os substratos. Foram testadas as seguintes técnicas: 1ª e 2ª derivadas com filtro de suavização Savitzky-Golay (SG), variando o número de pontos da janela e com polinômio de segunda ordem; Suavização com filtro SG, variando o número de pontos da janela de 3 a 15 e com polinômio de segunda ordem; SNV e centralização na média.

Após analisar individualmente o conjunto de espectros de cada substrato, foi avaliado o conjunto total de dados. A escolha do melhor pré-processamento foi baseada em inspeção visual dos espectros corrigidos, e pelos gráficos de escores e *loadings* obtidos a partir da Análise de Componentes Principais (PCA). Definido o melhor pré-processamento, a etapa seguinte consistiu na escolha do conjunto de treinamento e do conjunto de predição para a construção do modelo.

3.3.2 Definição dos Conjuntos de Treinamentos e de Predição

Para a construção do modelo e sua posterior validação externa, fez-se necessário dividir o conjunto de dados em dois grupos: o conjunto de treinamento, que foi utilizado para construir os modelos de classificação, e o conjunto de predição, que foi utilizado para ver a capacidade de predição do modelo construído.

A primeira escolha do conjunto de predição levou em consideração os seguintes critérios: uma classe de substratos que só tivesse amostras de sangue humano (PO5), outra classe com apenas amostras de sangue animal (CE4) e por último, uma classe de pisos que

tivesse amostras de falsos positivos comuns (CE6). Essa última foi escolhida também por não possuir espectros de sangue animal, pois essa classe teve o menor número de espectros, e a retirada de mais espectros de SA diminuiria a variabilidade do modelo de classificação. Assim, o primeiro conjunto de predição foi formado por amostras depositadas em 3 substratos totalmente diferentes dos substratos usados no conjunto de treinamento. O conjunto de treinamento foi formado pelas amostras depositadas nos 6 substratos restantes.

Para avaliar a escolha do conjunto de treinamento, foram testadas outras combinações de classes de substratos, considerando as diferenças e semelhanças entre eles. A escolha dessas combinações foi realizada a partir da construção de um modelo de análise de PCA usando todos os espectros e os espectros da superfície dos substratos limpos. Foram observados os escores e *loadings* das classes dos substratos. Dessa forma, foram propostos outros dois conjuntos de treinamento e, consequentemente, outros dois conjuntos de predição.

3.3.3 Construção do Modelo Hierárquico

No desenvolvimento do modelo de classificação, o conjunto de treinamento foi submetido a sucessivas construções de modelos de PCA, variando o número de componentes principais (PC), observando o gráfico de escores, *loadings* e de resíduos. A partir da avaliação desses modelos foi decidida a construção de uma modelagem hierárquica de 2 regras/nós, onde inicialmente seriam removidas as amostras de falsos positivos, e posteriormente seriam classificadas as diferentes origens dos espectros de sangue.

Foram construídos modelos hierárquicos com os espectros considerando a faixa espectral completa e também com a faixa reduzida, sendo que os mesmos conjuntos de treinamento foram utilizados na construção desses modelos.

3.3.4 Elaboração de um Protocolo para Avaliação do Substrato

Com o objetivo de garantir que o modelo construído para identificação e classificação de amostras de sangue possa ser utilizado na prática por peritos criminais, foi desenvolvido um protocolo inicial para avaliação prévia do substrato, de modo a garantir que a variabilidade do substrato esteja representada pelo conjunto de dados do modelo.

Esse protocolo seguiu uma metodologia similar a proposta para a construção dos modelos hierárquicos para identificação do sangue. Nesse caso, construiu-se um modelo hierárquico de nó único referente a um modelo PCA, cuja regra de decisão foi o Q-Estatístico.

4 RESULTADOS E DISCUSSÃO

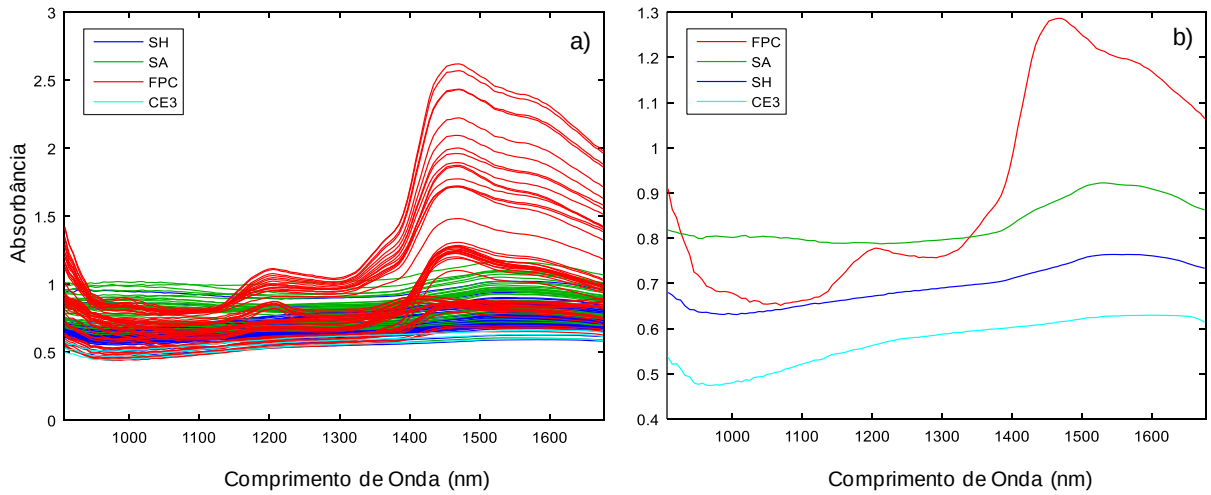
Neste capítulo, são discutidos, inicialmente, os resultados obtidos através da análise espectral e da escolha do pré-processamento, seguida pela definição dos conjuntos de treinamento e de predição, da avaliação dos substratos, da construção dos modelos hierárquicos e, por último, a elaboração de um protocolo de avaliação dos substratos.

4.1 ANÁLISE E PRÉ-PROCESSAMENTO ESPECTRAL

Devido à variabilidade dos substratos, a abordagem inicialmente escolhida, foi avaliar os espectros obtidos para cada substrato individualmente. Foram analisados os espectros originais e a sua média em relação ao conjunto de doadores. Será detalhado nessa primeira parte do trabalho o procedimento realizado para um único substrato, a cerâmica CE3. Contudo, o passo a passo foi o mesmo para os outros oito substratos.

Na Figura 8a são apresentados o gráfico dos espectros originais obtidos para o substrato CE3 onde, as linhas azuis representam os espectros do sangue humano (SH), as linhas verdes são os espectros do sangue animal (SA), as linhas vermelhas os espectros dos falsos positivos comuns (FPC) e as linhas ciano são os espectros do substrato puro (CE3). As médias desses espectros originais podem ser observadas na Figura 8b. No Apêndice A encontram-se todos os gráficos dos espectros originais e das médias dos espectros originais referentes a todos os substratos.

Figura 8 – Comparação entre os diferentes espectros originais de sangue humano (azul), sangue animal (verde) e falsos positivos comuns (vermelho) depositados no substrato CE3 e o substrato puro (ciano). a) Espectros originais b) Média dos espectros originais.

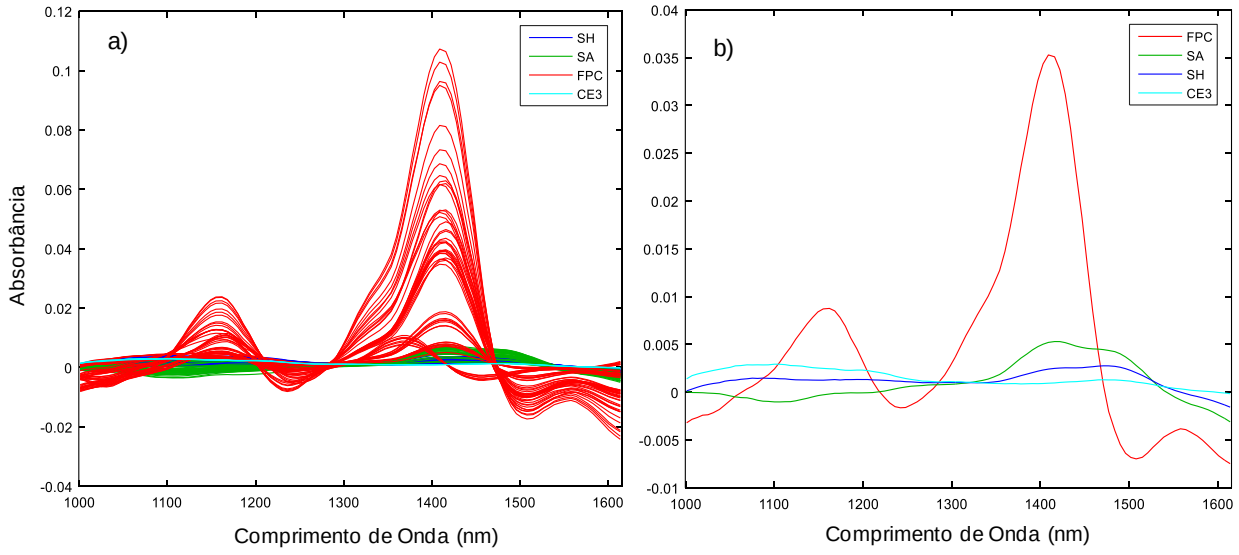


Alguns efeitos indesejados podem ser vistos na Figura 8a, como a presença de ruídos (até aproximadamente 1000 nm), a variação de linha de base e efeitos multiplicativos devido ao espalhamento da radiação. Sabe-se também que ocorrem efeitos de borda nos comprimentos de onda iniciais e finais da faixa espectral monitorada. Assim, para melhor avaliar as informações relevantes, foram cortados 15 pontos no início e 10 pontos no final dos espectros.

Foram testadas diferentes técnicas de pré-processamentos com o objetivo de corrigir os efeitos citados, como SNV, 1ª Derivada e 2ª Derivada com diferentes janelas de suavização. Os espectros originais e pré-processados foram constantemente comparados a fim de avaliar os efeitos das técnicas. Além disso, foram realizadas análises de componentes principais para cada um dos pré-processamentos, e os gráficos de escores foram avaliados com o objetivo de verificar a existência de algum padrão de agrupamento e a presença de *outliers*. Também foram avaliados os *loadings*, para observar quais as variáveis mais influentes nas componentes principais, e os gráficos de resíduos para avaliar a presença de amostras com comportamentos anômalos.

Os melhores resultados na correção dos efeitos indesejados observados nos espectros foram obtidos utilizando a técnica de 1ª Derivada de Savitzky-Golay com 15 pontos de suavização e polinômio de 2º grau. A Figura 9a apresenta os espectros pré-processados referentes ao substrato CE3, em que é possível observar que o ruído espectral e os efeitos de espalhamento da radiação foram minimizados, especialmente para as amostras de sangue.

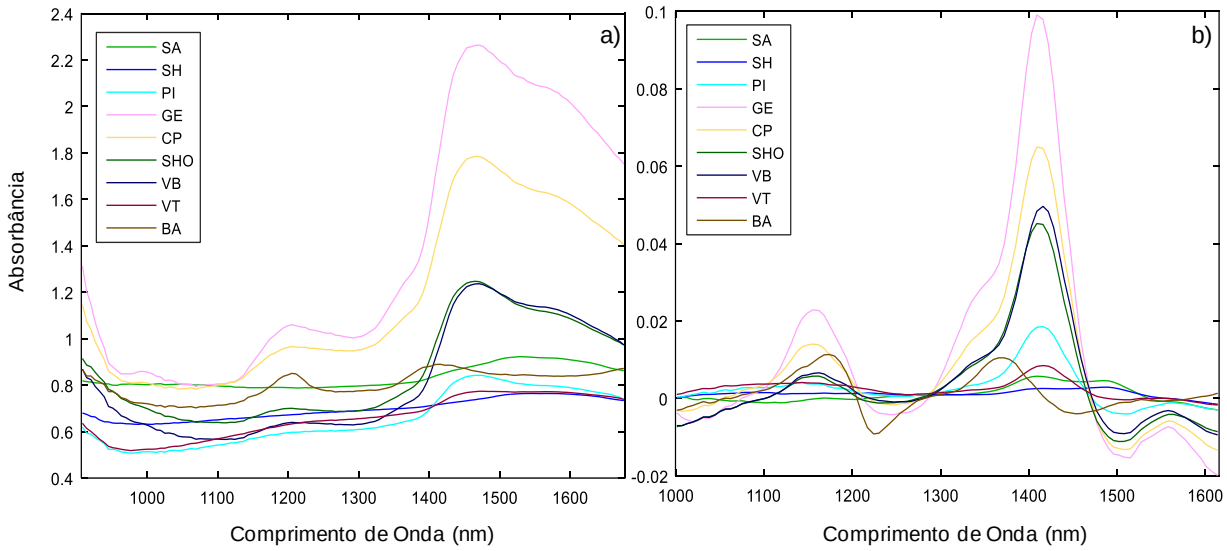
Figura 9 – Comparação entre os diferentes espectros pré-processados com 1ª Derivada de Savitzky-Golay de sangue humano (azul), sangue animal (verde) e falsos positivos comuns (vermelho) depositados no substrato CE3 e o substrato puro (ciano). a) conjunto de dados b) médias dos espectros.



A partir da Figura 9b pode-se constatar uma grande diferença na intensidade dos espectros de FPC em relação aos demais. As regiões de máximo nos espectros originais, aparecem como zero nos espectros pré-processados devido a aplicação do método de 1ª Derivada. Assim, os espectros de FPC apresentam bandas de absorção mais intensas nas regiões entre 1150 nm a 1250 nm, e entre 1400 e 1550 nm. Os demais espectros pré-processados referentes aos outros substratos podem ser visualizados no Apêndice B.

Os espectros de FPC são, na realidade, espectros de sete diferentes substâncias: molho de pimenta (PI), geleia (GE), ketchup (CP), molho de soja (SHO), vinagre balsâmico (VB), vinho tinto (VT) e batom (BA). Para visualizar melhor os espectros de cada FPC, alterou-se a legenda conforme mostra a Figura 10. A Figura 10a apresenta a média dos espectros originais de SH, SA e de cada uma dessas substâncias. Enquanto a Figura 10b apresenta a média dos espectros pré-processados.

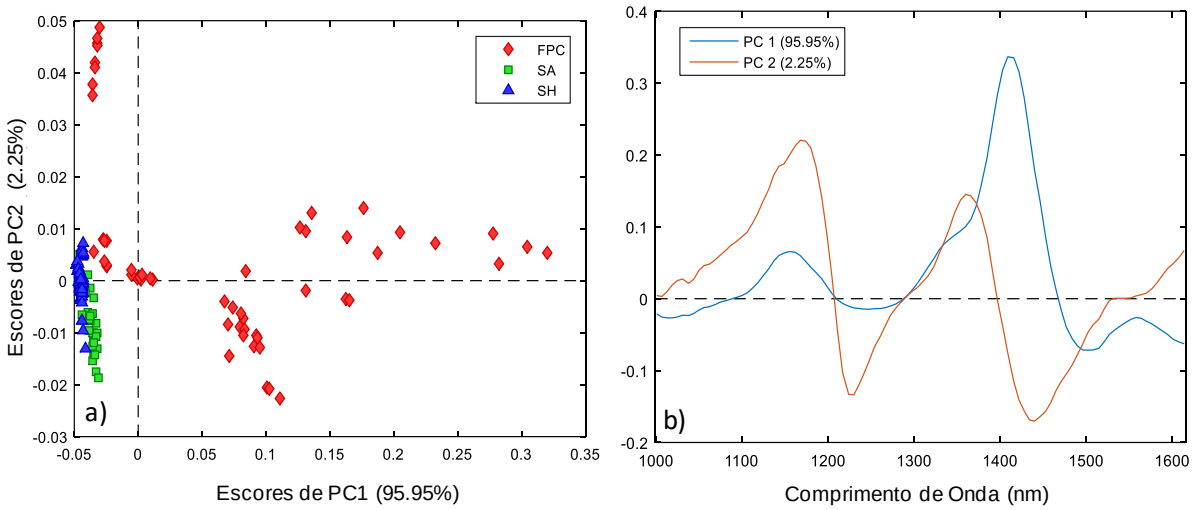
Figura 10 – Comparação entre as médias dos espectros originais e pré-processados de sangue humano (azul), sangue animal (verde), pimenta (anil), geleia (rosa), ketchup (amarelo), molho de soja (verde escuro), vinagre balsâmico (azul marinho), vinho tinto (vinho) e batom (marrom) depositados no substrato CE3. a) média dos espectros originais b) média dos espectros pré-processados.



Observando as Figura 10a e 10b, pode-se constatar que quatro substâncias têm uma maior absorção na região por volta de, aproximadamente, 1460 nm que pode ser atribuída ao primeiro sobretom de estiramento da hidroxila ($-OH$), provavelmente devido a presença de álcool e ácidos carboxílicos (WORKMAN; WEYER, 2009), como o ácido cítrico muito utilizado em alimentos como geleias e ketchup por ter ação antioxidante.

A Figura 11a apresenta os gráficos de escores e a Figura 11b os gráficos de *loadings* referentes as duas primeiras componentes principais da PCA construída com as amostras de SH, SA e FPC depositadas sob o substrato CE3.

Figura 11 – a) Gráfico dos escores e b) *loadings* da PCA para as amostras de sangue humano (azul), sangue animal (verde) e falsos positivos comuns (vermelho) depositados no substrato CE3.



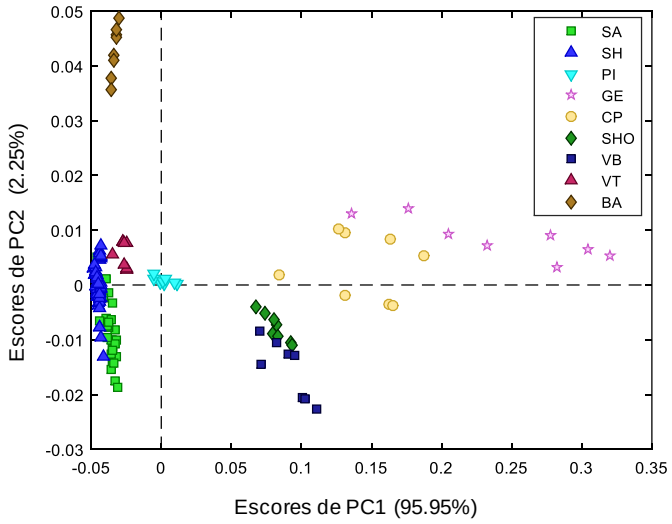
Observa-se na Figura 11a que os escores de PC1 da classe sangue apresentam uma tendência de separação dos escores de FPC, enquanto os escores da PC2 evidenciam a diferença entre um grupo de FPC das demais amostras. PC1 representa 95,95% da variabilidade explicada, enquanto PC2 apresenta apenas 2,25%.

As variáveis que mais influenciaram positivamente a primeira componente foram as regiões entre 1300 e 1500 nm, enquanto, na segunda componente, as variáveis contribuem com um peso parecido ao longo da região, conforme pode ser visto no gráfico dos *loadings* da Figura 10b.

A Figura 12 apresenta o mesmo gráfico de escores da Figura 11a, mas agora com a legenda de cada FPC. Através dessa figura é possível observar que os escores da classe de vinho tinto são os que mais se aproximam dos escores de sangue, e que o agrupamento com escores mais positivos em PC2 são de batom.

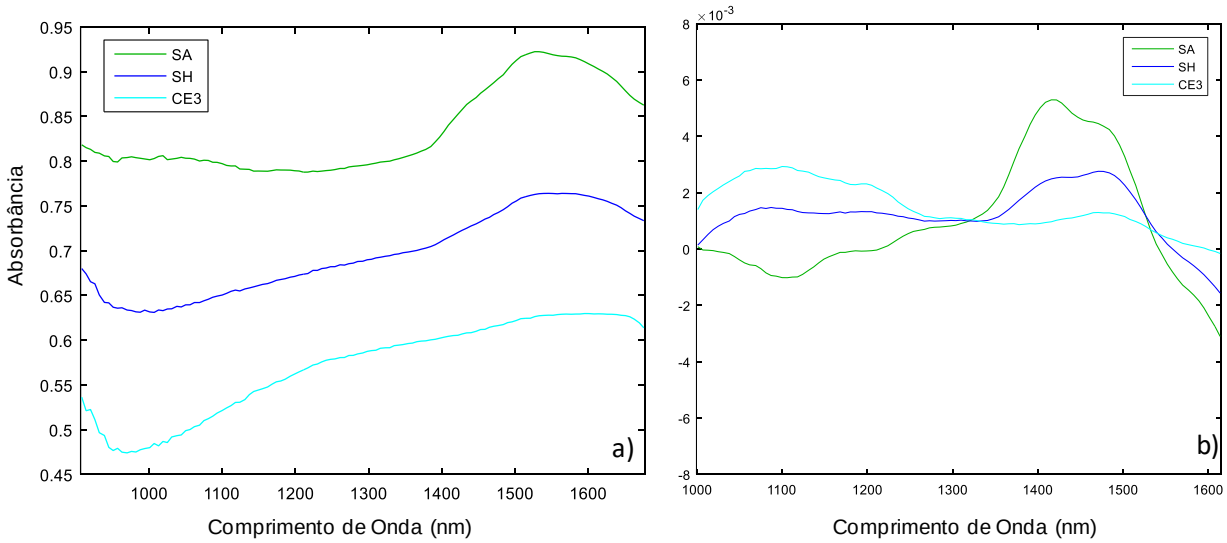
Nas Figuras 11a e 12, é possível perceber que os escores das classes de SH e SA têm coordenadas semelhantes em PC1, enquanto em PC2 os escores de SA apresentam a tendência de serem um pouco mais negativos que os escores de SH. Para visualizar melhor essas diferenças excluiu-se as amostras de FPC. E os espectros originais, pré-processados e dos gráficos obtidos através da PCA foram novamente analisados.

Figura 12 –Gráfico a) de escores e b) *loadings* da PCA para as amostras de SH (azul), SA (verde), PI (anil), GE (rosa), CP (amarelo), SHO (verde escuro), VB (azul marinho), VT (vinho) e BA (marrom) depositados no substrato CE3.



Os espectros originais e pré-processados das amostras de SH, SA e CE3 podem ser visualizados através das Figura 13a e 13b, respectivamente.

Figura 13 – Comparação entre os espectros originais e pré-processados das amostras de sangue humano (azul), sangue animal (verde) e do substrato puro (ciano) a) espectros originais b) espectros pré-processados.

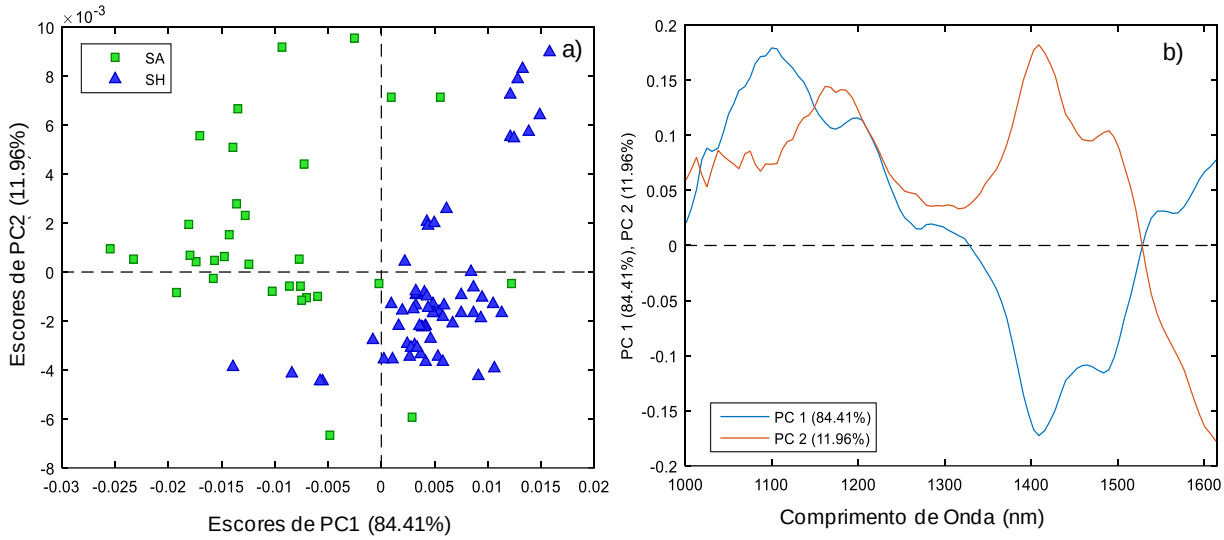


É possível visualizar na Figura 13a que a região dos espectros de sangue que apresenta maior absorbância ocorre entre 1300 nm e 1600 nm, essa também é a região menos influenciada pelos efeitos associados ao substrato. Apesar disso, na Figura 13b, podemos visualizar que ainda há alguma informação relevante em relação ao sangue capaz de auxiliar na distinção entre SH e SA no início do espectro.

Segundo Workman e Weyer (2009), a região entre 1440 e 1485 nm está associada ao primeiro sobretom de estiramento da ligação $O-H$ presente na água, entre 1347 e 1367 nm ocorre a faixa de absorção espectral associada a combinações do segundo sobretom da ligação $C-H$ em grupos metil ($-CH_3$) que pode ser associada a molécula de hemoglobina. Enquanto por volta de 1511 nm está localizada o primeiro sobretom de estiramento da ligação $-NH$ pertencentes a proteínas. Já no início do espectro, em 1007 nm se encontra a banda de estiramento do segundo sobretom da ligação $N-H$. E entre 1191 e 1194 nm ocorre absorção do segundo sobretom de estiramento da ligação $C-H$ (WORKMAN; WEYER, 2009).

Os escores de PC1 versus PC2 das classes de SH e SA podem ser vistos na Figura 14a. Nessa figura, observa-se que há uma tendência de separação entre as duas classes na PC1, em que o conjunto de SH têm coordenadas mais positivas do que o conjunto de SA.

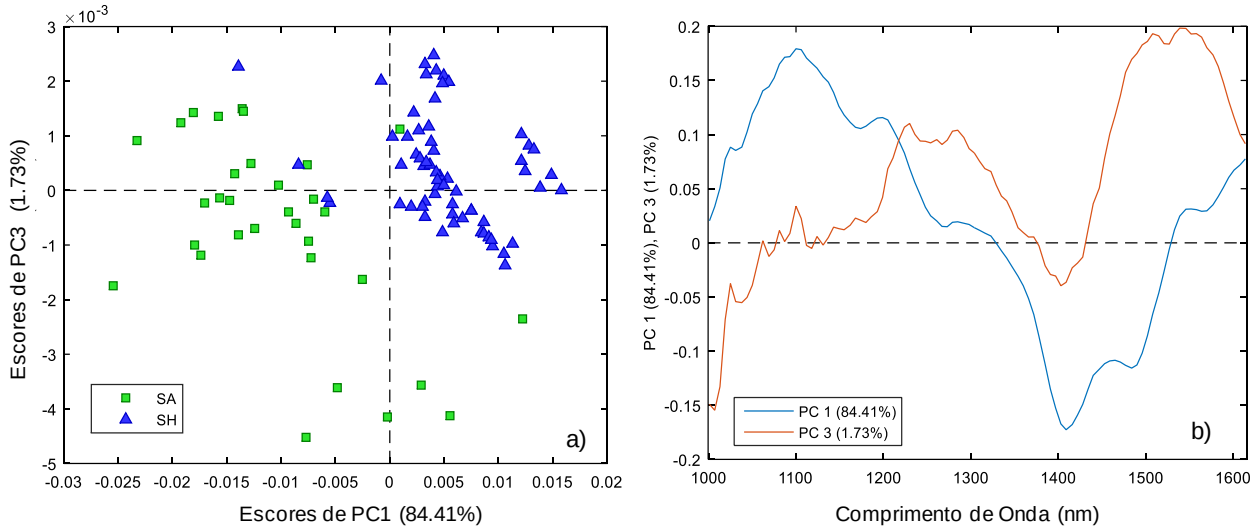
Figura 14 – Gráfico de a) escores e b) *loadings* de PC1 e PC2 das amostras de sangue humano (azul), sangue animal (verde) depositadas sob o substrato CE3.



A primeira componente representa 84,4% da variabilidade dos dados, enquanto a segunda componente representa 11,96%. Olhando os *loadings* dessas componentes apresentados na Figura 14b, conclui-se que as amostras de SA têm influência negativa das variáveis na região entre 1300 e 1500 nm (escores negativos), enquanto as amostras SH tem maiores valores para as variáveis entre 1000 e 1200 nm.

A tendência de separação entre as classes SH e SA em PC1 fica mais evidente quando a terceira componente principal é incluída, conforme pode ser constatado através da Figura 15a que apresenta os escores de PC1 versus PC3. Os *loadings* da Figura 15b mostram que a PC3 é influenciada mais fortemente pelas variáveis na faixa entre 1500 e 1600 nm.

Figura 15 – Gráfico de a) escores e b) *loadings* de PC1 e PC3 das amostras de sangue humano (azul), sangue animal (verde) depositadas sob o substrato CE3.

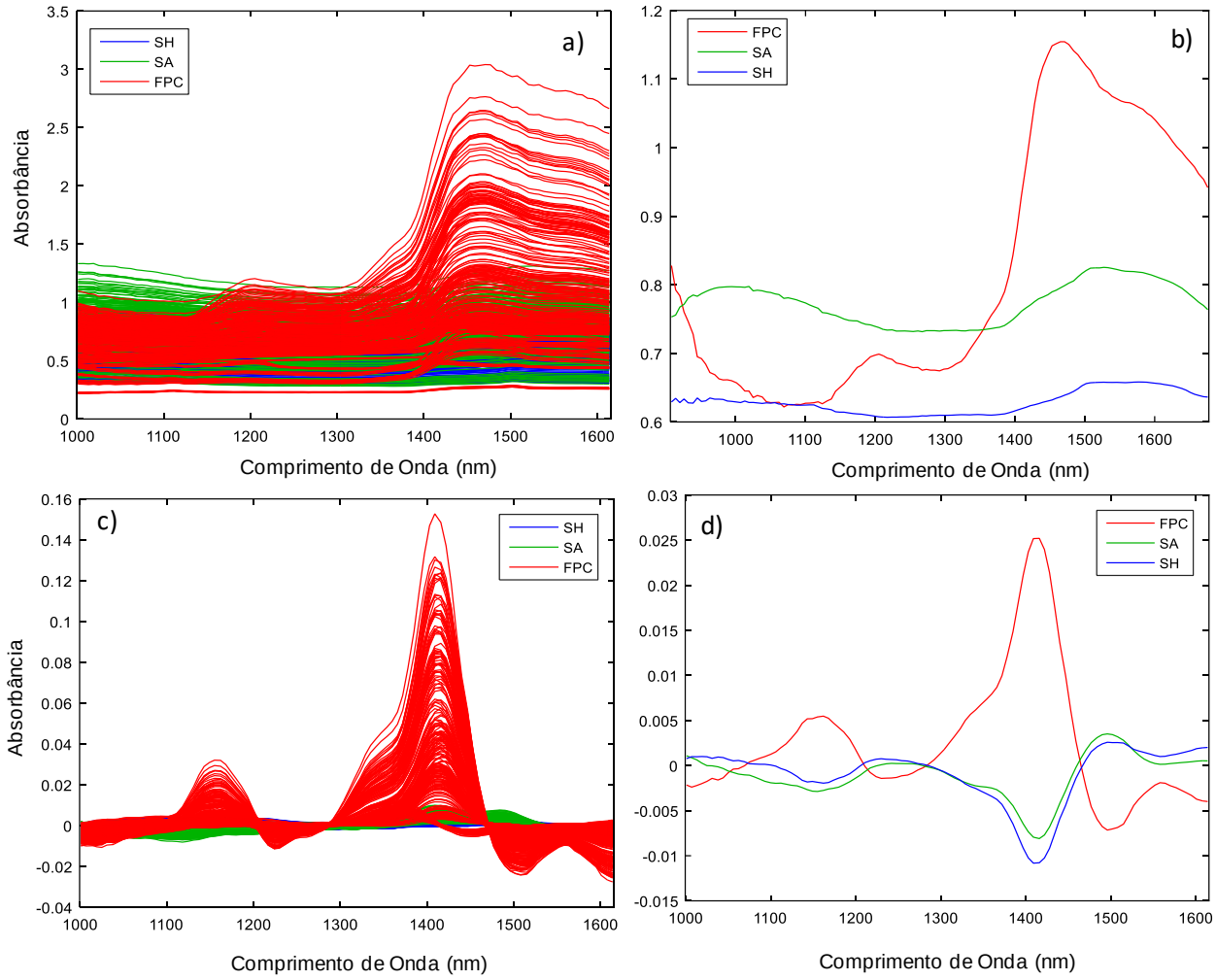


Todos os substratos tiveram seus espectros originais e pré-processados avaliados seguindo o procedimento que foi detalhado para o substrato CE3. O padrão de comportamento dos espectros pré-processados de SH, SA e FPC, assim como o perfil dos escores e *loadings* das PCA mantiveram alguma semelhança para todos os substratos, evidenciando um comportamento para cada tipo de amostra, porém com algumas influências dos substratos. Dessa forma, o passo seguinte foi unir todos os espectros em um único conjunto de dados e construir um modelo único para identificação de sangue humano em diferentes substratos, observando as possíveis diferenças entre os espectros das amostras nos diferentes pisos.

Os espectros originais, a média dos espectros originais, os espectros pré-processados e a média dos espectros pré-processados para o conjunto contendo todos os espectros podem ser visualizados nas Figuras 16a, 16b, 16c e 16d, respectivamente.

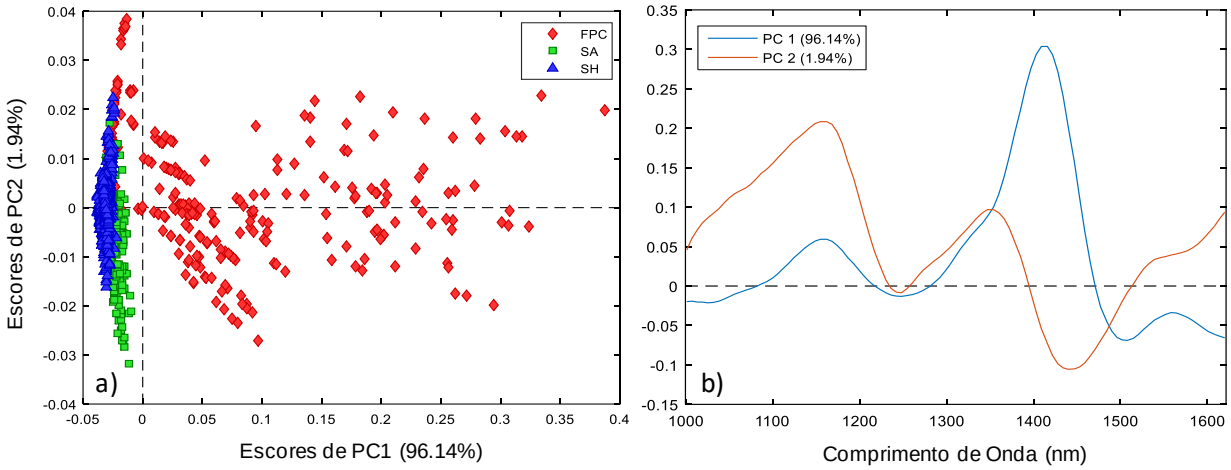
O comportamento dos espectros originais e pré-processados para todo o conjunto de dados é semelhante ao observado e discutido em relação aos espectros do substrato CE3.

Figura 16 – a) Espectros originais, b) média dos espectros originais, c) espectros pré-processados e d) média dos espectros pré-processados de todo o conjunto de dados.



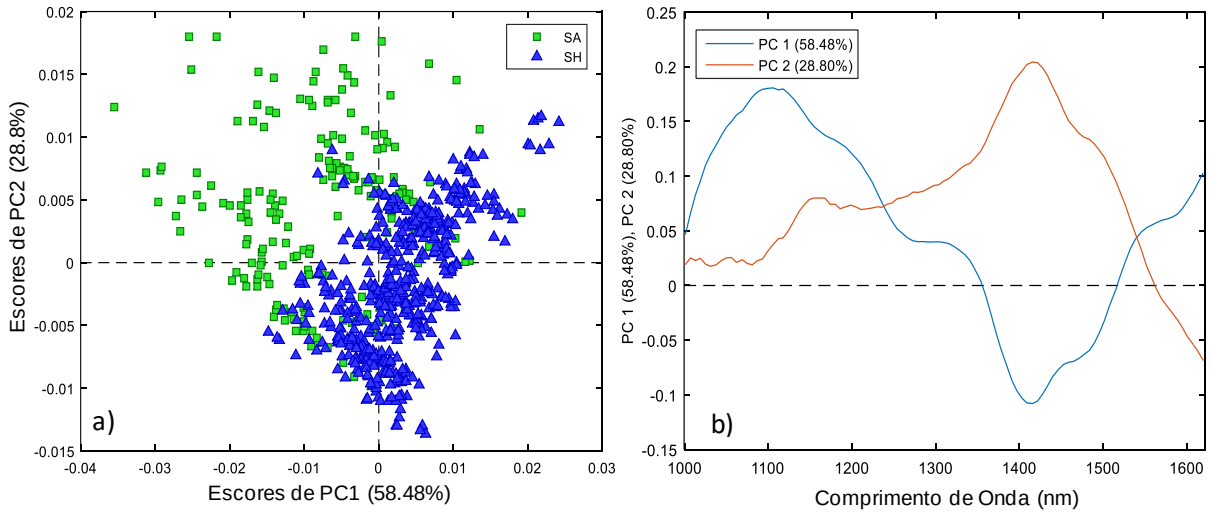
A PCA construída utilizando todo o conjunto de dados apresentou um padrão similar de escores e *loadings*. Os gráficos dos escores de PC1 versus PC2 para todos o conjunto de dados podem ser visto na Figura 17a, e os *loadings* dessas componentes estão na Figura 17b. PC1 explica 96,14% da variabilidade do conjunto de dados, e evidencia a tendência de separação da classe de FPC da classe de sangue, as amostras de FPC que mais se assemelham ao sangue continuam sendo as amostras de vinho tinto (Figura 12). PC2 explica 1,94% da variabilidade, e os *loadings* mostram que as variáveis com maior peso em PC1 estão na faixa entre 1300 e 1500 nm, e em PC2 entre 1000 e 1200 nm.

Figura 17 – Gráfico de a) escores e b) *loadings* de PC1 e PC2 das amostras de sangue humano (azul), sangue animal (verde) e falsos positivos comuns (FPC) de todo o conjunto de dados.



Excluindo-se as amostras de FPC e construindo a PCA utilizando apenas os espectros de sangue (Figura 18), observa-se a mesma tendência de separação apresentadas na Figura 14a. Os escores de SH têm a tendência de serem mais positivo em PC1 (Figura 18a), indicando uma maior contribuição das variáveis na faixa entre 1000 e 1200 nm (Figura 18b), enquanto os escores de SA têm um maior peso das variáveis na região entre 1300 e 1500 nm.

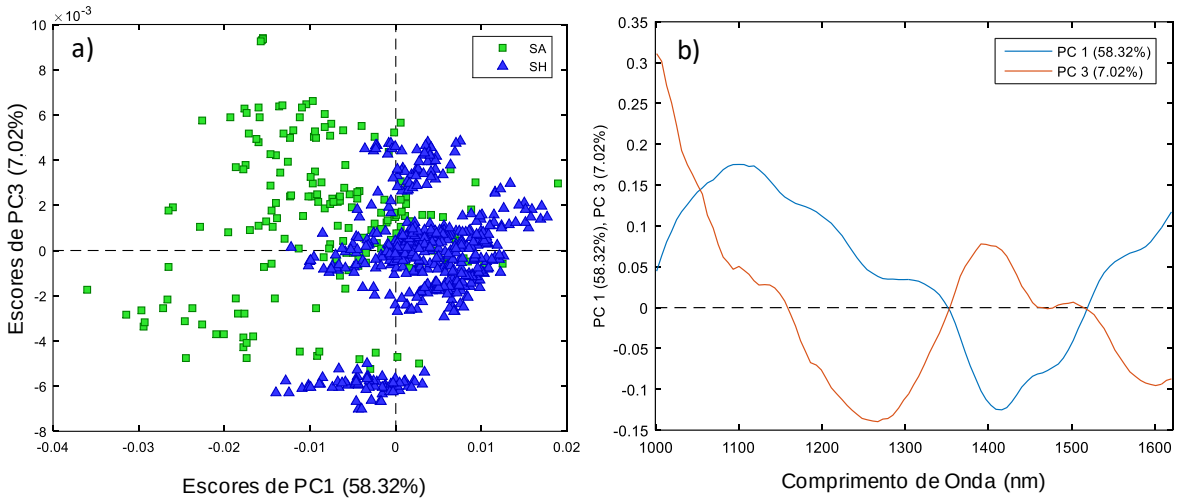
Figura 18 – Gráfico de a) escores e b) *loadings* de PC1 e PC2 das amostras de sangue humano (azul), sangue animal (verde) de todo o conjunto de dados.



Os escores de PC1xPC3 podem ser vistos na Figura 19a e seus *loadings* na Figura 19b. É possível visualizar melhor a tendência de separação entre as classes de SH e SA na primeira componente, enquanto em PC3 é visível a formação de um agrupamento de SH com escores mais negativos, sendo que a terceira componente principal possui 7,02% da variância explicada.

Os *loadings* de PC3 mostram que o início do espectro (1000 nm) exerce uma contribuição positiva nos escores, enquanto as variáveis entre 1200 e 1300 nm evidenciam uma contribuição mais negativa. Essa componente também evidencia a diferença entre um conjunto de amostras com escores com coordenadas mais negativas, esses espectros são de sangue animal e sangue humano depositados no substrato CE5, o que indica que a PC3 apresenta uma variabilidade referente a esse substrato.

Figura 19 – Gráfico de a) escores e b) *loadings* de PC1 e PC3 das amostras de sangue humano (azul), sangue animal (verde) de todo o conjunto de dados.



4.2 CONJUNTOS DE TREINAMENTO E DE PREDIÇÃO

A ideia inicial para definir o conjunto de predição (CP) foi optar pela maior quantidade de substratos possíveis que pudessem ser retirados do conjunto de dados, sem comprometer a sua variabilidade do conjunto de treinamento, de modo que ele fosse o mais representativo possível dos dados originais. Dessa forma, a primeira estratégia, foi retirar substratos que tivessem apenas uma única classe de amostras depositadas em sua superfície, como o caso do piso PO5 que possui 75 espectros de SH, e do piso CE4 que possui 32 espectros de SA. Como as amostras de FPC estão presente nas demais classes de substratos, a estratégia foi retirar o piso que não possuísse amostras de SA, pois essa classe é a que possui a menor quantidade de espectros. Assim, foi retirado o piso CE6.

O primeiro conjunto de predição foi formado então por esses três substratos, correspondendo a 23,3% do conjunto de dados. O primeiro conjunto de treinamento foi formado pelos seis substratos restantes, totalizando 757 espectros.

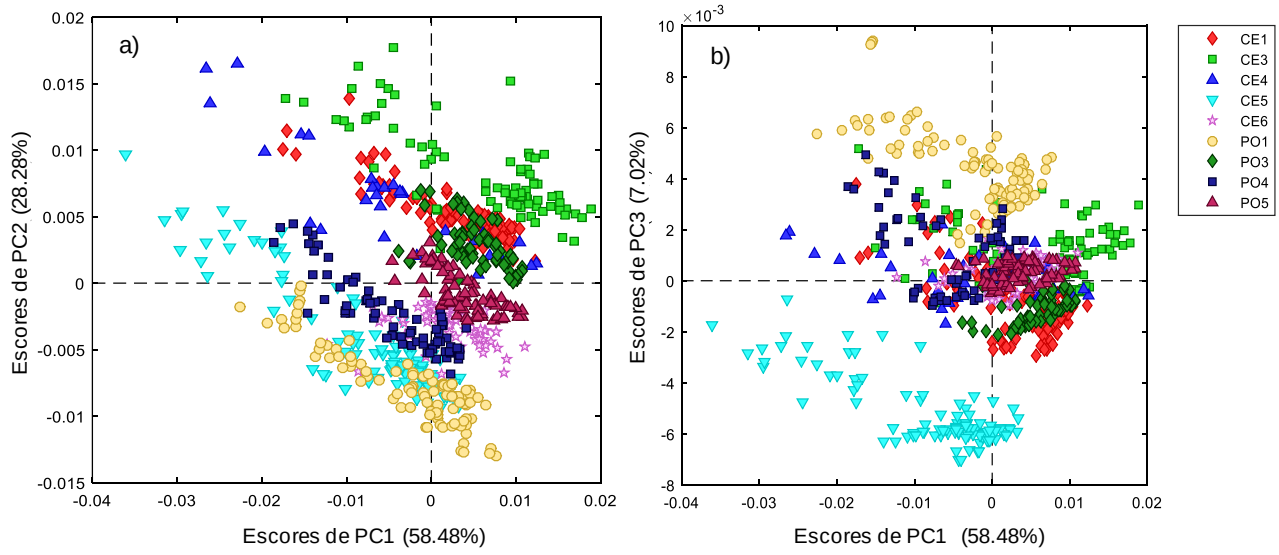
4.2.1 Avaliação da Robustez do Conjunto de Treinamento

Com a intenção de avaliar a escolha do conjunto de treinamento, foram testadas outras combinações de classes de substratos para formar novos conjuntos de treinamento e de predição. A partir da observação dos gráficos de escores de PC1xPC2 (Figura 20a) e PC1xPC3 (Figura 20b) construído usando todos os espectros e a legenda das classes de substratos (Figura 20).

O critério utilizado para a escolha dos novos conjuntos de treinamento e predição foi a diferença ou semelhança entre os pisos, observando-se visualmente os gráficos de escores com as legendas dos substratos. O conjunto de predição 2 (CP2) foi construído a partir da escolha de uma classe de substrato com escores localizados mais internamente e outra classe de substrato com escores localizados mais externamente na PCA. A primeira classe foi a classe PO4 (azul escuro), com 40 espectros de SH, 32 espectros de SA e 56 espectros de FPC. A segunda classe escolhida foi CE5 (anil), com 67 espectros de SH, 32 espectros de SA e 56 espectros de FPC.

Na Figura 20a, pode-se observar que os escores da classe CE5 assemelham-se com os escores de PO1 nas duas primeiras componentes. No entanto, quando observamos a PC3 na Figura 22b, os escores de CE5 são mais negativos e se encontram mais externos em relação aos demais pisos. O PO4 por sua vez, é uma classe que tem as coordenadas dos seus escores mais centralizadas e permanece sempre no interior do modelo.

Figura 20 - Gráfico de escores a) de PC1xPC2 e b) de PC1xPC3 construídos a partir dos espectros de sangue com a legenda das classes dos substratos.



Para o terceiro conjunto de predição (CP3) foram escolhidas duas classes de substratos com escores localizados mais internamente: CE1 (vermelho), com 69 espectros de SH, e 32 espectros de SA e, novamente, o substrato PO4. Comparando-se os escores dos substratos contidos em CP1 com os escores dos substratos de CP3, observa-se que os escores de CE4, CE6 e PO5 estão distribuídos ao longo de toda a PCA, e não estão concentrados internamente.

A Tabela 3 detalha os tipos de substratos e espectros que constituíram cada conjunto de treinamento e predição usados nesse trabalho.

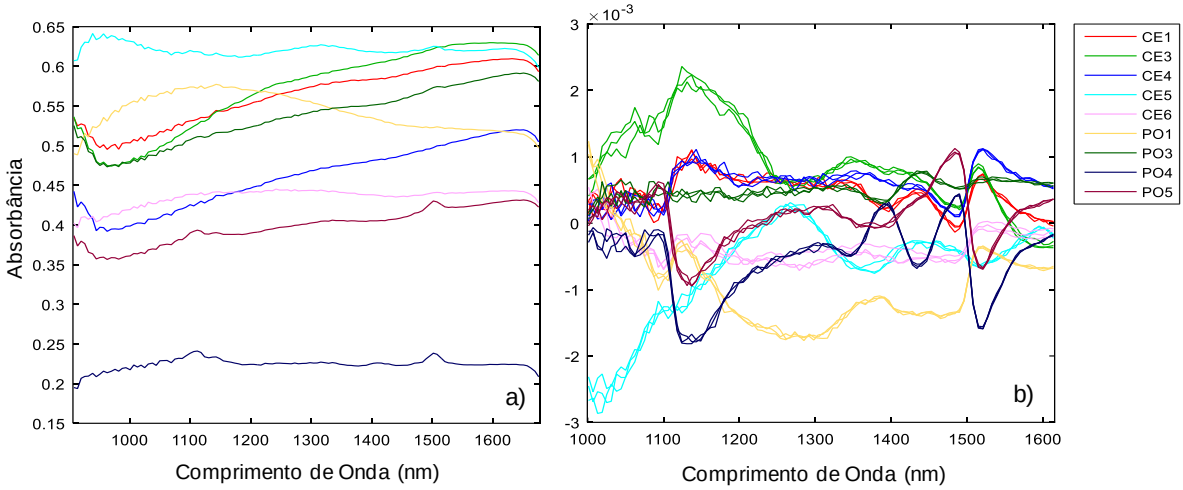
Tabela 3 - Distribuição dos tipos de espectros e de substratos nos diferentes conjuntos de treinamento e de predição.

Conjuntos de Treinamento					**Conjuntos de Predição**				
N°	**Substratos**	**Espectros**			**N°**	**Substratos**	**Espectros**		
		SH	**SA**	**FPC**			**SH**	**SA**	**FPC**
CT1	CE1,CE3,CE5, PO1,PO3,PO4	365	168	224	**CP1**	CE4,PO5,CE6	141	32	56
CT2	CE1,CE3,CE4,CE6, PO1,PO3,PO5	375	136	168	**CP2**	CE5 e PO4	131	64	112
CT3	CE3,CE4,CE5,CE6, PO1,PO3,PO5	397	136	224	**CP3**	CE1 e PO4	109	64	56

4.2.2 Avaliação dos Espectros de Substratos Puros

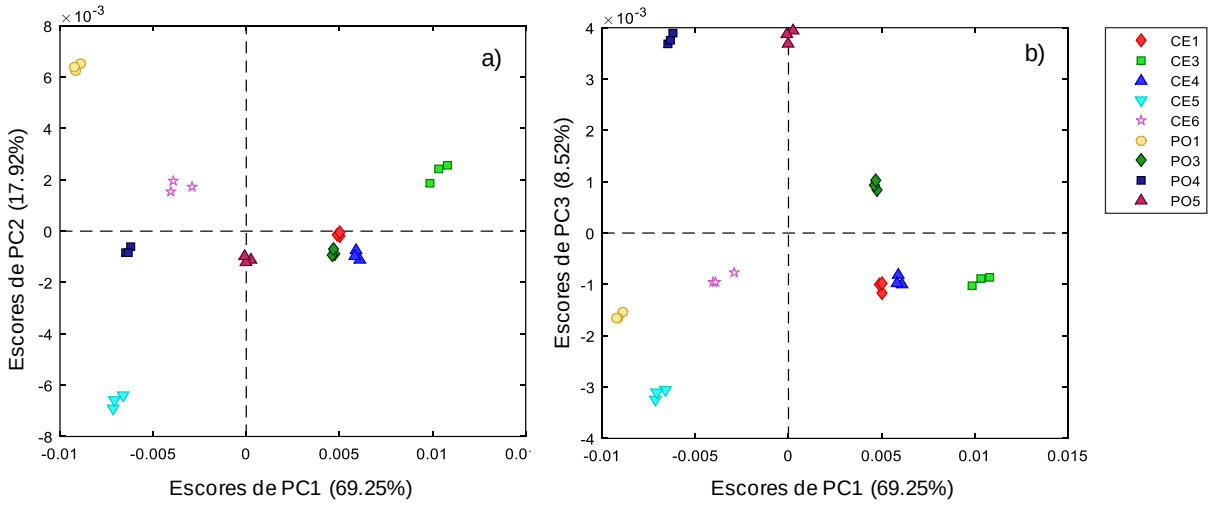
Ainda com relação a escolha do conjunto de treinamento e do conjunto de predição, e com o objetivo de avaliar quais os substratos apresentavam maiores semelhanças, foi realizada a PCA utilizando apenas os espectros dos substratos puros, sem nenhuma substância depositada em sua superfície. Na Figura 21, podem ser vistos os espectros originais (Figura 21a), pré-processados usando 1ª Derivada com suavização de 15 pontos (Figura 21b).

Figura 21 – Comparação entre os a) espectros originais e b) pré-processados dos substratos limpos.



Os gráficos de escores de PC1xPC2 (Figura 22a) e PC1xPC3 (Figura 22b). A partir da observação dos escores de PC2 da Figura 22a, pode-se constatar que alguns substratos apresentam coordenadas mais próximas nessa componente, como CE1, CE4, CE6, PO3, PO4 e PO5, enquanto na primeira componente os escores são mais similares entre CE1, CE4 e PO3, e entre CE5, CE6, PO4 e PO1.

Figura 22 - Gráficos de a) escores de PC1xPC2 b) PC1xPC3 da PCA dos espectros de substrato.

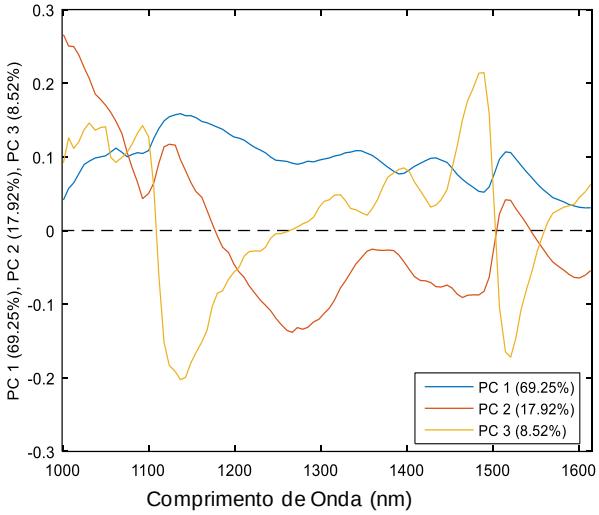


Observando PC3 (Figura 22b) é possível identificar que PO4 e PO5 têm escores muito parecidos nessa componente. As classes de substratos com escores próximos ao centro dos eixos das componentes principais serão chamados de substratos internos, enquanto as classes de substratos com escores mais positivos ou negativos, ou seja, que se distanciam do centro, serão chamados de substratos externos.

Dessa forma, o conjunto de predição 1 (CP1) formado pelos substratos CE4, CE6 e PO5, têm escores localizados mais internamente no gráfico de PC1xPC2, enquanto em PC3 os escores de PO5 estão localizados mais externamente, mas ainda próximos dos escores PO4. Em relação a CP2, formado por CE5 e PO4, as coordenadas dos escores de CE5 apresentam uma localização mais externas nas três componentes. Já CP3, formado por CE1 e PO4, têm ambos os substratos com escores localizados próximos ao centro em PC1 e PC2, e em PC3, os escores de PO4 são mais positivos, porém assemelham-se a PO5.

Na Figura 23 podem ser visto os *loadings* das três primeiras componentes do modelo de PCA criado com os espectros dos substratos limpos. A região espectral de maior influência na primeira componente encontra-se no início do espectro, entre 1000 e 1200 nm.

Figura 23 - Gráficos de *loadings* das três primeiras componentes da PCA dos espectros de substrato.



4.3 MODELOS HIERÁRQUICOS

Foram construídos três tipos de modelos hierárquicos para cada conjunto de treinamento, totalizando nove modelos. Todos eles foram formados por duas regras de decisão: a primeira consistiu em um modelo de PCA que foi responsável pela separação dos espectros de FPC e de alguns possíveis *outliers*, enquanto a segunda regra foi responsável por identificar as amostras de SH e SA, essa regra foi diferente para cada tipo de modelo hierárquico, o primeiro modelo usou um modelo de PCA, o segundo usou um modelo de SIMCA e o terceiro usou um modelo de PLS-DA.

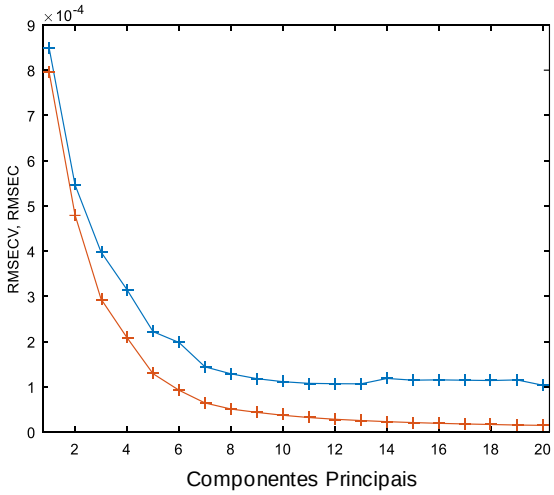
O procedimento descrito a seguir foi realizado igualmente para todos os conjuntos de treinamento e predição, contudo serão descritos apenas para os CT1 e CP1. No Apêndice C e D podem ser visualizados os gráficos obtidos através de CT2 e CT3.

4.3.1 Separando falsos positivos comuns dos espectros de sangue

Como foi observado ao longo do tópico 5.1, os espectros de FPC são muito diferentes dos espectros de SH e SA. Dessa forma, uma estratégia para facilitar a identificação das amostras de sangue, foi criar um primeiro filtro, capaz de separar os espectros de amostras muito diferentes, como os de FPC. Para isso, foi construída uma PCA utilizando apenas os espectros de sangue do conjunto de treinamento, o que resultou no total de 533 espectros.

O pré-processamento utilizado foi o mesmo discutido ao longo do Tópico 5.1. Além disso, o modelo foi construído usando validação cruzada por blocos aleatórios. O número de componentes principais foi selecionado observando os gráficos de RMSEC e RMSECV (Figura 24), os gráficos de escores (Figura 26a), e os *loadings* (Figura 26b) de cada componente.

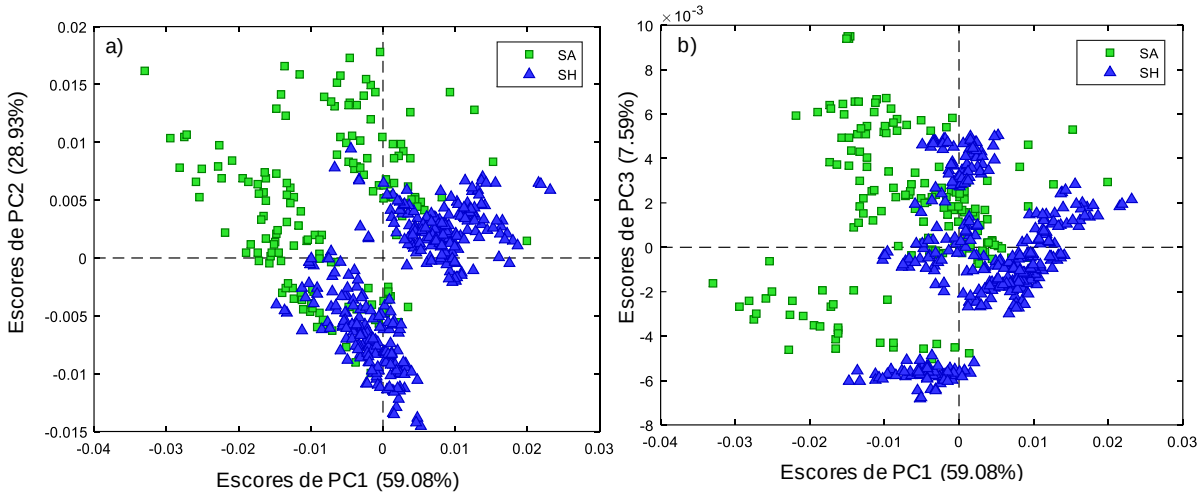
Figura 24 – Gráfico de RMSEC (laranja) e RMSECV (azul) versus o Número de Componentes Principais.



A diferença entre os valores de RMSEC e RMSECV é muito pequena, independentemente da quantidade de componentes escolhidas, visto que a escala do gráfico é de valores elevados a 10^{-4}. Dessa forma, para escolher o número de componentes observou-se os *loadings* e a variância explicada pelas componentes, após a retirada das amostras com comportamentos anômalos no conjunto de espectros.

Foram usadas as três primeiras PCs, que juntas explicaram 94,12% da variabilidade dos espectros de sangue. Na Figura 25 podem ser vistos os escores de PC1xPC2 (Figura 25a) e PC1xPC3 (Figura 25b) dos espectros de sangue presentes no conjunto de treinamento 1, o comportamento dos escores é muito semelhante ao observado na Figura 20, onde os escores da classe SH também possuem coordenadas mais positivas em PC1 em relação aos escores da classe SA.

Figura 25 - Gráfico de escores a) de PC1xPC2 e b) de PC1xPC3 construídos a partir dos espectros de sangue de CT1.

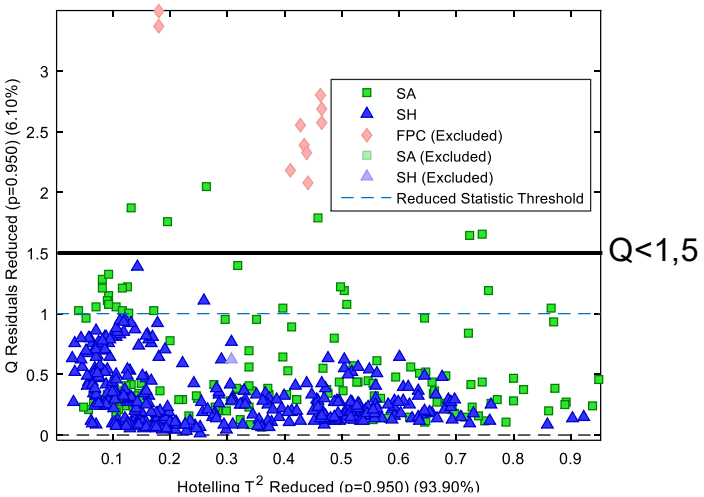


Uma vez definido o número de componentes, avaliou-se a presença de possíveis *outliers* e projetou-se os espectros de FPC do conjunto de treinamento na PCA e o Gráfico de Influência (Figura 24) foi usado para estabelecer o critério que seria aplicado como regra de decisão no modelo hierárquico que, no caso, foi o Q-Estatístico.

Essa escolha baseou-se na localização das projeções dos FPC no gráfico de resíduos da PCA. A linha preta no gráfico na Figura 26 ilustra o limite estipulado por esse critério, com o qual o é possível reter as amostras de FPC ou outras amostras apresentam comportamento fora do que se considera comportamento padrão das amostras de SA e SH. De modo que as amostras com Q<1,5 seguiram para a etapa seguinte e as demais foram classificados como não-sangue (NS). Essa estratégia para definir limites ou critérios de classificação no modelo de PCA é muito parecida com a utilizada pelo método de classificação por SIMCA que se baseia nos valores de Q-Estatístico e T^2 de Hotteling com 95% de confiança para definir se uma amostra pertence ou não a uma classe.

É importante enfatizar que para melhorar a visualização e a definição do critério, apenas uma região do gráfico de influência foi ampliada, sendo que a grande maioria das amostras de FPC apresentam valores de Q-Estatístico e T^2 de Hotteling muito elevados e por isso não aparecem na figura. Algumas amostras de SH e SA apresentaram valor de Q-Estatístico acima do valor estipulado, no entanto, essa restrição imposta não trouxe prejuízos para classificação dos espectros de sangue presentes nos conjuntos de predição.

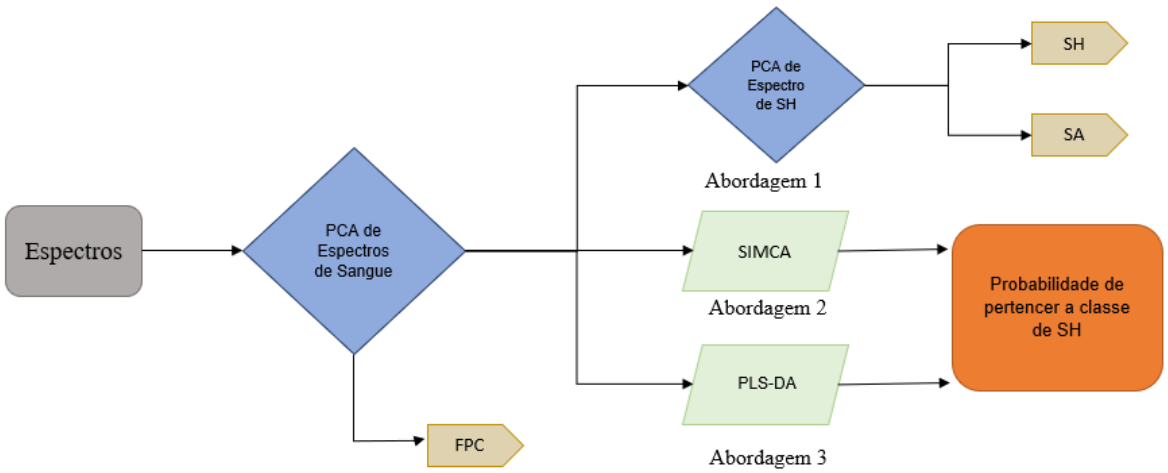
Figura 26 - Gráfico de Influência (Q-Residual Reduzido x T² de Hotteling Reduzido) da PCA construída com os espectros de SH e SA do CT1.



4.3.2 Identificando o Sangue Humano

Na segunda regra do modelo hierárquico, foram realizadas três abordagens diferentes: a primeira delas usou como critério escolha um modelo PCA, a segunda usou um modelo SIMCA e a última, um modelo PLS-DA. Conforme pode ser visto a partir da Figura 27.

Figura 27 - Fluxograma dos diferentes tipos de abordagem usados para construção dos modelos.



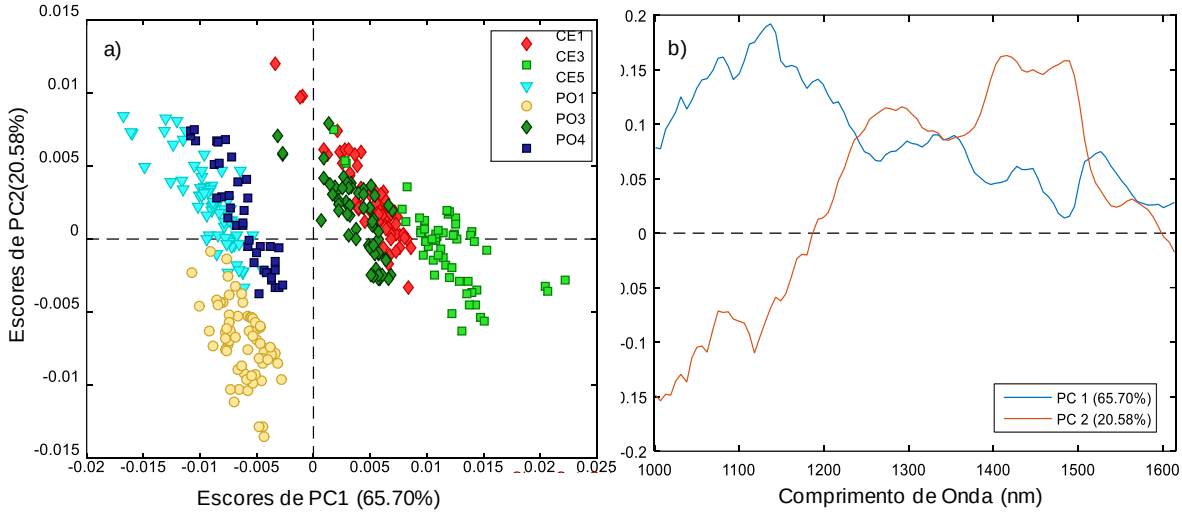
4.3.2.1 PCA

Com o objetivo de identificar as amostras de SH e de SA, foi construído um modelo de PCA com apenas os espectros de sangue humano do conjunto de treinamento. Os gráficos de escores e o gráfico de influência foram analisados, de modo semelhante ao explicado na

primeira regra, e os espectros de sangue animal foram projetados no modelo para observar o comportamento dessas amostras.

Também foram usadas as três primeiras componentes principais para definir o modelo de PCA. A Figura 28a mostra os gráficos de escores de PC1xPC2, enquanto a Figura 28b mostra os *loadings* dessas componentes. Pode-se observar uma tendência de separação entre os escores de SH, que se justificam pelo fato das amostras pertencerem a classes de substratos diferente, sendo que os escores mais positivos em PC1 são de espectros de SH dos substratos CE1, CE3 e PO3. O grupo com escores mais negativos correspondem aos espectros de manchas de SH depositados nos substratos PO1, PO4 e CE5.

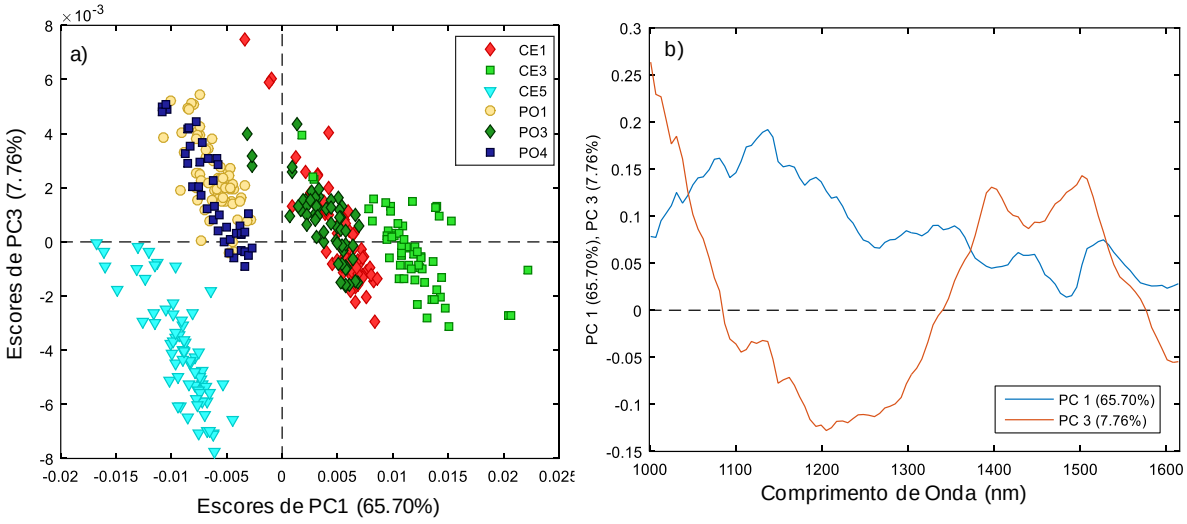
Figura 28 - Gráfico de a) escores de PC1xPC2 e b) os *loadings* construídos a partir dos espectros de sangue humano de CT1 e legenda com as classes de substratos.



Os *loadings* de PC1 e PC2 mostram que há alguma influência de ruídos nas variáveis iniciais, o que coincide com a região em que há maior interferência espectral devido aos substratos.

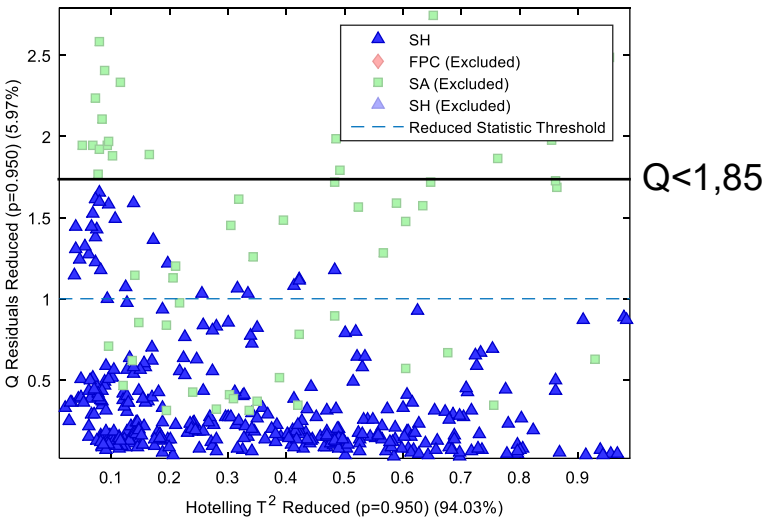
A Figura 29a mostra os gráficos de escores de PC1xPC3, enquanto a Figura 29b mostra os *loadings* dessas componentes. Através do gráfico de escores de PC3 pode ser percebido que os escores de CE5 apresentam coordenadas mais negativas nessa componente do que os demais substratos. Os *loadings* da terceira componente têm uma maior contribuição negativa da faixa entre 1100 nm e 1300 nm, o que indica que essa região tem uma maior influência do substrato CE5 do que informações espectrais referentes ao sangue.

Figura 29 - Gráfico de a) escores de PC1xPC3 e b) os *loadings* construídos a partir dos espectros de sangue humano de CT1 e legenda com as classes de substratos.



Os espectros de SA do CT1 foram projetados no modelo de PCA e o gráfico de influência (Figura 30) foi novamente utilizado para estabelecer o critério de escolha: Q-Estatístico <1,85, conforme ilustra a linha preta. Assim as amostras que atenderam este critério foram classificadas como SH, e as demais como SA.

Figura 30 - Gráfico de Influência (Q-Residual Reduzido x T² de Hotteling Reduzido) da PCA construída apenas com os espectros de SH do CT1.



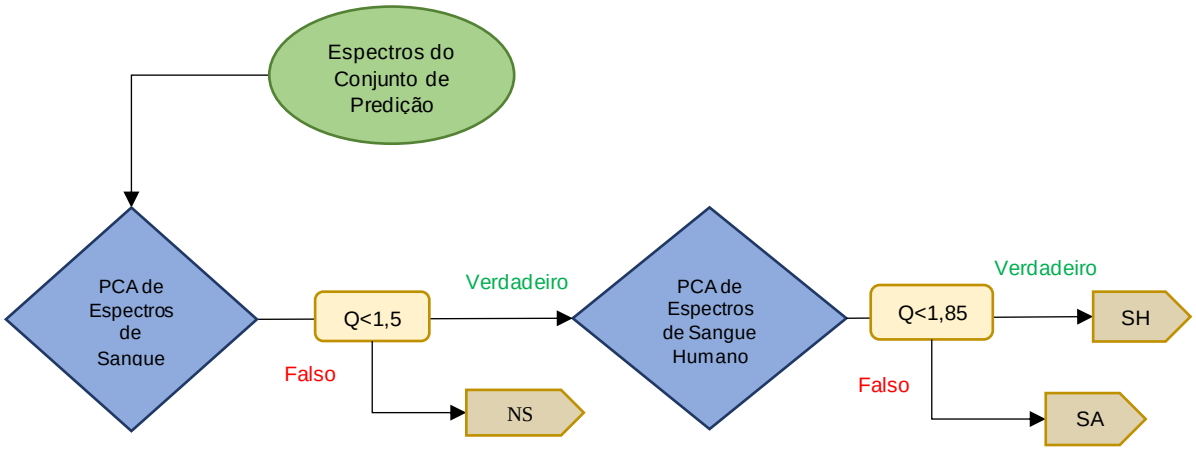
É possível visualizar na Figura 30 que algumas amostras de SA do CT1 atenderam ao critério estabelecido e seriam classificadas como SH se fossem utilizadas para predição. No entanto, é importante salientar que em uma cena de crime nenhum vestígio de sangue pode deixar de ser coletado, dessa forma é preferível classificar erroneamente uma amostra de sangue animal do que deixar de classificar uma amostra de sangue humano.

O fluxo de classificação de amostras de predição, ocorreu da seguinte maneira: os espectros do conjunto de predição foram introduzidos no MH e foram projetados no primeiro modelo de PCA, as amostras que não atenderam ao critério Q<1,5 não passam para etapa seguinte e são classificadas como não-sangue (NS), as amostras que atenderam foram projetadas no segundo modelo de PCA e tiveram seus valores de Q-Estatístico comparados ao critério Q<1,85 e as que atenderam foram classificadas como SH, e as demais como SA.

A partir das constatações obtidas através das observações do modelo de PCA construído para espectros de sangue humano, pôde-se desenvolver o modelo hierárquico. Os nós e critérios utilizados no Modelo Hierárquico 1.A (MH1.A) foram os modelos de PCA descritos anteriormente, e podem ser visualizados no fluxograma da Figura 31.

Os valores de sensibilidade e especificidade da validação interna obtidos através da projeção das amostras do CT1 no modelo MH1.A foram, respectivamente, 0,98 e 0,88. Esses valores representam um bom ajuste do modelo. Como esperado, o valor da especificidade é menor do que o valor de sensibilidade, isso ocorre porque algumas amostras de SA foram classificadas como SH, devido a escolha do critério de decisão, conforme pôde ser visto na Figura 30.

Figura 31 - Fluxograma do Modelo Hierárquico 1. A construído usando o CT1.



As amostras de predição de CP1 foram então submetidas ao MH1.A, e os valores de sensibilidade e especificidade calculados foram iguais a 1,0 e 0,78, respectivamente. Todos os espectros de SH e FPC foram classificados corretamente, contudo, 69% dos espectros de SA foram classificadas como SH. Esse resultado poderia ser melhorado se o valor do Q-Estatístico usado como critério de decisão fosse mais restritivo. Por exemplo, adotando o critério de Q<1,0, a sensibilidade cai para 0,69 e a especificidade aumenta para 0,85. Contudo, o ganho na

especificidade não justifica a redução da sensibilidade, visto que isso aumenta o risco de classificar incorretamente espectros de SH, o que seria um erro muito mais grave em um contexto forense.

Usando esse mesmo procedimento, foram construídos os modelos hierárquicos para os outros conjuntos de treinamento, CT2 e CT3 (esses modelos serão chamados aqui de MH2.A e MH3.A, respectivamente). E os conjuntos de predição foram submetidos, respectivamente, a cada um deles. Os resultados de sensibilidade e especificidade para cada modelo podem ser comparados na Tabela 4. As figuras de mérito obtidas a partir da validação interna para todos os modelos apresentam sensibilidade acima de 0,98 e especificidade acima de 0,84. Esses valores indicam um bom ajuste para os modelos de classificação.

Tabela 4 – Resumo dos resultados dos modelos hierárquicos que usaram um modelo de PCA como segunda regra de decisão.

Conjunto de Treinamento	**Modelo Hierárquico**	**Validação Interna**		**Predição**	
		Sensibilidade	**Especificidade**	**Sensibilidade**	**Especificidade**
CT1	MH1.A	0,98	0,88	1	0,78
CT2	MH2.A	0,98	0,84	0,33	0,7
CT3	MH3.A	0,98	0,94	1	0,94

Os valores de Sn e Sp para as amostras preditas por MH2.A foram de 0,33 e 0,7, respectivamente. Já era esperada uma piora nos resultados devido a escolha do conjunto de predição CP2, visto que uma classe dos substratos apresentava escores com coordenadas mais externas em relação aos demais.

E o melhor resultado foi obtido usando o MH3.A, com Sn igual a 1,0 e Sp igual a 0,94. O conjunto de predição CP3 foi formado por duas classes de substratos com variabilidade mais próxima aos demais, cujos escores estavam localizados mais internamente no modelo de PCA. De modo que o conjunto de treinamento CT3 possui uma maior variabilidade, capaz de contemplar a variabilidade dos substratos do conjunto de predição CP3.

4.3.2.2 SIMCA

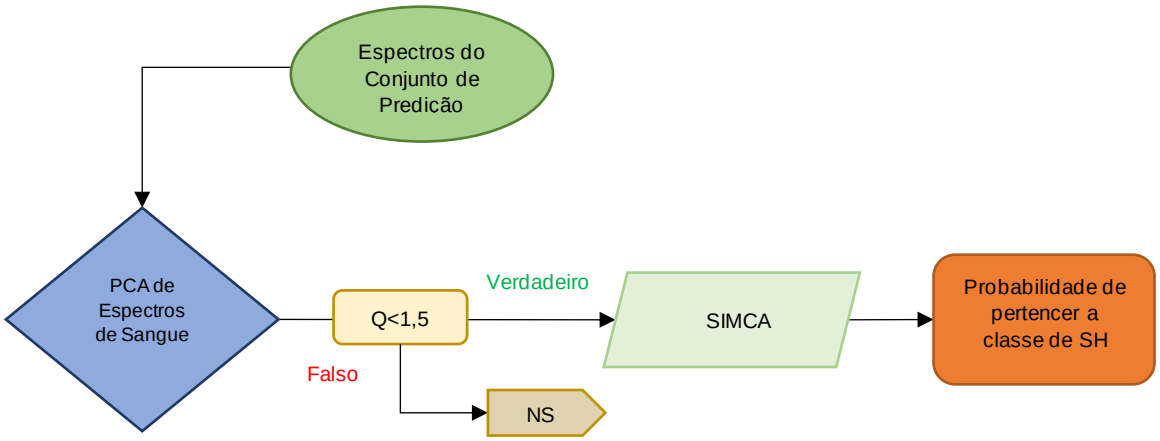
Outra estratégia para identificar o sangue humano foi construir um modelo SIMCA de classe única e usá-lo como segunda regra do modelo hierárquico no lugar do modelo de PCA descrito no tópico anterior.

O modelo SIMCA foi construído para uma única classe, a de sangue humano, e o número de componentes principais usados para definir a classe foi o mesmo da PCA de SH

descrita no tópico 5.3.2.1. Por ser um modelo de classificação flexível, ele permite que amostras desconhecidas não sejam classificadas na classe modelada, e a resposta final obtida nesta etapa foi a probabilidade de a amostra pertencer a classe de SH. De modo que as amostras que obtiveram um valor de probabilidade acima de 50% foram classificadas como pertencentes a classe SH.

O fluxograma apresentado na Figura 32 mostra como foi construído o modelo hierárquico para o CT1 com a segunda regra de decisão sendo o modelo SIMCA (MH1.B). De modo similar, foram construídos os modelos hierárquicos usando o SIMCA para o CT2 e CT3, que serão chamados de MH2.B e MH3.B, respectivamente.

Figura 32 - Fluxograma do Modelo Hierárquico 1.B construído usando o CT1.



Os resultados de sensibilidade e especificidade da validação interna e das amostras de predição para cada modelo podem ser visualizados na Tabela 5.

Tabela 5 – Resumo dos resultados dos modelos hierárquicos que usaram um modelo SIMCA como segunda regra de decisão.

Conjunto de Treinamento	Modelo Hierárquico	Validação Interna		Predição	
		Sensibilidade	Especificidade	Sensibilidade	Especificidade
CT1	MH1.B	0,91	0,93	0,92	0,85
CT2	MH2.B	0,96	0,91	0,03	0,69
CT3	MH3.B	0,99	0,92	0,94	1

Os valores de sensibilidade obtidos para a validação interna foram menores em todos os MH usando SIMCA como segunda regra em relação aos MH que usaram PCA como segunda regra de decisão. Isso ocorre devido ao fato de que o SIMCA se baseia nos valores de Q-Estatístico e T^2 de Hotteling definidos para 95% confiança, o que restringe ainda mais os limites de classificação.

Os valores de Sn e Sp para MH1.B foram iguais a 0,92 e 0,85, o que representa um aumento da especificidade em relação aos valores obtidos para o modelo MH1.A. Contudo, ocorre uma diminuição da sensibilidade de 1,0 para 0,92, ou seja, um aumento do erro na classificação de amostras de SH. Quando comparado ao modelo anterior, MH2.B também teve uma diminuição na sensibilidade de 0,33 para 0,03. Enquanto a especificidade não sofreu alteração estatística relevante se mantendo em aproximadamente 0,7. Para o modelo MH3.B, a especificidade aumentou para 1,0. No entanto, também ocorreu uma diminuição da sensibilidade em relação a MH3.A de 1,0 para 0,94.

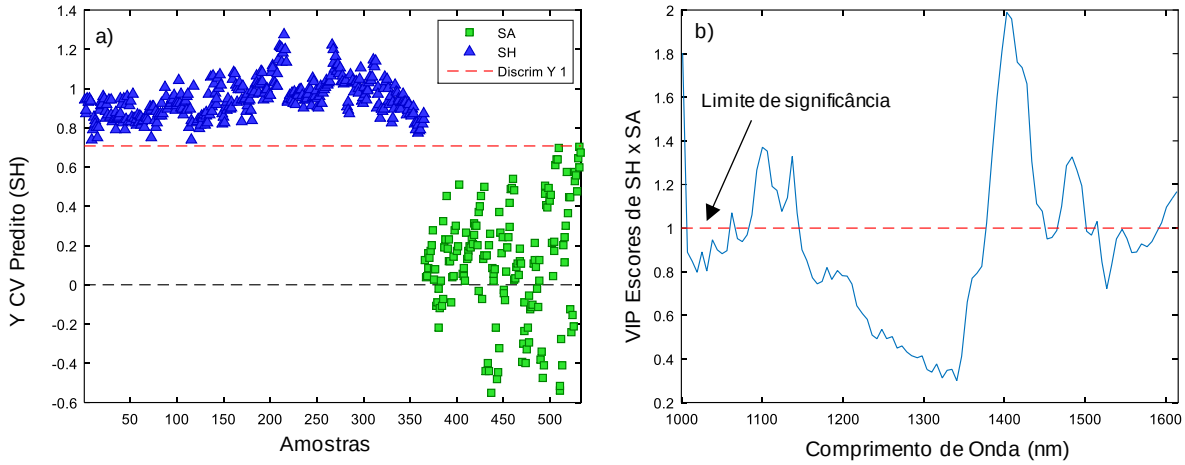
4.3.2.3 PLS-DA

A última estratégia utilizada para diferenciar o sangue humano do sangue animal foi construir um modelo de classificação usando o PLS-DA. Por ser baseado em análise discriminante, essa técnica de classificação precisa de ao menos duas classes para que seja delimitada uma fronteira entre elas. Dessa forma, esse modelo não pôde ser construído apenas com os espectros de sangue humano, de modo que os espectros de sangue animal também foram considerados para a construção do limiar de discriminação entre as duas classes.

Foram usadas cinco variáveis latentes para construir o modelo, esse número foi escolhido a partir da comparação dos valores de RMSEC e RMSECV. A validação interna foi realizada através de validação cruzada usando blocos aleatórios. A Figura 33a mostra o gráfico de escores de SH e SA, todas as amostras foram classificadas corretamente, e a especificidade e sensibilidade do modelo foram iguais a 1. Na Figura 33b pode ser visto o gráfico de VIP escores do modelo, a partir dele pode-se interpretar quais as variáveis tiveram maior influência para a discriminação das duas classes observadas.

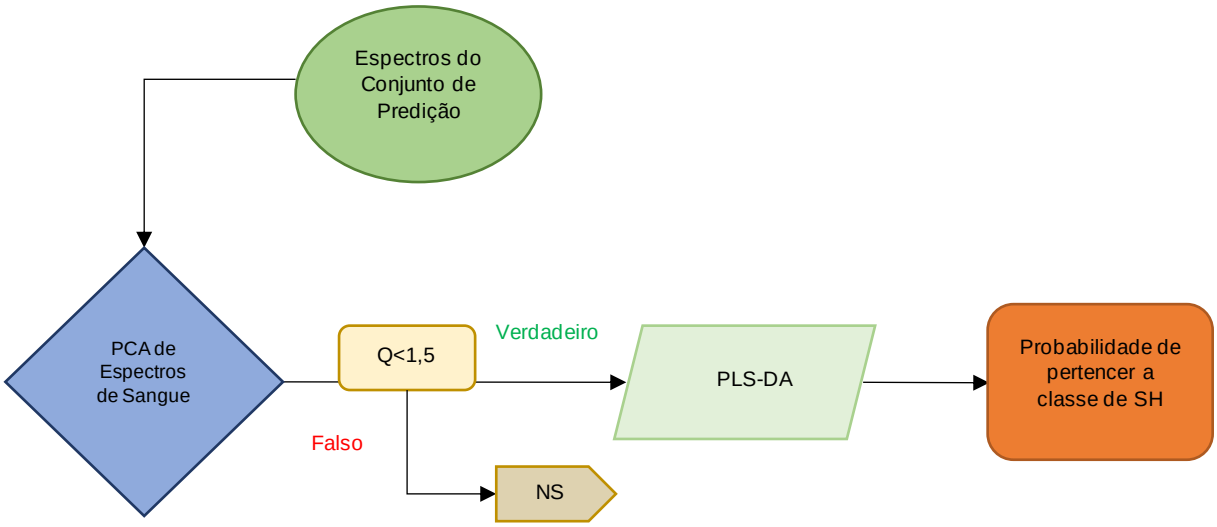
Observa-se que as variáveis acima do limite de significância pertencem aproximadamente a região entre 1370 nm e 1510 nm, é nessa faixa espectral que estão localizadas as regiões do primeiro sobretom de estiramento da ligação $-OH$ proveniente da água em 1440-1485 nm e do primeiro sobretom de estiramento da ligação $-NH$ presentes em proteínas (1511 nm) (WORKMAN; WEYER, 2009).

Figura 33 – Gráfico a) dos escores de classificação PLS-DA para as amostras de SH e SA e b) dos VIP escores para a discriminação das amostras de SH e de SA do CT1.



A resposta de saída do modelo hierárquico foi a probabilidade de pertencer a classe SH. Como pode ser visto no fluxograma de MH1.C representado na Figura 34. De maneira semelhante, foram construídos os modelos MH2.C e MH3.C para os conjuntos CT2 e CT3.

Figura 34 - Fluxograma do Modelo Hierárquico 1.C construído usando o CT1.



Os resultados de classificação da validação interna e dos conjuntos de predição podem ser vistos na Tabela 6. Com relação as figuras de mérito referentes a validação interna há um aumento da sensibilidade e especificidade em relação aos modelos construídos usando PCA ou SIMCA. Os valores de sensibilidade e especificidade para os conjuntos de predição do modelo MH1.C foram iguais a 1,0. Também foram obtidos valores elevados para MH3.C, sendo Sn igual a 0,97 e Sp igual a 1. Já para MH2.C esses valores foram 0,22 e 0,34, respectivamente.

Tabela 6 – Resumo dos resultados dos modelos hierárquicos que usaram um modelo PLS-DA como segunda regra de decisão.

Conjunto de Treinamento	Modelo Hierárquico	Validação Interna		Predição	
		Sensibilidade	Especificidade	Sensibilidade	Especificidade
CT1	MH1.C	1,00	1,00	1	1
CT2	MH2.C	0,99	1,00	0,22	0,74
CT3	MH3.C	0,99	1,00	1	1

Conforme pôde ser visto na Figura 28b, a região dos espectros de sangue que aparentam ter maior influência dos substratos se encontra no início do espectro, na faixa entre 1100 e 1300 nm. O gráfico de VIP escores da Figura 33b também indica que a região do espectro que possui maior influência na discriminação de SH e SA é a faixa a partir de 1300 nm.

Baseado nessas observações, e com o objetivo de avaliar se haveria uma melhora na classificação do sangue humano, decidiu-se reduzir os espectros e usá-los para construir novos modelos hierárquicos a partir da faixa de 1304 nm até 1614 nm. As classes de substrato dos conjuntos de treinamento foram mantidas, e a única alteração além do corte foi o pré-processamento realizado que, nesse caso, consistiu em 1ª Derivada com polinômio do 2º grau e suavização de 7 pontos. Os espectros reduzidos e pré-processados, assim como os modelos de PCA podem ser visualizados no Apêndice E.

4.3.3 Comparando os Resultados de Classificação

Os resultados de classificação obtidos para cada tipo de MH podem ser comparados através das Tabela 7 e 8, onde a primeira foi construída pelos conjuntos de treinamento que usaram a faixa espectral entre 1000 e 1614 nm, e a segunda pelos que usaram a faixa entre 1304 e 1614nm. Ambas mostram detalhes dos modelos, os percentuais de classificação correta para cada uma das classes e os valores de sensibilidade e especificidade calculados

Observando as tabelas, fica evidente que os piores resultados foram obtidos usando o CT2, isso era esperado devido a menor variabilidade dos substratos que formaram esses conjuntos, visto que os escores do piso CE5 apresentavam coordenadas mais distantes dos demais. Isso evidencia a importância da escolha do conjunto de treinamento.

Tabela 7 – Resultados de classificação dos diferentes modelos hierárquicos construídos com os espectros na faixa de 1000 a 1614 nm.

Conjunto de Treinamento	Modelo	ETAPA 1			ETAPA 2			PREDIÇÃO								
		Técnica	N° de PCs	%Variab. Acumulada	Técnica	%Variab. Acumulada	N° de PCs/LVs	Conjunto de Predição	Espectros			Classificação Correta			Sn	Sp
									SH	AS	FPC	SH	SA	FPC		
1	MH1.A	PCA	3	94,12%	PCA	94,02%	3	CE4 CE6 PO5	141	32	56	100%	39%	100%	1	0,78
	MH1.B	PCA			SIMCA	94,02%	3					92%	58%	100%	0,92	0,85
	MH1.C	PCA			PLSDA	96,94% X 82,89% Y	5					100%	100%	100%	1	1
2	MH2.A	PCA	3	95,13%	PCA	94,62%	3	CE5 PO4	107	64	112	34%	17%	100%	0,34	0,7
	MH2.B	PCA			SIMCA	94,62%	3					3%	16%	100%	0,03	0,69
	MH2.C	PCA			PLSDA	96,06% X 84,61% Y	3					22%	30%	100%	0,22	0,74
3	MH3.A	PCA	3	96,52%	PCA	92,56%	3	CE1 PO4	109	64	56	100%	77%	100%	1	0,9
	MH3.B	PCA			SIMCA	92,56%	3					94%	89%	100%	0,94	0,94
	MH3.C	PCA			PLSDA	96,79% X 76,24% Y	4					100%	97%	100%	1	0,99

Tabela 8 – Resultados de classificação dos diferentes modelos hierárquicos construídos com os espectros na faixa de 1300 a 1600 nm

Conjunto de Treinamento	Modelo	ETAPA 1			ETAPA 2			Conjunto de Predição	PREDIÇÃO						Sn	Sp
		Técnica	N° de PCs	%Variab. Explicada	Técnica	%Variab. Acumulada	N° de PCs/LVs		Espectros			Classificação Correta				
									SH	AS	FPC	SH	SA	FPC		
1	MH4.A	PCA	3	95,43%	PCA	93,45%	3	CE4 CE6 PO5	141	32	56	100%	58%	100%	1	0,85
	MH4.B	PCA			SIMCA	93,45%	3					100%	71%	100%	1	0,9
	MH4.C	PCA			PLSDA	94,95% X 81,8% Y	4					100%	100%	100%	1	1
2	MH5.A	PCA	3	95,86%	PCA	94,52%	3	CE5 PO4	107	64	112	100%	41%	100%	1	0,78
	MH5.B	PCA			SIMCA	94,52%	3					61%	56%	100%	0,61	0,84
	MH5.C	PCA			PLSDA	94,47% X 77,28% Y	3					63%	56%	100%	0,63	0,84
3	MH6.A	PCA	3	96,78%	PCA	93,13%	3	CE1 PO4	109	64	56	100%	55%	100%	1	0,76
	MH6.B	PCA			SIMCA	93,13%	3					62%	81%	100%	0,62	0,9
	MH6.C	PCA			PLSDA	94,49% X 76,57% Y	4					100%	86%	100%	1	0,94

Os melhores resultados foram obtidos utilizando os modelos MH1.C e MH4.C, que classificaram corretamente 100% das amostras de cada classe e obtiveram resultado de sensibilidade e especificidade iguais a 1. Lembrando que esses modelos foram construídos usando o CT1 formado pelas mesmas classes de substrato (CE1, CE3, C35, PO1, PO3 e PO4), de modo que a diferença entre eles é apenas a faixa espectral reduzida para o modelo MH4.C.

Com a exceção dos modelos hierárquicos construídos com o conjunto CT2, todos os modelos tiveram um percentual de classificação correta para sangue humano acima de 90%. Quando se avalia os modelos que usaram como segunda regra de decisão um modelo de PCA ou um modelo de PLS-DA, tem-se que 100% das amostras de sangue humano foram classificadas corretamente, ou seja, não foram obtidos nenhum resultado falso negativo.

Os modelos que usaram o PLS-DA foram os que tiveram melhor performance em relação a identificação do sangue animal, isso era esperado porque essa técnica baseia-se em ambas as classes SH e SA para delimitar as fronteiras entre elas. Contudo, deve-se destacar que a classe de sangue animal foi formada não apenas por uma menor quantidade de espectros, mas também por amostras fornecidas por duas espécies de animais, o que pode indicar um sobreajuste nos modelos que usaram essa técnica de classificação como regra de decisão.

Deve se destacar que todos os MH construídos usando os conjuntos CT3 também obtiveram excelentes resultados para classificação de SH e SA, mesmo nos modelos que usaram um modelo de PCA construído apenas com SH como regra de decisão (todos os resultados de sensibilidade e especificidade foram acima de 0,94). Isso demonstra que essas técnicas são bastante efetivas e podem ser utilizadas para identificação e classificação de sangue, com a vantagem de não apresentar os mesmos riscos de sobreajuste que os modelos construídos com PLS-DA.

Com relação aos resultados obtidos por trabalhos encontrados na literatura, nenhum artigo estudou a identificação de sangue em diferentes substratos usando um mesmo modelo de classificação para todos eles simultaneamente. Dessa forma, os resultados aqui apresentados serão comparados com os resultados obtidos por Pereira e colaboradores (2017) em que os modelos de classificação foram construídos individualmente para cada tipo de substrato.

Os modelos construídos por Pereira *et al.* (2017) para classificar amostras de SH, SA e FPC depositados em um tipo de porcelanato obtiveram sensibilidade e especificidade iguais a 1, classificando corretamente 100% das amostras de usando as técnicas de SPA-LDA, GA-LDA e PLS-DA. Usando a técnica de SIMCA, 100% das amostras de SH e FPC foram classificadas

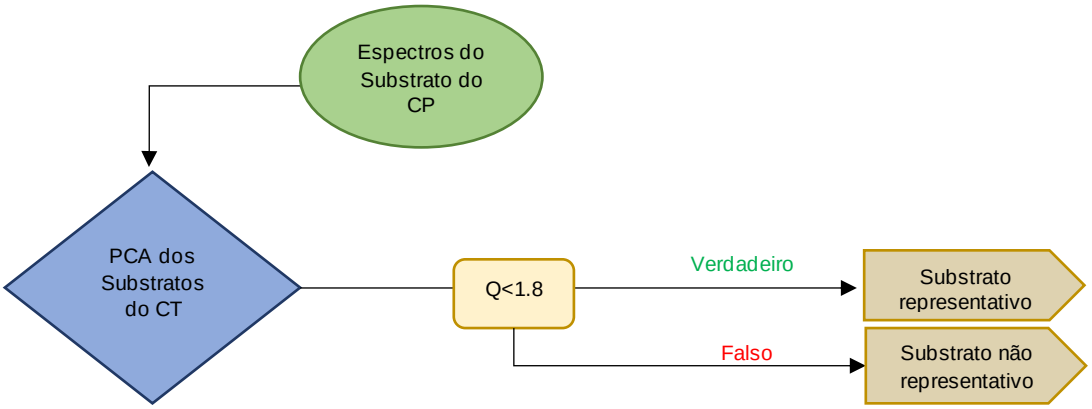
Esse estudo foi realizado usando apenas um único tipo de porcelanato e um único tipo de cerâmica. Quando comparamos com os resultados obtidos no presente trabalho, usando o modelo construído a partir de diferentes pisos de cerâmica e porcelanato, observa-se uma melhora nos resultados de classificação de SH para as amostras depositadas em cerâmica.

4.3.4 Protocolo de Representatividade dos Pisos

Observando os diferentes resultados de sensibilidade e especificidade, após a construção dos diferentes modelos hierárquicos e a constatação de que a escolha dos substratos a serem utilizados no conjunto de treinamento influenciava na identificação dos vestígios de sangue, percebeu-se a necessidade de realizar um procedimento anterior a classificação: avaliar se o substrato onde o vestígio de sangue foi encontrado tem a variabilidade representada pelos substratos que formam o conjunto de treinamento.

Dessa forma, criou-se uma etapa anterior a classificação, em que foram construídos três modelo hierárquicos: um para cada conjunto de substratos representados nos três conjuntos de treinamento descritos no tópico 5.2 (CT1, CT2 e CT3). Esses modelos possuíram um único nó onde a regra de decisão foi o Q-Estatístico obtido através da PCA construída usando os espectros dos substratos usados em cada conjunto de treinamento, conforme pode ser visto no fluxograma da Figura 35.

Figura 35 – Modelo Hierárquico para o protocolo de avaliação da representatividade dos pisos.



Os modelos hierárquicos criados para esse protocolo foram capazes de identificar se os substratos do conjunto de predição poderiam ser previstos pela variabilidade dos substratos do conjunto de treinamento.

Ao fazer a leitura dos espectros do piso CE5 no modelo de PCA construído com os espectros dos substratos das mesmas classes do conjunto CT2, ele foi capaz de identificar que os espectros de CE5 não eram representados por esse conjunto. De maneira análoga, os espectros de CE4, CE6, e PO5 foram submetidos ao protocolo dos substratos formados pelas classes de CT1, e a resposta obtida para todos esses pisos foi que eles eram representativos.

Dessa forma, uma estratégia para aumentar a confiabilidade no uso do modelo hierárquico para identificação de sangue humano é utilizar esse protocolo como uma garantia de que o piso onde o vestígio foi encontrado tem a variabilidade representada pelo modelo.

5 CONCLUSÕES E PERSPECTIVAS FUTURAS

A partir da construção dos modelos de PCA para o conjunto total de dados foi possível constatar que apesar da influência dos substratos, os escores de SH, SA e FPC apresentavam uma tendência de separação que permitiu o desenvolvimento de um modelo hierárquico único para identificação e classificação de vestígios de sangue em diferentes pisos de cerâmica e porcelanato.

O modelo hierárquico construído a partir dos conjuntos de treinamento com uma variabilidade mais abrangente e utilizando um modelo PLS-DA como segunda regra de decisão para classificar o sangue humano foi o que apresentou os melhores resultados de sensibilidade e especificidade, ambos iguais a 1. Esses mesmos valores foram obtidos pelos modelos construídos para os conjuntos com a faixa espectral reduzida. No entanto, deve-se ressaltar que o modelo PLS-DA é um modelo discriminante e, por esse motivo, para criar a fronteira da classe de SH ele utiliza as informações dos espectros da classe SA. Dessa forma, é importante destacar que os modelos construídos a partir dos conjuntos de treinamento com uma variabilidade mais abrangente e usando um modelo PCA ou SIMCA para discriminar as amostras de SH e SA, também obtiveram excelentes resultados de sensibilidade para classificação do SH, com valores entre 0,94, 0,97 e 1. Diferentemente do PLS-DA, como já mencionado, esses modelos utilizam as informações da classe de interesse, que nesse caso é o SH.

Deve-se ressaltar que o nosso principal interesse é identificar as amostras de sangue, visto que em um contexto forense onde um possível vestígio de sangue foi encontrado é mais importante sua identificação e coleta para testes posteriores, ainda que seja sangue de origem animal. Nesse sentido, uma especificidade mais baixa para a classe SH não seria um grande problema, visto que todos os modelos obtiveram 100% de classificação correta para as amostras de FPC.

Os modelos construídos por conjuntos de treinamento mais abrangentes tiveram melhor desempenho. Isso evidenciou a importância da escolha do conjunto de treinamento e a necessidade de avaliar a representatividade do substrato o que levou ao desenvolvimento de uma metodologia para avaliação prévia do substrato. Essa metodologia se mostrou bastante efetiva e possibilitou que o analista avaliasse a confiabilidade do resultado fornecido pelo modelo construído de acordo com a representatividade do substrato em que o vestígio foi depositado.

Tendo em vista os resultados obtidos, é possível concluir que os modelos hierárquicos apresentam um avanço bastante significativo para o desenvolvimento de uma metodologia

robusta e confirmatória para identificação das manchas de sangue humano em cenas de crime, com o mínimo de interferência nas evidências e conhecimento técnico.

Como perspectivas para continuidade do trabalho propomos avaliação da capacidade do modelo construído para identificar amostras envelhecidas com tempos de deposição diferentes e/ou construídos modelos que incluam no conjunto de treinamento diferentes tempos de deposição.

Outras faixas espectrais presentes na região do infravermelho próximo podem ser utilizadas para identificar componentes do sangue. Propõe-se avaliar equipamento com faixa espectral ampliada até 2500 nm, abrangendo outras regiões que não foram aplicadas no projeto atual. Além disso, propõe-se também incluir no estudo novas amostras de sangue de outras espécies animais, e avaliar a capacidade do modelo de identificar amostras com diferentes concentrações de sangue e de falsos positivos comuns.

Uma outra proposta é avaliar outros substratos comuns em cenas de crime, como tecidos. Além disso, pretende-se também incluir no estudo novas amostras de sangue de outras espécies animais, e avaliar a capacidade do modelo de identificar amostras com diferentes concentrações de sangue e de falsos positivos comuns.

REFERÊNCIAS

BALLABIO, D.; CONSONNI, V. Classification tools in chemistry. Part 1: linear models. PLS-DA. **Analytical Methods**, n. June 2009, p. 3790–3798, 2013.

BALLABIO, D.; GRISONI, F.; TODESCHINI, R. Multivariate comparison of classification performance measures. **Chemometrics and Intelligent Laboratory Systems**, v. 174, n. April 2017, p. 33–44, 2018.

BEEBE, K. R.; PELL, R. J.; SEASHOLTZ, M. B. **Chemometrics: A Practical GuideTechnometrics**John Wiley & Sons Ltd, , 1998.

BEVEL, T.; M. GARDNER, R. **Bloodstain Pattern Analysis with an Introduction to Crime Scene Reconstruction**. [s.l: s.n.].

BREMMER, R. H. et al. Forensic quest for age determination of bloodstains. **Forensic Science International**, v. 216, n. 1–3, p. 1–11, 2012.

BRERETON, R. G. **Chemometrics: Data Analysis for the Laboratory and Chemical Plant**. Bristol, UK: John Wiley & Sons Ltd, 2003. v. 8

BRO, R. et al. Cross-validation of component models: A critical look at current methods. **Analytical and Bioanalytical Chemistry**, v. 390, n. 5, p. 1241–1251, 2008.

BRO, R.; SMILDE, A. K. Principal component analysis. **Analytical Methods**, v. 6, n. 9, p. 2812–2831, 2014.

BURNS, D. A.; CIURCZAK, E. W. **Handbook of Near-Infrared Analysis**. 3rd. ed. [s.l.] CRC Press: Taylor & Francis Group, 2009.

CADD, S. et al. Age Determination of Blood-Stained Fingerprints Using Visible Wavelength Reflectance Hyperspectral Imaging. **Journal of Imaging**, p. 5–13, 2018.

CASALI, F. et al. Validation of presumptive tests for non-human blood using Kastle Meyer and Hemastix reagents. **Science and Justice**, v. 60, n. 1, p. 30–35, 2020.

CERQUEIRA, D. et al. **Atlas da Violência 2019**. Brasília: Instituto de Pesquisa Econômica Aplicada (IPEA), 2019.

CHRISTIE, A. **The Murder of Roger Ackroyd**. Kindle ed. London, UK: Harpercollins, 2013.

EDELMAN, G. et al. Identification and age estimation of blood stains on colored backgrounds by near infrared spectroscopy. **Forensic Science International**, v. 220, n. 1–3, p. 239–244, 2012.

EDELMAN, G. J.; VAN LEEUWEN, T. G.; AALDERS, M. C. Visualization of latent blood stains using visible reflectance hyperspectral imaging and chemometrics. **Journal of Forensic Sciences**, v. 60, n. s1, p. S188–S192, 2015.

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

EIGENVECTOR. **Hierarchical Model Builder**. Disponível em: <http://wiki.eigenvector.com/index.php?title=Hierarchical_Model_Builder#Introduction_to_Node_Types>. Acesso em: 13 maio. 2020.

ELKINS, K. M. **Introduction to Forensic Chemistry**. [s.l.] Taylor & Francis Group, 2019.

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

FERREIRA, M. M. C. **Quimiometria: Conceitos, Métodos e Aplicações**. [s.l.] Editora da Unicamp, 2015.

GEMPERLINE, P. **Practical Guide to Chemometrics**. Boca Raton, FL: CRC Press: Taylor & Francis Group, 2006.

HAYASHI, S. et al. Acceleration effect of the forensic luminol reaction induced by visible light irradiation of whole human blood aqueous solutions. **Forensic Science International**, v. 299, p. 208–214, 2019.

HOTELLING, H. The Generalization of Student's Ratio. p. 54–65, 1992.

KUMAR, R.; SHARMA, V. Chemometrics in Forensic Science. **Trends in Analytical Chemistry**, v. 105, p. 191–201, 2018.

LI, H. et al. Identification of blood species based on diffuse reflectance and transmission joint spectra with machine learning method. **Infrared Physics and Technology**, v. 88, p. 200–205, 2018.

LIN, H. et al. Estimation of the age of human bloodstains under the simulated indoor and outdoor crime scene conditions by ATR-FTIR spectroscopy. **Scientific Reports**, v. 7, n. 1, p. 1–9, 2017.

LÓPEZ, M. I.; CALLAO, M. P.; RUISÁNCHEZ, I. A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. **Analytica Chimica Acta**, v. 891, p. 62–72, 2015.

MALEGORI, C. et al. Identification of invisible biological traces in forensic evidences by hyperspectral NIR imaging combined with chemometrics. **Talanta**, v. 215, n. March, p. 120911, 2020.

MARINI, F. **Chemometrics in Food Chemistry**. 1st. ed. [s.l.] ELSEVIER, 2013. v. 53

MISTEK, E.; HALÁMKOVÁ, L.; LEDNEV, I. K. Phenotype profiling for forensic purposes : Nondestructive potentially on scene attenuated total reflection Fourier transform-infrared (ATR FT-IR) spectroscopy of bloodstains. **Forensic Chemistry**, v. 16, n. April, p. 100176, 2019.

MISTEK, E.; LEDNEV, I. K. Identification of species' blood by attenuated total reflection (ATR) Fourier transform infrared (FT-IR) spectroscopy. **Analytical and bioanalytical chemistry**, v. 407, n. 24, p. 7435–7442, 2015.

MURO, C. K. et al. Forensic body fluid identification and differentiation by Raman spectroscopy. **Forensic Chemistry**, v. 1, p. 31–38, 2016.

OLIVERI, P. Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues – A tutorial. **Analytica Chimica Acta**, v. 982, p. 9–19, 2017.

OLIVERI, P.; DOWNEY, G. **Discriminant and class-modelling chemometric techniques for food PDO verification**. 1. ed. [s.l.] Copyright © 2013 Elsevier B.V. All rights reserved., 2013. v. 60

OTTO, M. **Chemometrics: Statistics and Computer Application in Analytical Chemistry**. 3rd. ed. Leipziger, Germany: Wiley-VCH Verlag GmbH &Co, 2017.

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.

PEREIRA, J. F. Q. et al. Evaluation and identification of blood stains with handheld NIR spectrometer. **Microchemical Journal**, v. 133, p. 561–566, 2017.

PEREIRA, J. F. Q. **Espectroscopia no Infravermelho Próximo e Quimiometria em problemas forense: Identificação de manchas de sangue humano e plantações de Cannabis sativa L.** [s.l.] Universidade Federal de Pernambuco (UFPE), 2019.

REECE, J. B. et al. **Biologia de Campbell**. 10. ed. Porto Alegre: Artmed, 2015.

RINNAN, Å.; BERG, F. VAN DEN; 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, 2009.

ROSENBLATT, R. et al. Raman spectroscopy for forensic bloodstain identification : Method validation vs . environmental interferences. **Forensic Chemistry**, v. 16, n. August, p. 100175, 2019.

S. DE MARTINIS, B.; F. DE OLIVEIRA, M. **Química Forense Experimental**. São Paulo: Cencage Learning, 2016.

SHARMA, S.; CHOPHI, R.; SINGH, R. Forensic discrimination of menstrual blood and peripheral blood using attenuated total reflectance (ATR)-Fourier transform infrared (FT-IR) spectroscopy and chemometrics. **International Journal of Legal Medicine**, v. 134, n. 1, p. 63–77, 2020.

SHARMA, V.; KUMAR, R. Trends of chemometrics in bloodstain investigations. **Trends in Analytical Chemistry**, v. 107, p. 181–195, 2018.

SIEGEL, J. A. **Forensic Chemistry: Fundamentals and Applications**. [s.l.] Wiley Blackwell, 2016.

SILVA, C. S.; BRAZ, A.; PIMENTEL, M. F. Vibrational spectroscopy and chemometrics in forensic chemistry: Critical review, current trends and challenges. **Journal of the Brazilian Chemical Society**, v. 30, n. 11, p. 2259–2290, 2019.

SKOOG, D.; HOLLER, J.; CROUCH, S. **Principles of Instrumental Analysis**. 7th. ed. [s.l.] Cencage Learning, 2018.

TAKAMURA, A. et al. Soft and Robust Identification of Body Fluid Using Fourier Transform Infrared Spectroscopy and Chemometric Strategies for Forensic Analysis. **Scientific Reports**, v. 8, n. 1, p. 1–10, 2018.

TAKAMURA, A. et al. Comprehensive modeling of bloodstain aging by multivariate Raman spectral resolution with kinetics. **Communications Chemistry**, v. 2, n. 115, p. 1–10, 2019.

VIRKLER, K.; LEDNEV, I. K. Analysis of body fluids for forensic purposes: From laboratory testing to non-destructive rapid confirmatory identification at a crime scene. **Forensic Science International**, v. 188, n. 1–3, p. 1–17, 2009a.

VIRKLER, K.; LEDNEV, I. K. Blood Species Identification for Forensic Purposes Using Raman Spectroscopy Combined with Advanced Statistical Analysis. **Analytical Chemistry**, v. 81, n. 18, p. 7773–7777, 2009b.

VIRKLER, K.; LEDNEV, I. K. Raman spectroscopic signature of blood and its potential application to forensic body fluid identification. **Analytical and bioanalytical chemistry**, p. 525–534, 2010.

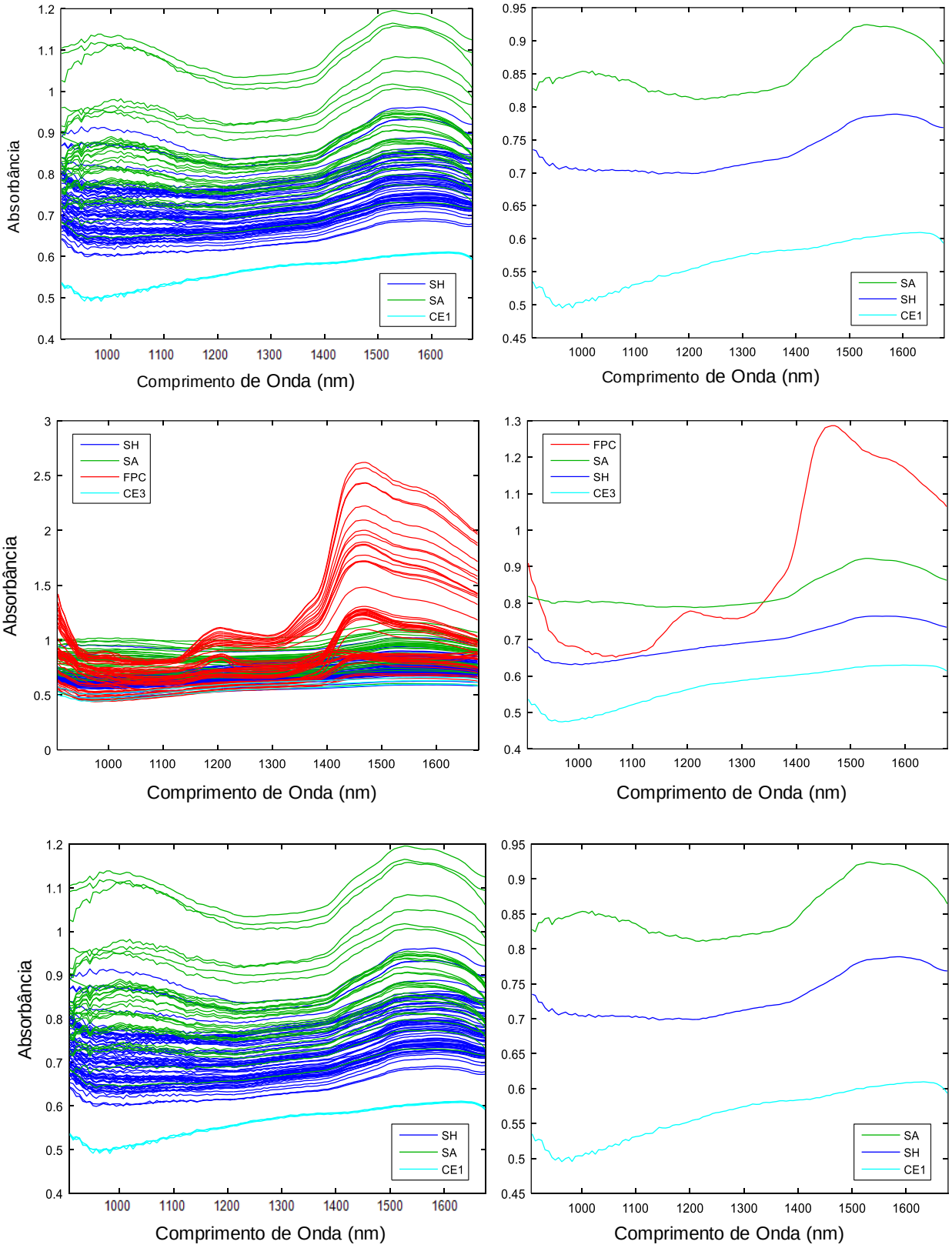
WORKMAN, J.; WEYER, L. **Practical Guide to Interpretive Near-Infrared Spectroscopy**. [s.l.] CRC Press: Taylor & Francis Group, 2009.

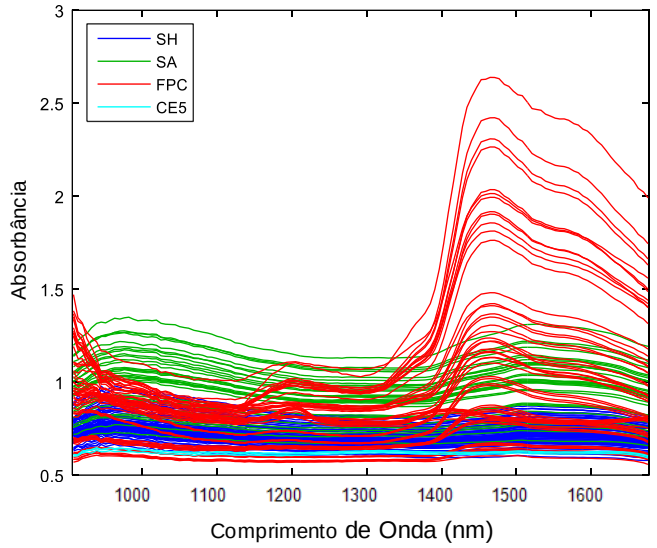
ZADORA, G.; MENZYK, A. In the pursuit of the holy grail of forensic science e Spectroscopic studies on the estimation of time since deposition of bloodstains. **Trends in Analytical Chemistry**, v. 105, p. 137–165, 2018.

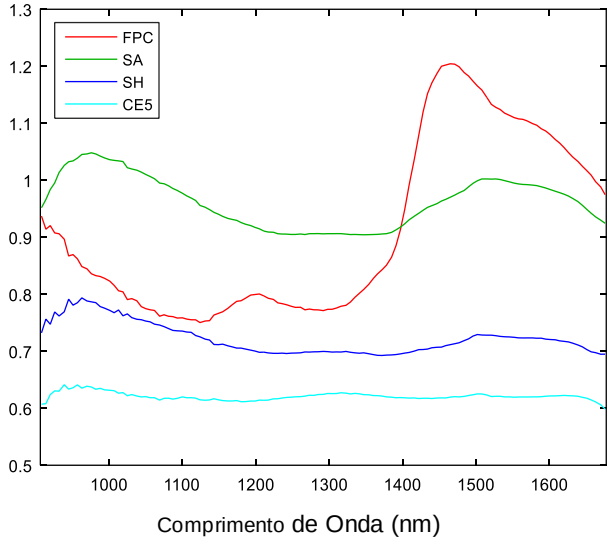
ZHANG, L. et al. Blood species identification using Near-Infrared diffuse transmitted spectra and PLS-DA method. **Infrared Physics and Technology**, v. 76, p. 587–591, 2016.

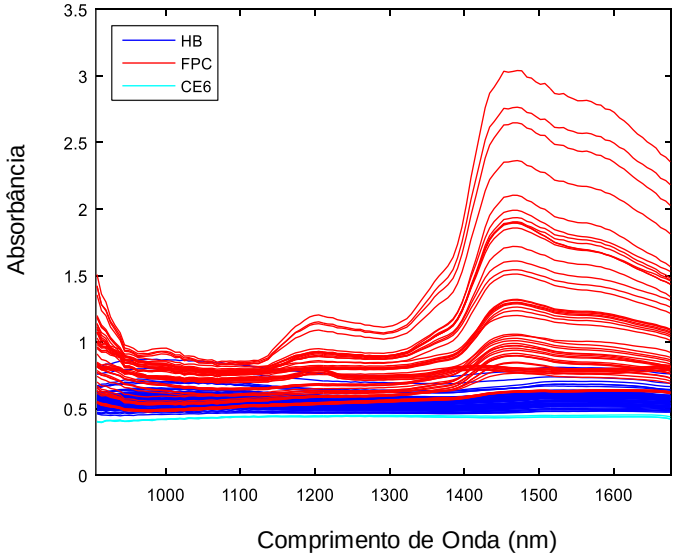
ZOU, Y. et al. Whole blood and semen identification using mid-infrared and Raman spectrum analysis for forensic applications. **Analytical Methods**, v. 8, n. 18, p. 3763–3767, 2016.

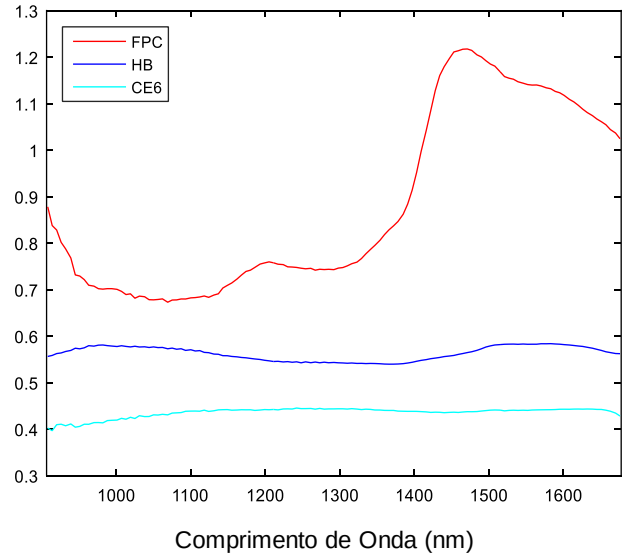
APÊNDICE A - ESPECTROS ORIGINAIS (COLUNA DA ESQUERDA) E MÉDIA DOS ESPECTROS ORIGINAIS (COLUNA DA DIREITA) DE TODAS AS AMOSTRAS SOB OS DIFERENTES SUBSTRATOS. ESPECTROS DE SH (AZUL), ESPECTROS DE SA (VERDE), ESPECTROS DE FPC (VERMELHO), ESPECTROS DO SUBSTRATO PURO (AZUL) .

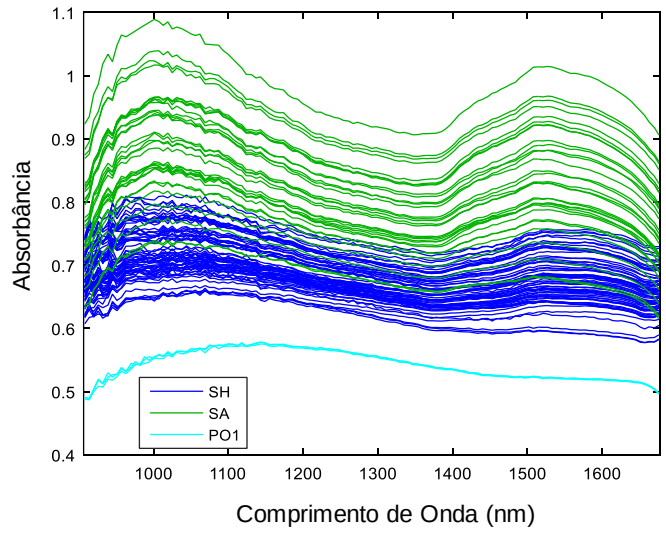


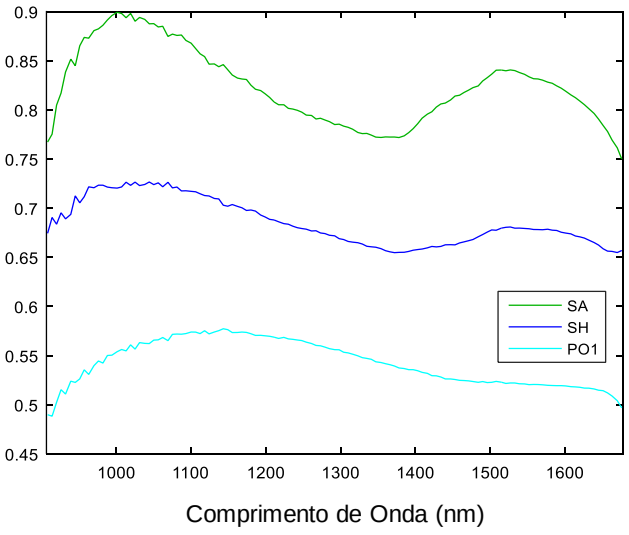

SH
SA
FPC
CE5
Absorbância
Comprimento de Onda (nm)

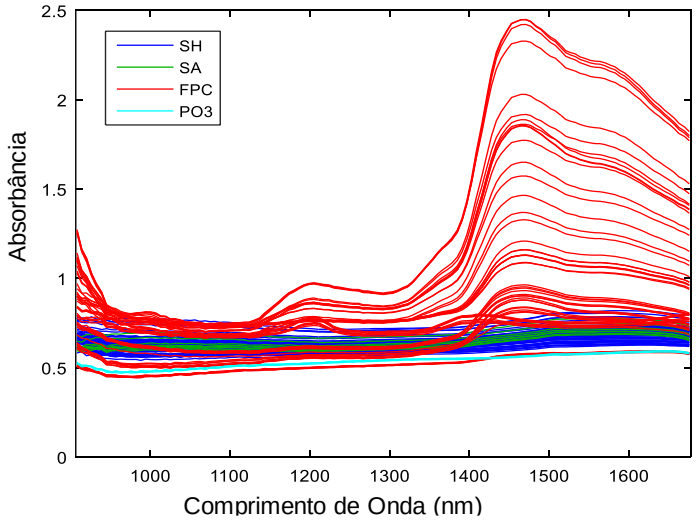

FPC
SA
SH
CE5
Comprimento de Onda (nm)

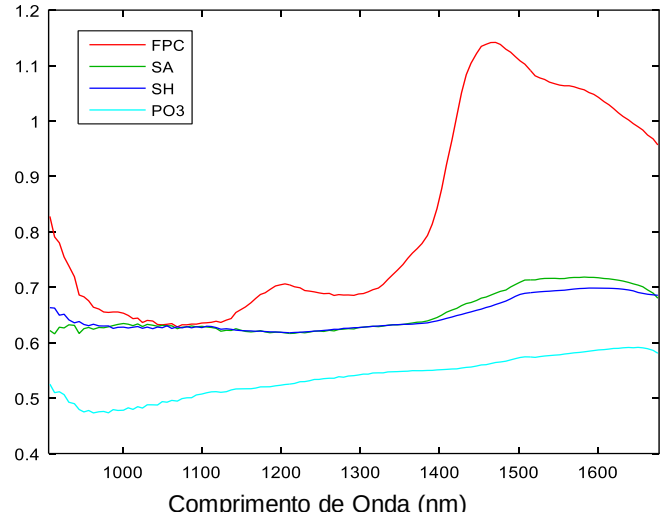

HB
FPC
CE6
Absorbância
Comprimento de Onda (nm)

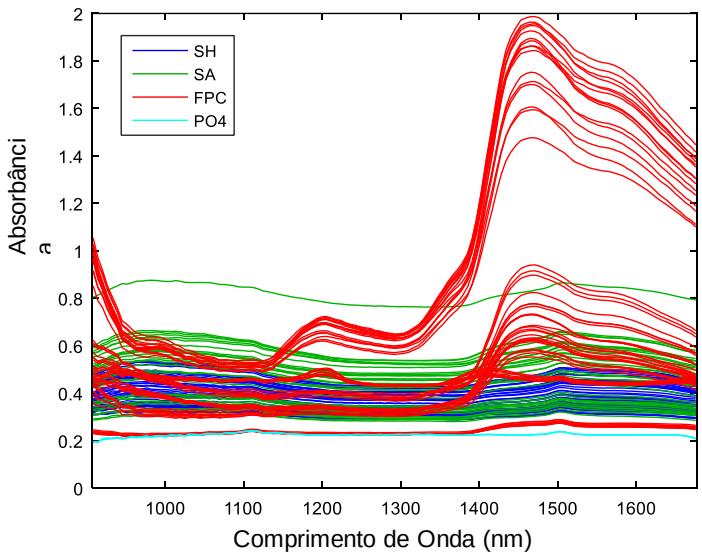

FPC
HB
CE6
Comprimento de Onda (nm)

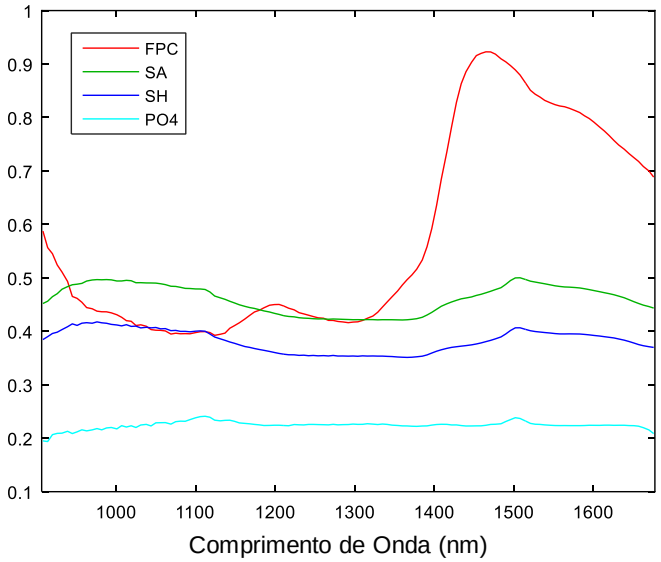

SH
SA
PO1
Absorbância
Comprimento de Onda (nm)

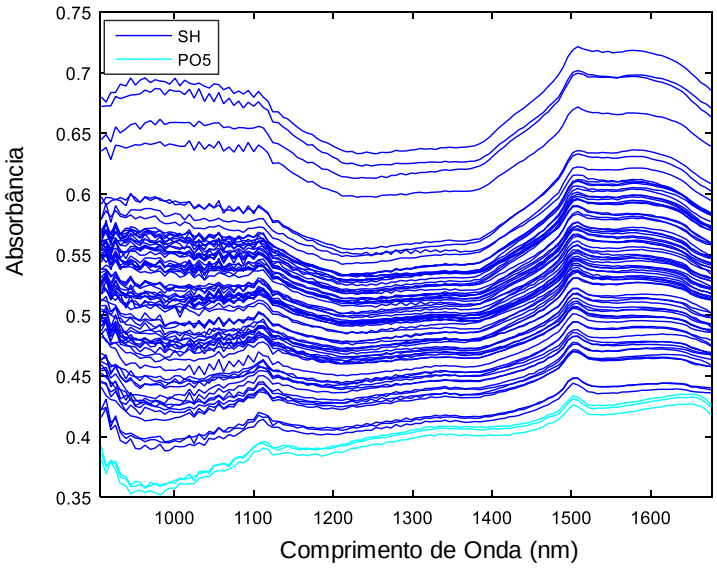

SA
SH
PO1
Comprimento de Onda (nm)

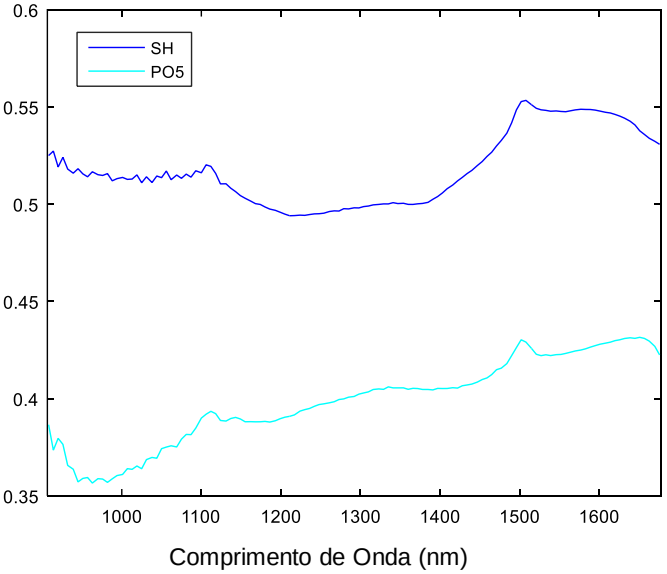

SH
SA
FPC
PO3
Absorbância
Comprimento de Onda (nm)


FPC
SA
SH
PO3
Comprimento de Onda (nm)

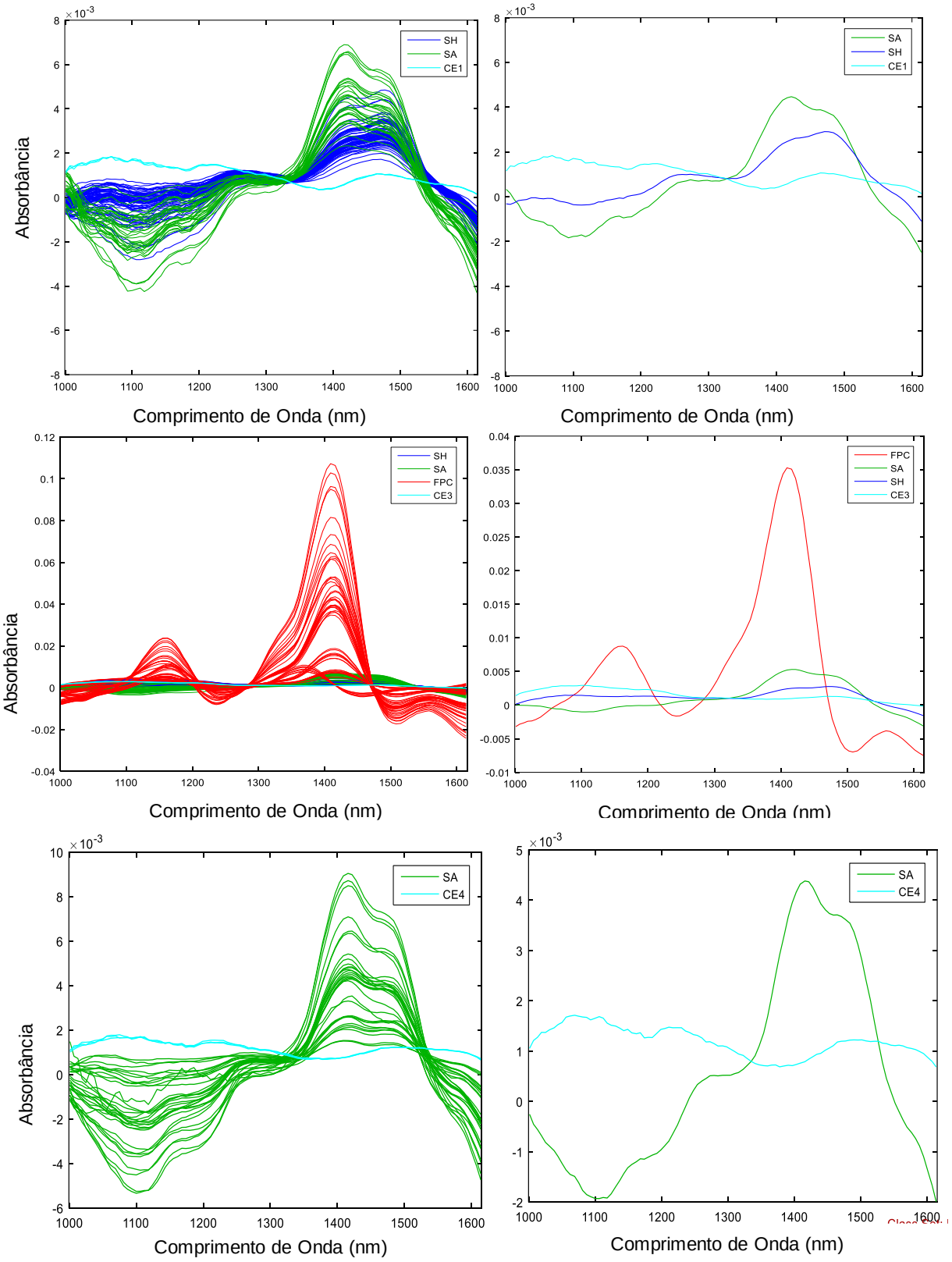

SH
SA
FPC
PO4
Absorbância
Comprimento de Onda (nm)

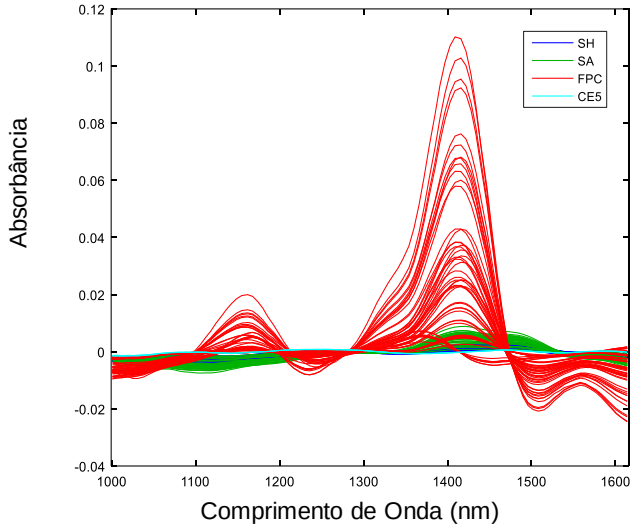

FPC
SA
SH
PO4
Comprimento de Onda (nm)

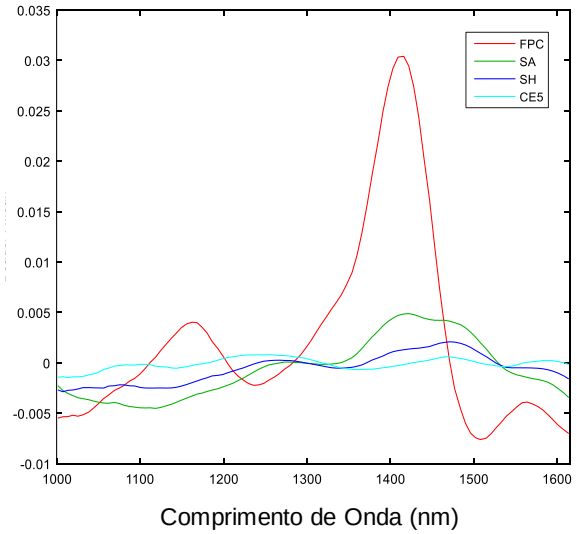

SH
PO5
Absorbância
Comprimento de Onda (nm)

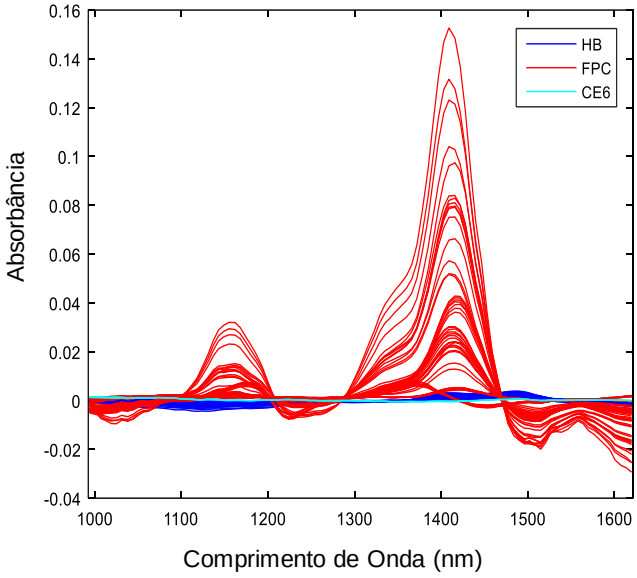

SH
PO5
Comprimento de Onda (nm)

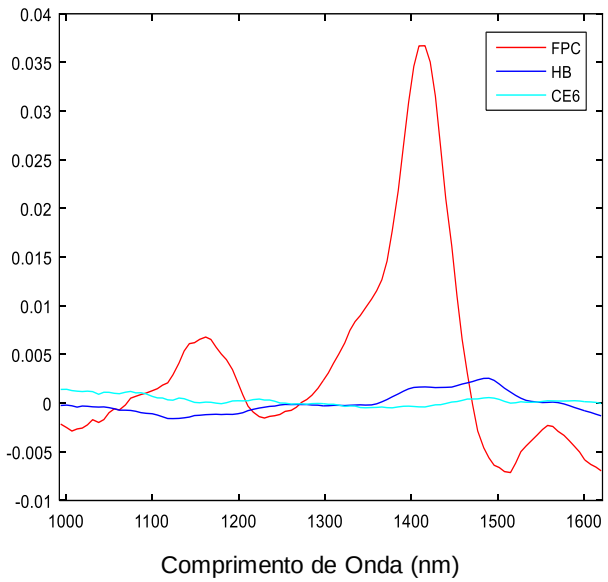
APÊNDICE B - ESPECTROS PRÉ-PROCESSADOS (COLUNA DA ESQUERDA) E MÉDIA DOS ESPECTROS PRÉ-PROCESSADOS (COLUNA DA DIREITA) DE TODAS AS AMOSTRAS SOB OS DIFERENTES SUBSTRATOS. ESPECTROS DE SH (AZUL), ESPECTROS DE SA (VERDE), ESPECTROS DE FPC (VERMELHO), ESPECTROS DO SUBSTRATO PURO (AZUL) PARA CADA UM DOS SUBSTRATOS.

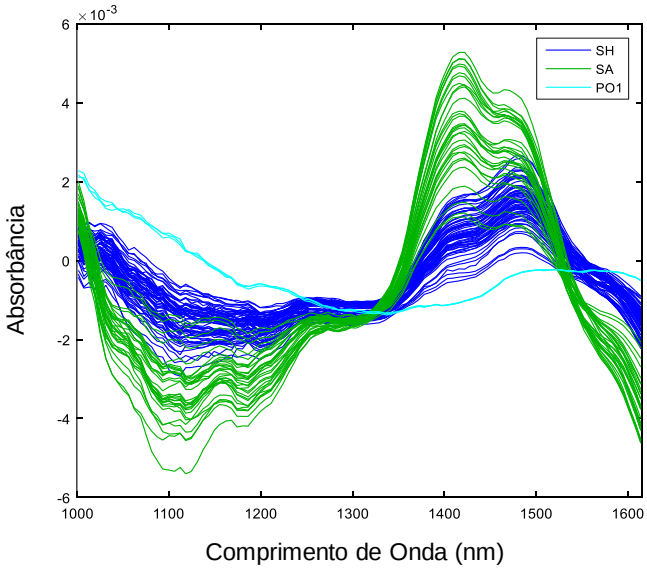


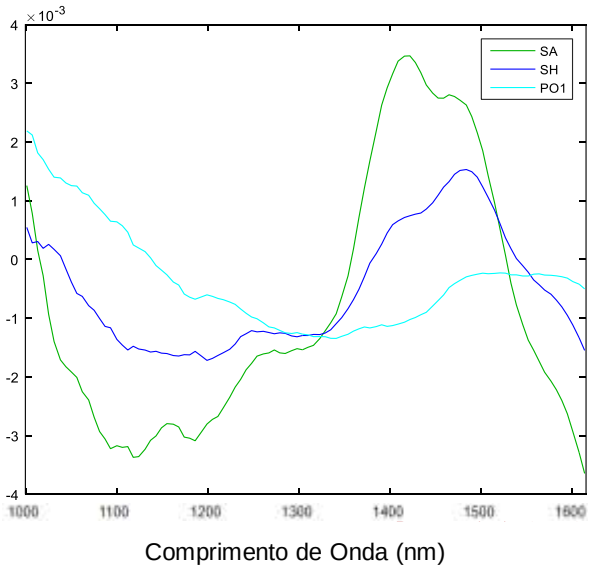

Absorbância
Comprimento de Onda (nm)
SH
SA
FPC
CE5

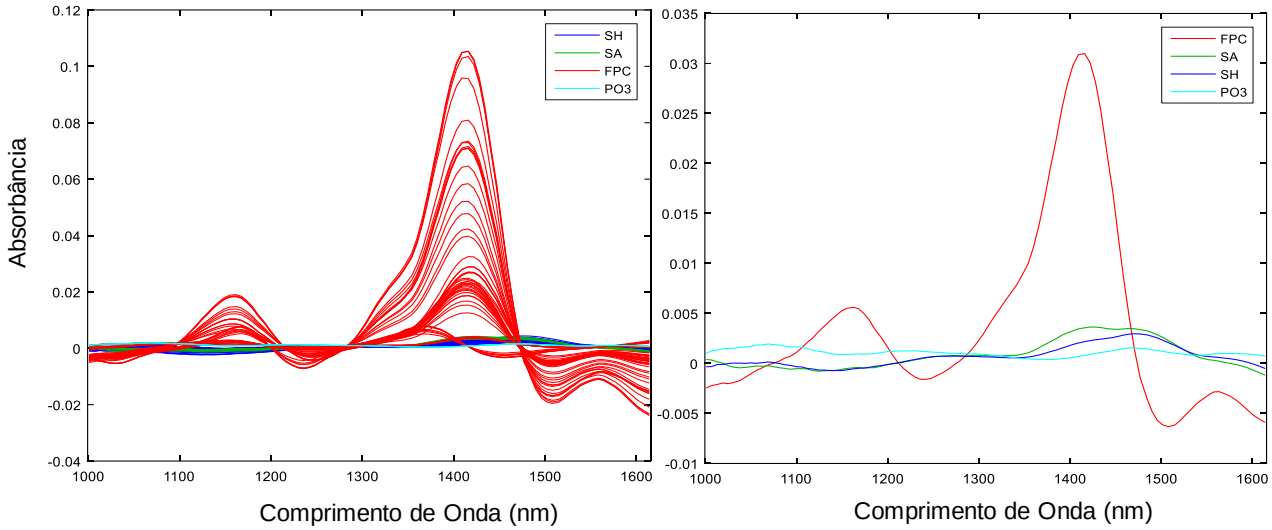

Comprimento de Onda (nm)
FPC
SA
SH
CE5

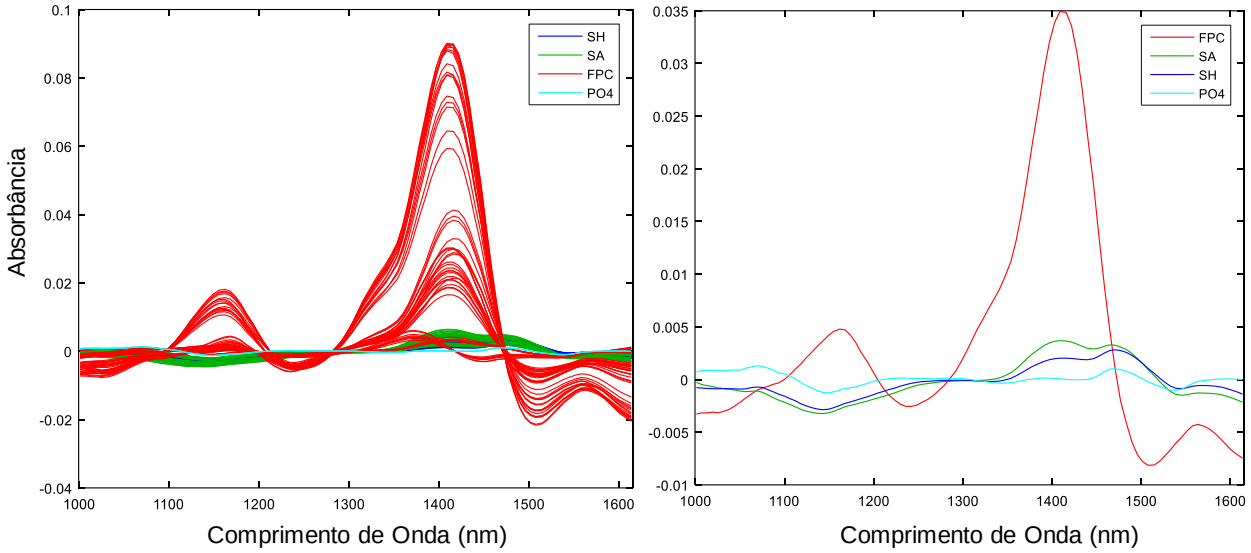

Absorbância
Comprimento de Onda (nm)
HB
FPC
CE6

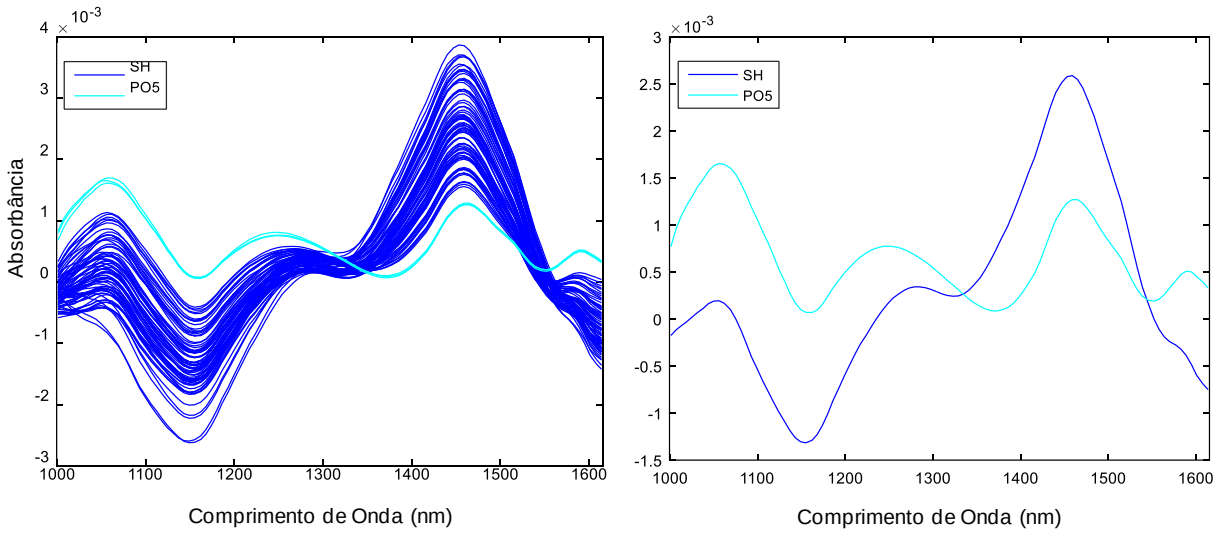

Comprimento de Onda (nm)
FPC
HB
CE6


Absorbância
Comprimento de Onda (nm)
SH
SA
PO1

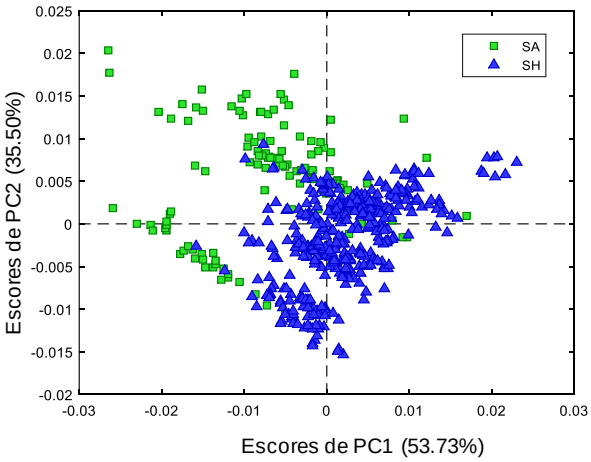

Comprimento de Onda (nm)
SA
SH
PO1

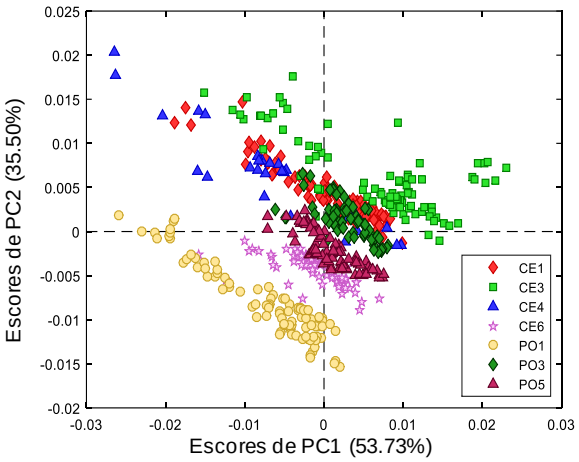

Absorbância
Comprimento de Onda (nm)
SH
SA
FPC
PO3
Comprimento de Onda (nm)
FPC
SA
SH
PO3

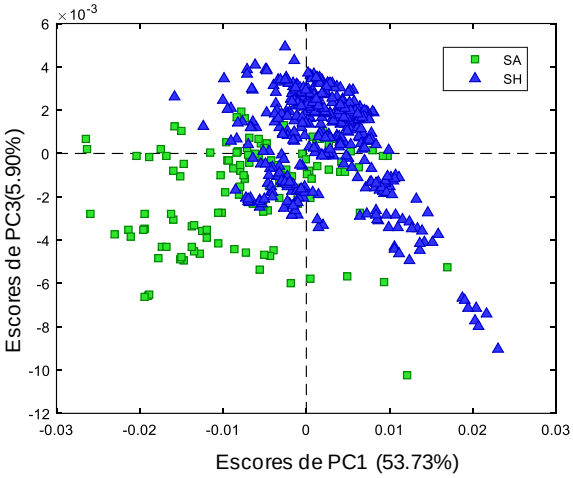

Absorbância
Comprimento de Onda (nm)
SH
SA
FPC
PO4
Comprimento de Onda (nm)
FPC
SA
SH
PO4

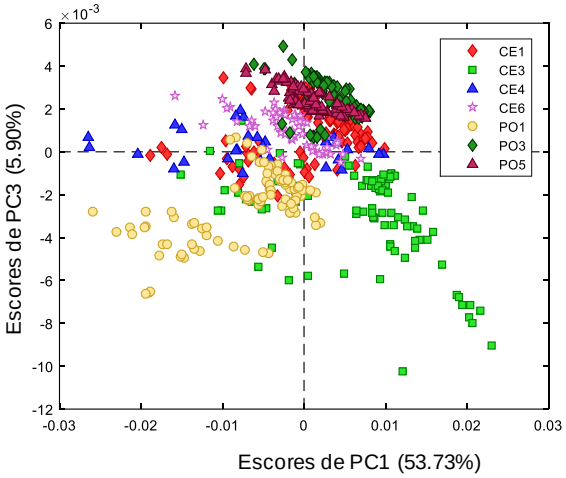

Absorbância
Comprimento de Onda (nm)
SH
PO5
Comprimento de Onda (nm)
SH
PO5

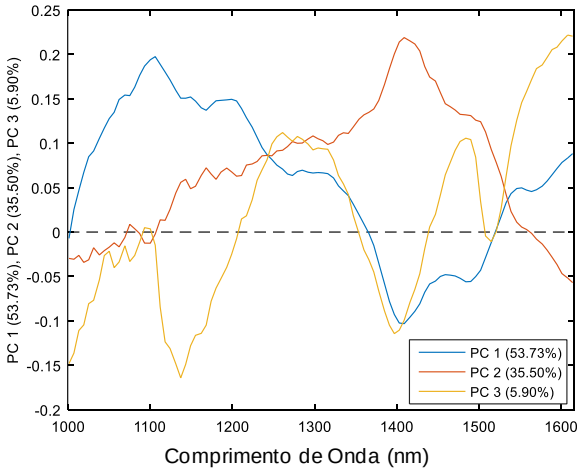
APÊNDICE C - GRÁFICOS DE ESCORES DE PC1XPC2 E PC1XPC3 E SEUS RESPECTIVOS *LOADINGS*, E GRÁFICOS DE INFLUÊNCIA CONSTRUÍDOS A PARTIR DAS AMOSTRAS DE SANGUE DO CONJUNTO DE TREINAMENTO CT2.

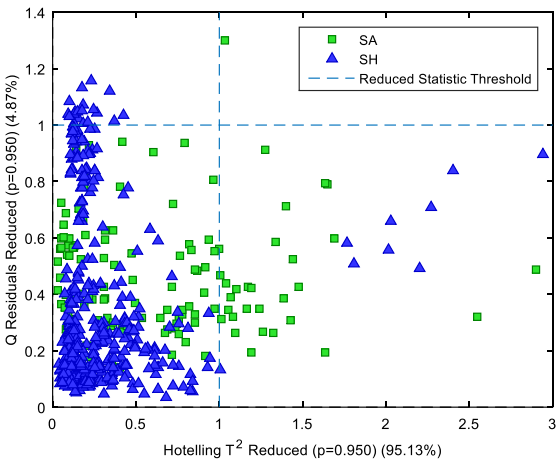


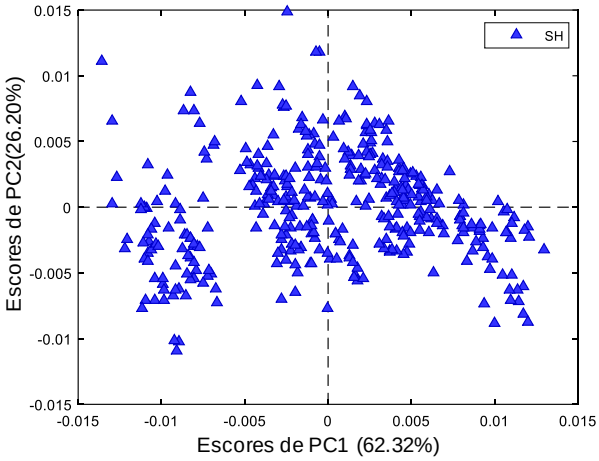


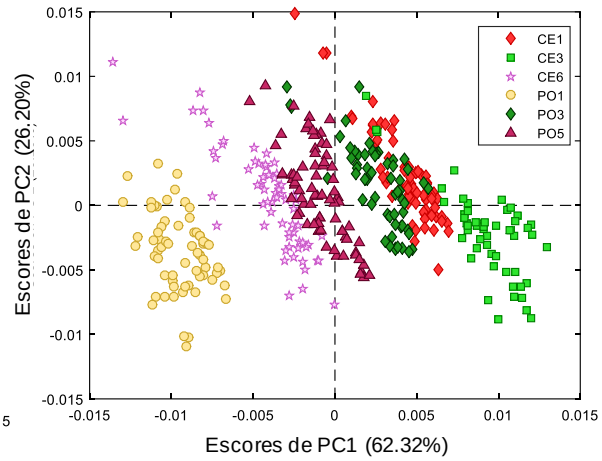


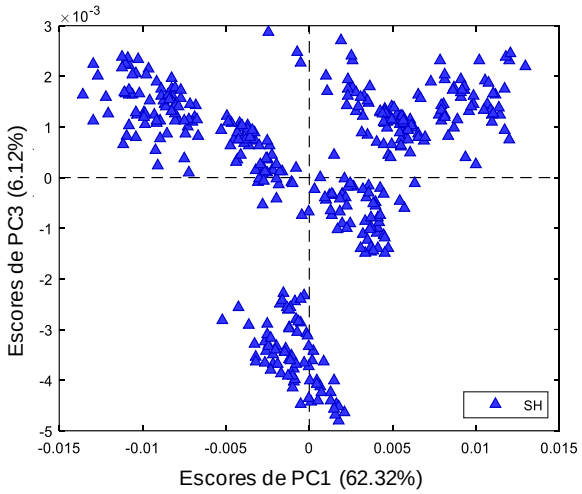


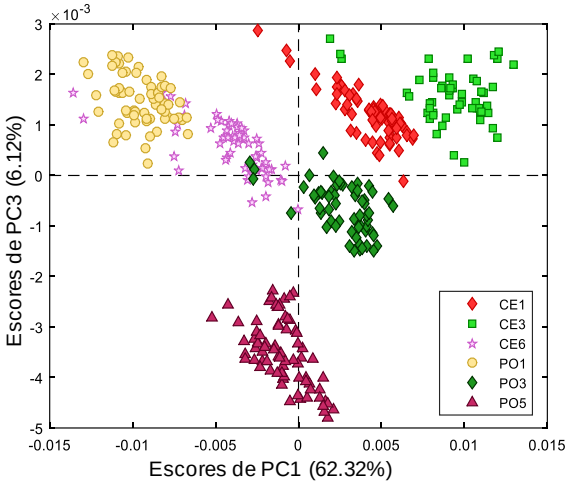


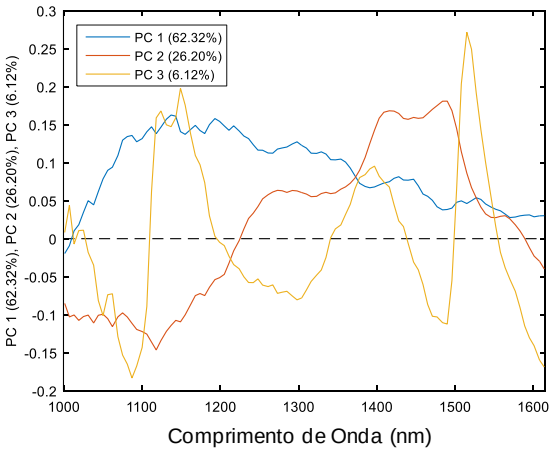


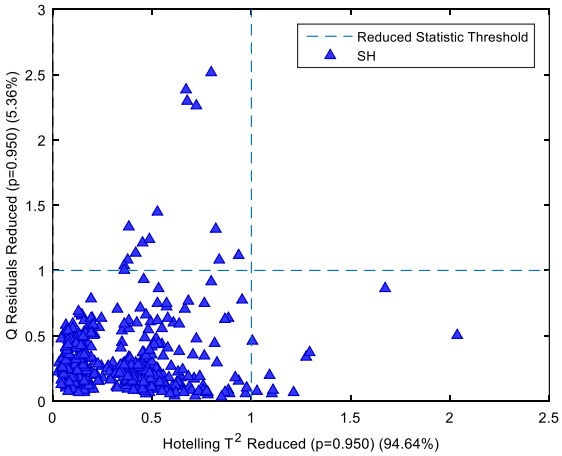

SH
Escores de PC2(26.20%)
Escores de PC1 (62.32%)


CE1
CE3
CE6
PO1
PO3
PO5
Escores de PC2 (26.20%)
Escores de PC1 (62.32%)

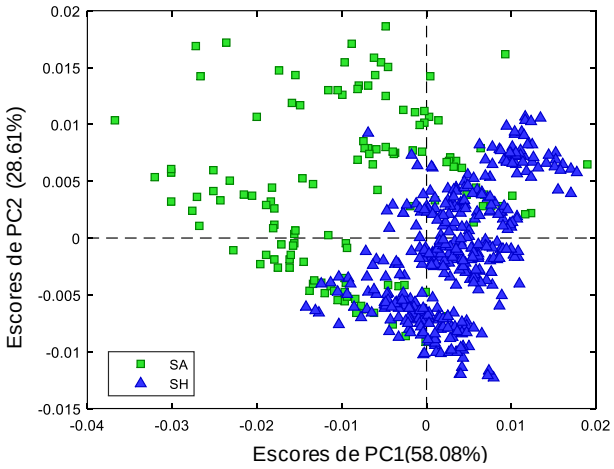

SH
Escores de PC3 (6.12%)
Escores de PC1 (62.32%)

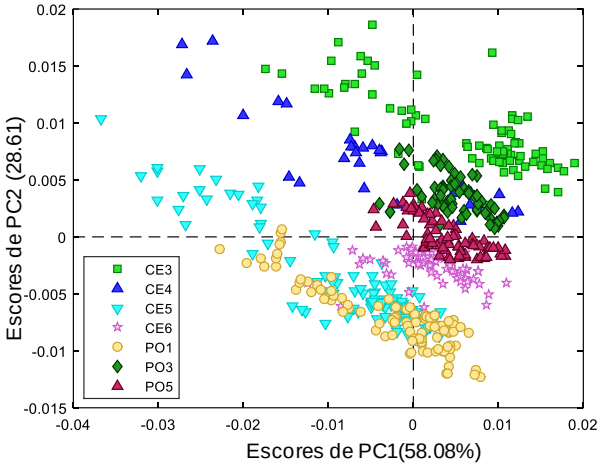

CE1
CE3
CE6
PO1
PO3
PO5
Escores de PC3 (6.12%)
Escores de PC1 (62.32%)

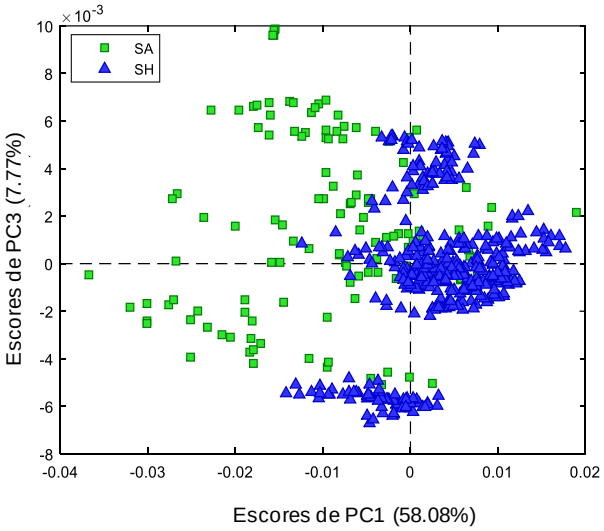

PC 1 (62.32%)
PC 2 (26.20%)
PC 3 (6.12%)
PC 1 (62.32%), PC 2 (26.20%), PC 3 (6.12%)
Comprimento de Onda (nm)

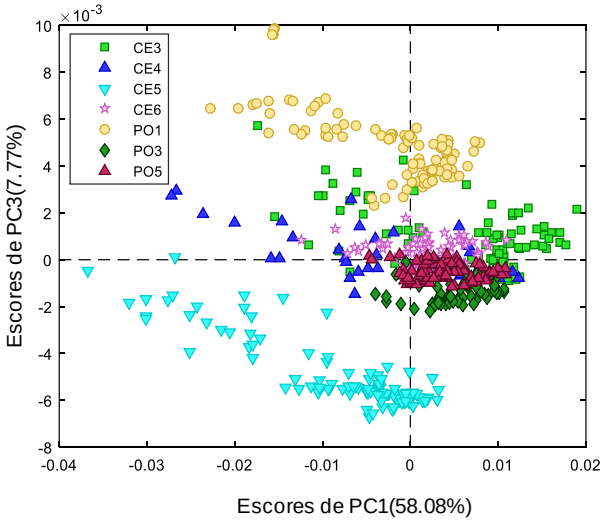

Reduced Statistic Threshold
SH
Q Residuals Reduced (p=0.950) (5.36%)
Hotelling T2 Reduced (p=0.950) (94.64%)

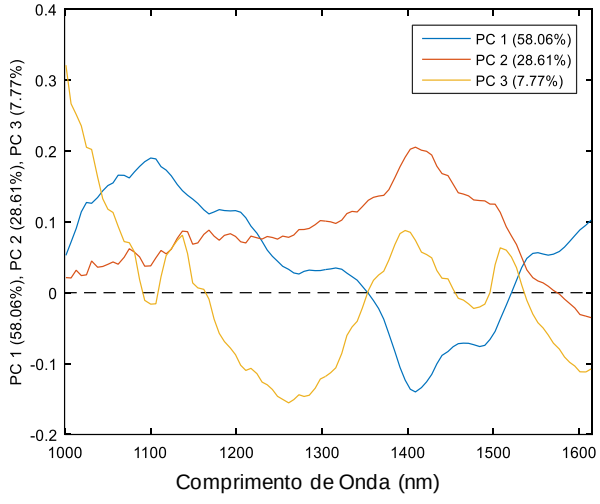
APÊNDICE D - GRÁFICOS DE ESCORES DE PC1XPC2 E PC1XPC3 E SEUS RESPECTIVOS *LOADINGS*, E GRÁFICO DE INFLUÊNCIA CONSTRUÍDOS A PARTIR DAS AMOSTRAS DE SANGUE DO CONJUNTO DE TREINAMENTO CT3.

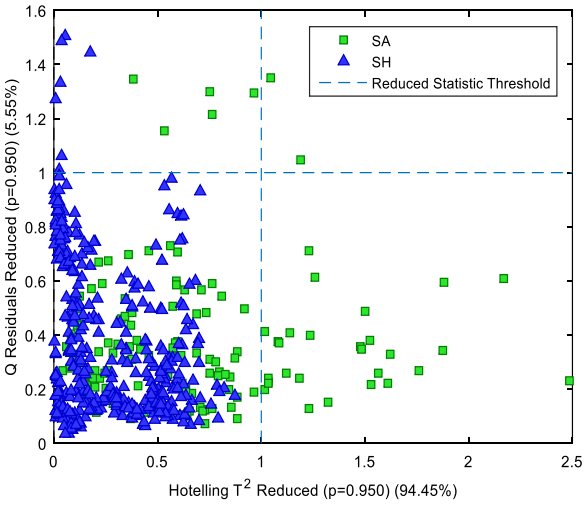


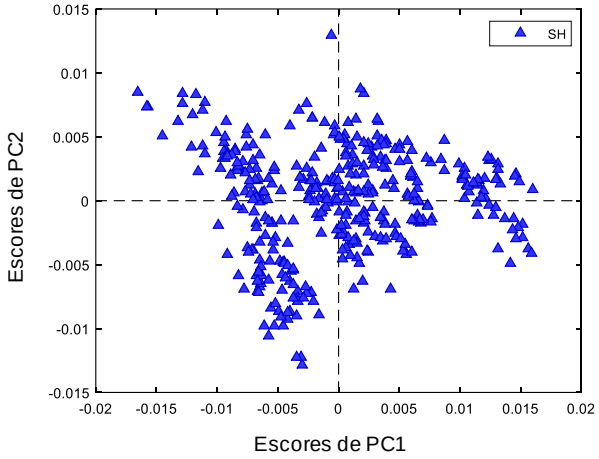


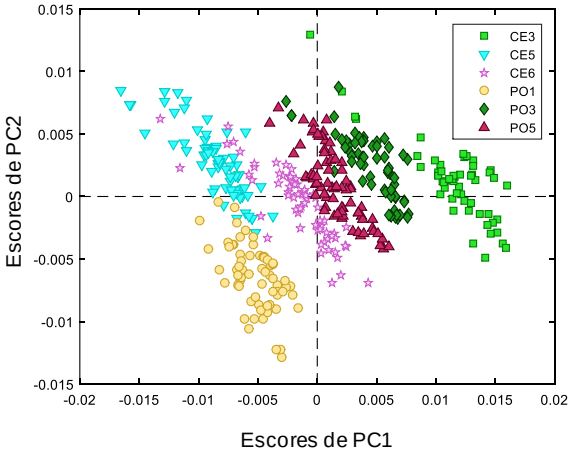


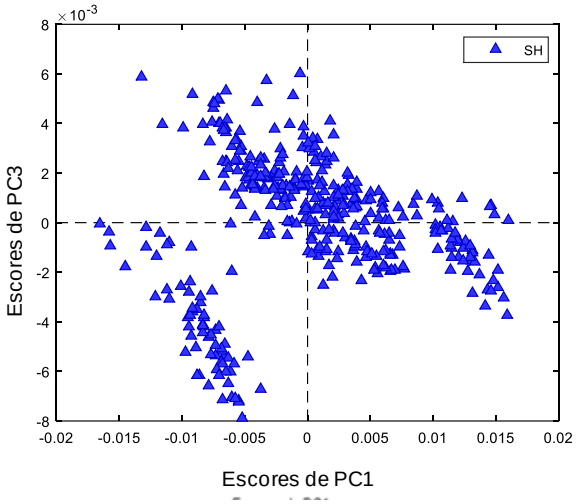


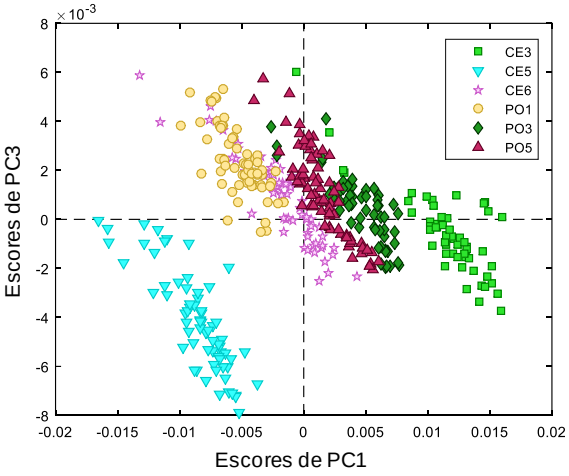


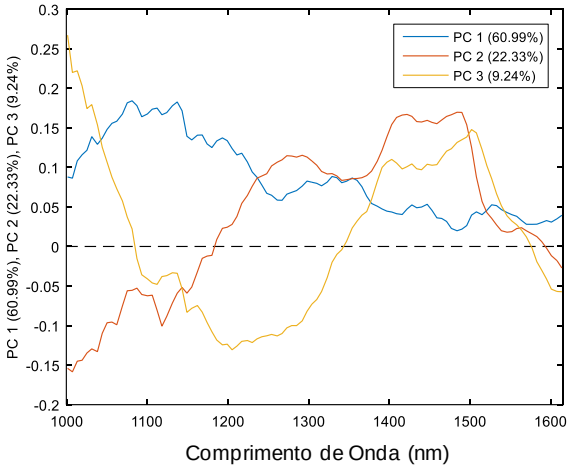


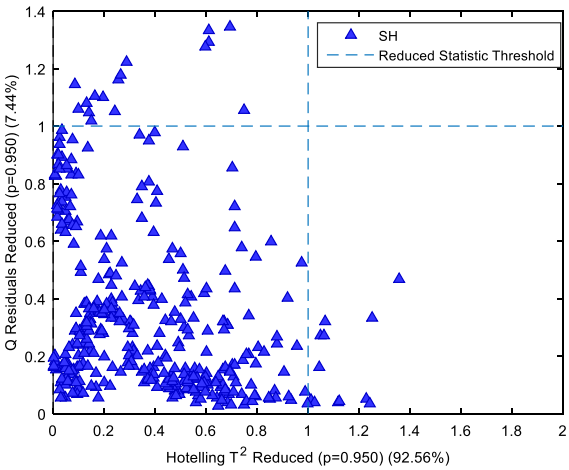

SH
Escores de PC2
Escores de PC1


CE3
CE5
CE6
PO1
PO3
PO5
Escores de PC2
Escores de PC1

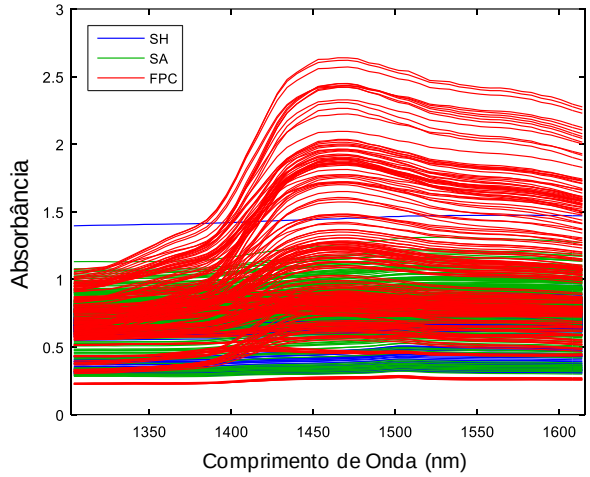

SH
Escores de PC3
Escores de PC1

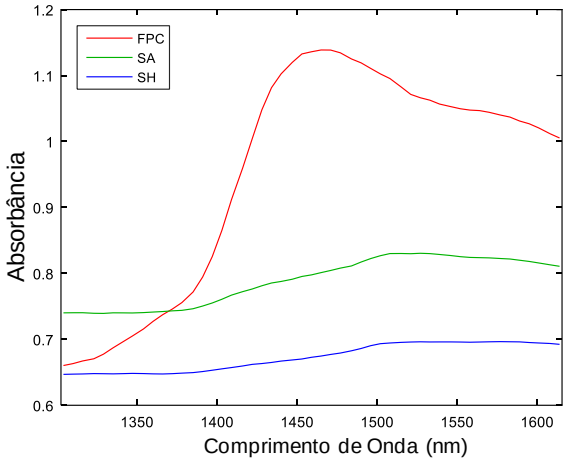

CE3
CE5
CE6
PO1
PO3
PO5
Escores de PC3
Escores de PC1

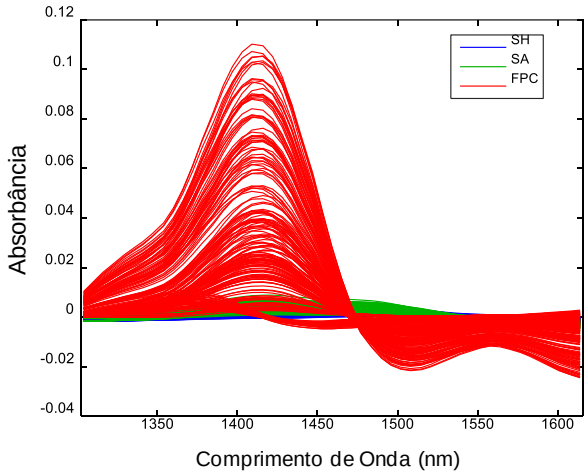

PC 1 (60.99%)
PC 2 (22.33%)
PC 3 (9.24%)
PC 1 (60.99%), PC 2 (22.33%), PC 3 (9.24%)
Comprimento de Onda (nm)

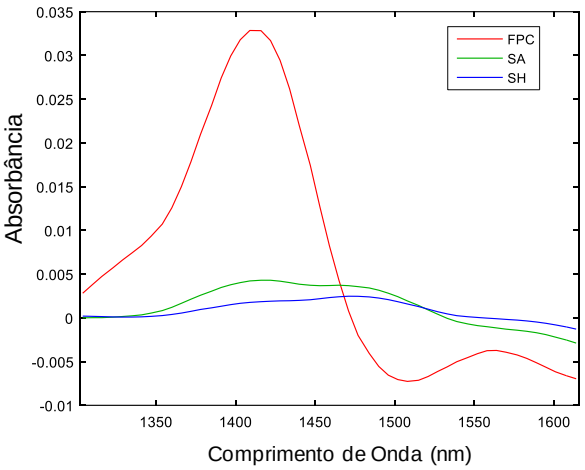

SH
Reduced Statistic Threshold
Q Residuals Reduced (p=0.950) (7.44%)
Hotelling T2 Reduced (p=0.950) (92.56%)

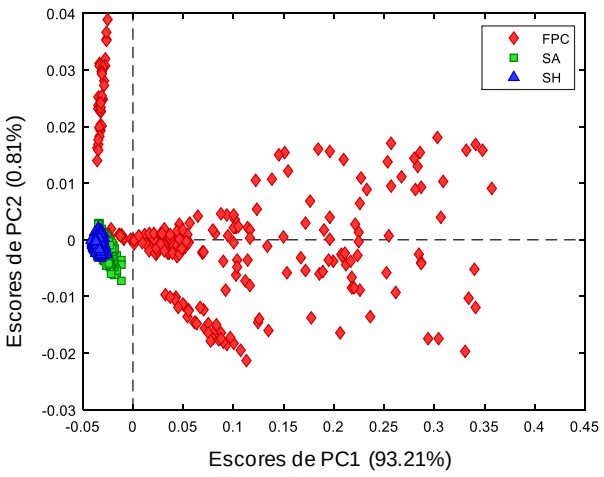
APÊNDICE E - GRÁFICOS DOS ESPECTROS ORIGINAIS, MÉDIA DOS ESPECTROS ORIGINAIS, GRÁFICOS DOS ESPECTROS PRÉ-PROCESSADOS E MÉDIA DOS ESPECTROS PRÉ-PROCESSADOS, GRÁFICOS DE ESCORES DE PC1XPC2 E PC1XPC3 E SEUS RESPECTIVOS *LOADINGS*, E GRÁFICO DE INFLUÊNCIA CONSTRUÍDOS A PARTIR DO CONJUNTO DE DADOS COM FAIXA ESPECTRAL REDUZIDA.

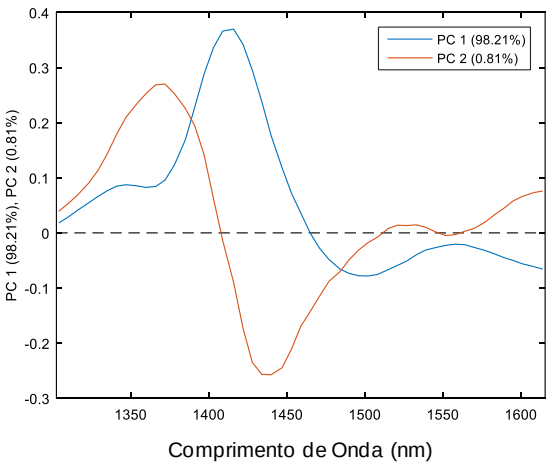


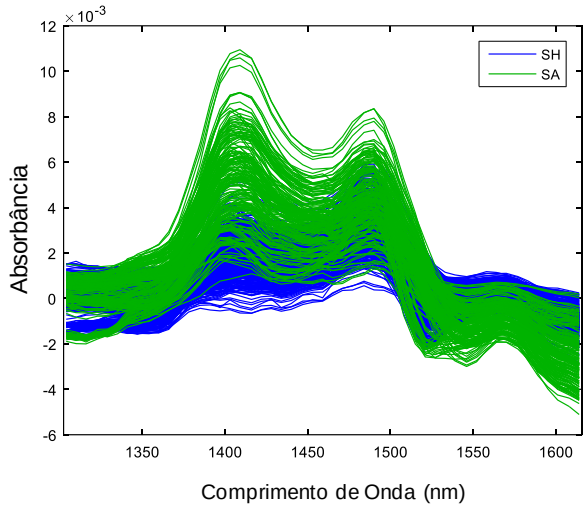


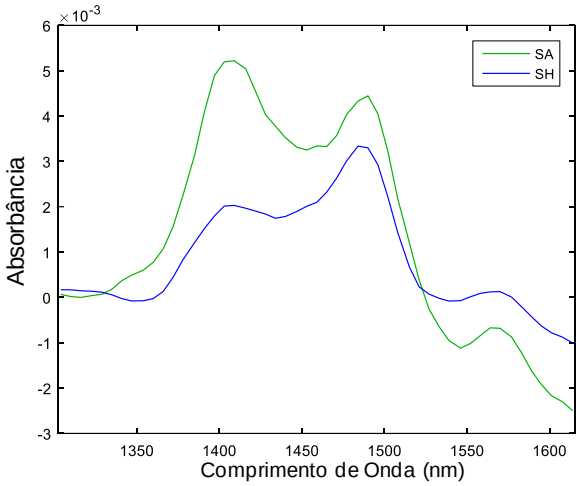


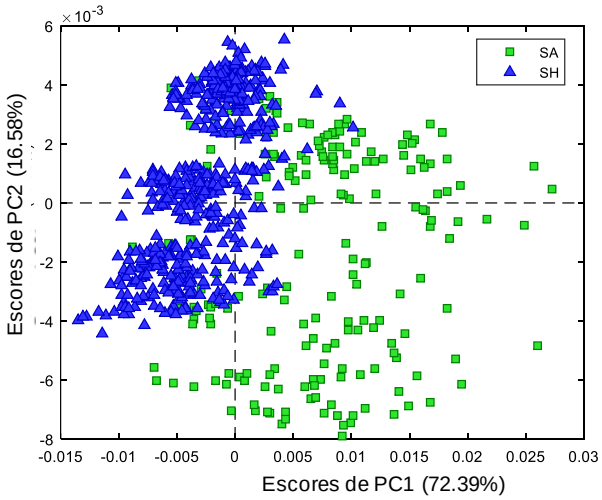


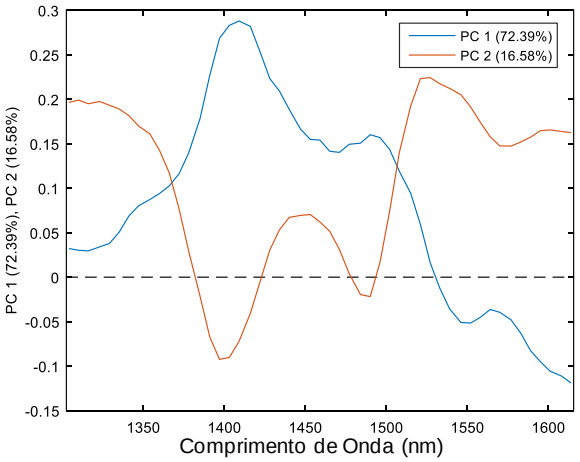


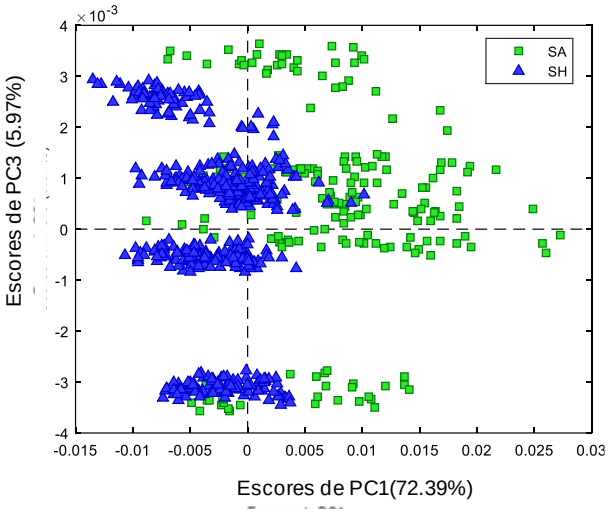


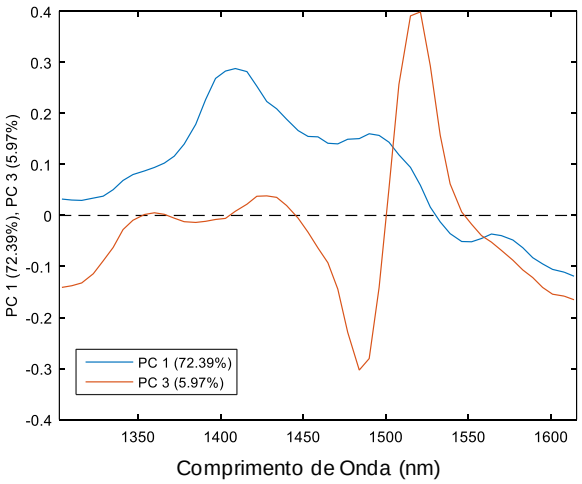

×10^-3
SH
SA
Absorbância
Comprimento de Onda (nm)


×10^-3
SA
SH
Absorbância
Comprimento de Onda (nm)


×10^-3
SA
SH
Escores de PC2 (16.58%)
Escores de PC1 (72.39%)


PC 1 (72.39%)
PC 2 (16.58%)
PC 1 (72.39%), PC 2 (16.58%)
Comprimento de Onda (nm)


×10^-3
SA
SH
Escores de PC3 (5.97%)
Escores de PC1(72.39%)


PC 1 (72.39%)
PC 3 (5.97%)
PC 1 (72.39%), PC 3 (5.97%)
Comprimento de Onda (nm)