UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS - ICEx
DEPARTAMENTO DE QUÍMICA

Larissa Batista dos Santos

APLICAÇÃO DE TÉCNICAS ESPECTROSCÓPICAS E MÉTODOS DE MODELAGEM DE CLASSE NA DISCRIMINAÇÃO GEOGRÁFICA DE GRÃOS DE CAFÉ VERDE DA REGIÃO DO CERRADO MINEIRO

Belo Horizonte
2022

UFMG/ICEX/DQ. 1.488
D. 809

Larissa Batista dos Santos

APLICAÇÃO DE TÉCNICAS ESPECTROSCÓPICAS E MÉTODOS DE MODELAGEM DE CLASSE NA DISCRIMINAÇÃO GEOGRÁFICA DE GRÃOS DE CAFÉ VERDE DA REGIÃO DO CERRADO MINEIRO

Dissertação apresentada ao Departamento de Química do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Química.

Orientadora: Prof.ª Dr.ª Mariana Ramos de Almeida

Belo Horizonte
2022

Ficha Catalográfica

S237a
2022
D

Santos, Larissa Batista dos
Aplicação de técnicas espectroscópicas e métodos de modelagem de classe na discriminação geográfica de grãos de café verde da região do Cerrado Mineiro [manuscrito] / Larissa Batista dos Santos. 2022.
99 f. : il., gráfs., tabs.

Orientadora: Mariana Ramos de Almeida.

Dissertação (mestrado) – Universidade Federal de Minas Gerais – Departamento de Química.
Bibliografia: f. 88-96.
Anexos: f. 97-99.

1. Química analítica – Teses. 2. Café – Teses. 3. Certificados de procedência – Teses. 4. Fluorescência de raio X – Teses. 5. Espectroscopia de infravermelho – Teses. 6. Espectrometria de massa – Teses. 7. Espectroscopia de absorção atômica – Teses. 8. Modelagem de dados – Teses. 9. Mínimos quadrados – Teses. I. Almeida, Mariana Ramos de, Orientadora. II. Título.

CDU 043

Elaborada por Sérgio Ferreira da Silva – CRB6-2719.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

UFMG
Programa de Pós-Graduação em Química
Departamento de Química - ICEx

química
pós-graduação-ufmg

"Aplicação de Técnicas Espectroscópicas e Métodos de Modelagem de Classe Na Discriminação Geográfica de Grãos de Café Verde da Região do Cerrado Mineiro"

Larissa Batista dos Santos

Dissertação aprovada pela banca examinadora constituída pelos Professores:

Profa. Mariana Ramos de Almeida - OrientadoraUFMG

Carolina Sheng Whei Miaw BotelhoIndustria MARDICÔ

Profa. Elionai Cassiana de Lima GomesUFMG

Belo Horizonte, 22 de março de 2022.

Documento assinado eletronicamente por **Mariana Ramos de Almeida**, **Professora do Magistério Superior**, em 22/03/2022, às 11:31, conformehorário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Elionai Cassiana de Lima Gomes**, **Professora do Magistério Superior**, em 22/03/2022, às 11:32, conformehorário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carolina Sheng Whei Miaw Botelho**, **Usuário Externo**, em 22/03/2022, às 12:20, conforme horário oficialde Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1315641** e o código CRC **A5521823**.

Referência: Processo nº 23072.214457/2022-75

SEI nº 1315641

AGRADECIMENTOS

Agradeço a Deus pela elegância da vida e por ter me dado a honra de compartilhá-la com grandes pessoas.

Aos meus pais, Fernanda e Wander, minha maior fonte de apoio. A vocês, o meu mais puro amor. Obrigada por sonharem junto comigo e me fazerem acreditar que tudo é possível para aquele que crê. Sou grata por dividir esta vida com vocês.

Às minhas irmãs, Milena e Stephanie, obrigada por serem minhas companheiras de conversas e risadas. Os últimos dois anos foram mais fáceis graças a vocês.

À minha avó Fátima, ou Fafá, por ser meu maior exemplo de força, superação e independência, você é incrível, mas você já sabe disso.

Ao meu amigo de quatro patas, Dexter, obrigada pela companhia e por entender as vezes que não posso brincar e dar toda atenção que você merece. Você é o melhor pet.

Aos meus amigos e familiares, obrigada por todo o carinho e palavras de incentivo. Sei o tanto que torcem por mim e me acompanham mesmo apesar da distância.

Aos meus amigos da UFMG, principalmente a Amanda. Obrigada por me escutar e dividir comigo os desafios da vida de pós-graduando.

Em especial, agradeço a minha orientadora, Dra. Mariana Ramos de Almeida, pela orientação, dedicação e disponibilidade. Obrigada pela confiança e amizade construída ao longo desses anos. Sua orientação fez toda diferença para este trabalho.

Aos professores, Dra. Clésia Nascentes, Dr. Marcelo Sena e Dr. Rodinei Augusti, por toda parceira e contribuição para o desenvolvimento deste trabalho.

Aos membros do GQQATE e colegas de laboratório, em especial Ana Gabriela, Camila Glaucimar e Karen por toda ajuda e boa vontade.

Agradeço a Federação de Cafeicultores do Cerrado Mineiro pela parceira e incentivo no desenvolvimento deste estudo.

Agradeço a CNPq por todo apoio financeiro.

Agradeço ao Departamento de Química pela infraestrutura e ao colegiado de pós-graduação por todo apoio e entendimento.

Agradeço a UFMG, meu segundo lar a quase 10 anos. Sou infinitamente grata pela oportunidade de estudar aqui e por todo conhecimento adquirido desde os tempos de graduação.

A todos, minha sincera gratidão.

“Eu lhe tenho ensinado o caminho da sabedoria e a maneira certa de viver. Se você andar sabiamente, nada atrapalhará o seu caminho, e você não tropeçará quando correr. Lembre-se sempre daquilo que aprendeu. A sua educação é a sua vida; guarde-a bem.”

Provérbios 4:11-13

RESUMO

O café é uma das bebidas mais consumidas e apreciadas no mundo. No cenário econômico, a matéria prima é de grande relevância para o país, principalmente no estado de Minas Gerais. Com o grande avanço tecnológico e melhora na qualidade de vida, cada vez mais aumenta a busca por produtos ou serviços, que tenham algum diferencial, os cafés produzidos na Região do Cerrado Mineiro possuem certificado de Denominação de Origem que garante a qualidade e o diferencial dos grãos. Nesse contexto, o objetivo deste trabalho foi desenvolver modelos de classificação para caracterizar e discriminar os grãos de café provenientes do Cerrado Mineiro. Foram utilizadas as técnicas espectroscópicas, fluorescência de raios X por reflexão total (TXRF), espectroscopia no infravermelho médio com refletância total atenuada (ATR-MIR), espectrometria de massas por *paper-spray* (PS-MS) e espectroscopia de absorção no ultravioleta-visível (UV-Vis). Foram construídos planejamentos de experimentos para otimização da extração dos compostos presentes nos grãos de café verde a serem utilizados nas análises por PS-MS e UV-Vis. Métodos de modelagem de classe, SIMCA (modelagem independente e flexível por analogia de classe), DD-SIMCA (modelagem flexível e independente por analogia de classes orientada aos dados) e OCPLS (mínimos quadrados parciais de uma classe) foram empregados para a construção de modelos com os dados individuais de cada técnica e com os dados concatenados, de modo a aproveitar a sinergia entre os dados provenientes de diferentes técnicas. Foi aplicado o método de seleção de variáveis, seleção dos preditores ordenados (OPS), visando melhorar a performance dos modelos. Em geral, os modelos construídos com os dados de UV-Vis e fusão de dados das outras técnicas apresentaram melhores desempenho. O método de seleção de variáveis foi capaz de selecionar as variáveis mais importantes para os modelos melhorando seu desempenho. A interpretação dos modelos foi realizada por meio do poder de modelagem das variáveis em que foi possível observar que as substâncias trigonelina e ácidos clorogênicos foram responsáveis na discriminação dos grãos de café da região do Cerrado em relação aos grãos de café da região do Caparaó, Mogiana e Sul de Minas. Em relação aos elementos inorgânicos, P, Cl, Ti, Cu, Zn e Rb foram selecionados como sendo os mais importantes. O desempenho dos modelos foi interpretado por meio das figuras de mérito sensibilidade, especificidade e eficiência.

Palavras-chave: café, denominação de origem, SIMCA, poder de modelagem, fusão de dados.

ABSTRACT

Coffee is one of the most consumed, appreciated beverages in the world. In the economic context, the raw material is widely relevant for Brazil, especially Minas Gerais. With the technological advance and improvement in the life's quality, the search for products or services that have differential is increasing. The green coffee beans produced in the Cerrado Mineiro region has Protected Designation of Origin (PDO) certificate that guarantees the quality and differential of the beans. In this context, the objective of this work was to develop classification models to characterize coffee beans from the Cerrado Mineiro. Spectroscopic techniques, total reflection X-ray fluorescence (TXRF), attenuated total reflectance mid-infrared spectroscopy (ATR-MIR), *paper-spray* mass spectrometry (PS-MS) and ultraviolet-absorption spectroscopy (UV-Vis) were used in this work. Design of experiments were constructed to optimize the extraction of compounds present in green coffee beans to be used in PS-MS and UV-Vis analyses. Class-modelling methods SIMCA (*Soft Independent Modelling by Class Analogy*), DD-SIMCA (*Data driven Soft Independent Modelling by Class Analogy*) and OCPLS (*One class partial least squares*) were built for the individual data block of each technique and with the concatenated data to take advantage of the synergy between the data from different techniques. The OPS variable selection method was applied to improve the performance of the model. In general, models built with UV-Vis data and data fusion from the other techniques performed better. The variable selection method was able to select the most important variables for the models, aiming to improve their performance. The interpretation of the models was carried out through the modeling power of the variables in which it was possible to observe that trigonelline and chlorogenic acids substances were responsible for the discrimination of coffee beans from the Cerrado region in relation to coffee beans from the Caparaó, Mogiana and South of Minas region. Regarding the inorganic elements, P, Cl, Ti, Cu, Zn and Rb were selected as the most important variables from this dataset. The performance of the models was interpreted estimating the figures of merit, sensitivity, specificity, and efficiency.

Keywords: coffee, protected designation of origin, SIMCA, modelling power, data fusion.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ACR	Resíduos absolutos centralizados (*Absolute Centered Residual*)
ANOVA	Análise de variância
ATR	Reflectância total atenuada (*Attenuated Total Reflection*)
EMBRAPA	Empresa Brasileira de Agropecuária
CV	Validação cruzada (*Cross Validation*)
DD-SIMCA	SIMCA orientados aos dados (*Data Driven SIMCA*)
DO	Denominação de Origem
EFC	Eficiência
ESI	Ionização por electrospray (*Electrospray Ionization*)
ESP	Especificidade
FN	Falso negativo
FP	Falso positivo
FTIR	Infravermelho com transformada de Fourier (*Fourier Transform Infrared*)
HCA	Análise por agrupamento hierárquico (*Hierarchical Cluster Analysis*)
HOMO	Orbital molecular de maior energia (*Highest Occupied Molecular Orbital*)
HPLC	Cromatografia líquida de alta eficiência (*High Performance Liquid Chromatography*)
ICO	Organização Internacional do Café (*International Coffee Organization)*
IGs	Indicações Geográficas
INPI	Instituto Nacional de Propriedade Industrial
IP	Indicação de Procedência
IR	Infravermelho (*Infrared*)
IT	*Íon trap*
KNN	K-ésimo vizinho mais próximo (K^{th} *Nearest Neighbour*)
KS	Kennard-Stone
LUMO	Orbital molecular de menor energia não ocupado (*Lowest Unoccupied Molecular Orbital*)
MIR	Infravermelho médio (*Mid-Infrared*)
MS	Espectrometria de massas (*Mass Spectrometry*)
MSC	Correção do espalhamento multiplicativo (*Multiplicative Scatter Correction*)
NIR	Infravermelho próximo (*Near-Infrared*)

OCPLS	Mínimos quadrados parciais de uma classe (*One Class Partial Least Squares*)
OD	Distância ortogonal (*Orthogonal Distance*)
OPS	Seleção dos preditores ordenados (*Ordered Predictors Selection*)
OT	*Orbitrap*
PC	Componente principal (*Principal Component*)
PCA	Análise de componentes principais (*Principal Component Analysis*)
PDO	Proteção de Denominação de Origem (*Protected Denomination of Origin)*
PLS	Mínimos quadrados parciais (*Partial Least Squares*)
PLS-DA	Análise discriminante por mínimos quadrados (*Partial Least Square Discriminant Analysis*)
PS-MS	Espectrometria de massas por *paper spray* (*Paper Spray Mass Spectrometry*)
Q	Quadrupolo
R^2	Coeficiente de Determinação
SD	Distância dos escores (*Scores Distance*)
SDD	Detector de desvio de silício (*Silicon Drift Detector*)
SEN	Sensibilidade
SIMCA	Modelagem independente e flexível por analogia de classe (*Soft Independent Modelling by Class Analogy*)
SNV	Variação normal padrão (*Standard Normal Variate*)
TOF	Tempo de voo (*Time of Flight*)
TXRF	Fluorescência de raios X por reflexão total (*Total Reflection X-ray Fluorescence)*
UCL	Limites de confiança superiores (*Upper Confidence Level*)
VL	Variáveis latentes
VN	Verdadeiro negativo
VP	Verdadeiro positivo
UV-Vis	Ultravioleta-Visível
XRF	Fluorescência de raios X (*X-Ray Fluorescence)*

SUMÁRIO

1. INTRODUÇÃO

Pertencente à família Rubiacea, o café é uma das bebidas mais populares além de ser uma das matérias-primas mais importantes para a economia do país. O Brasil é o maior produtor dos grãos de café e um dos maiores consumidores da bebida. Grande parte da produção de café é realizada em Minas Gerais, sendo 25% na região do Cerrado Mineiro (ICO, 2021). Os grãos de café verde produzidos na região do Cerrado são conhecidos por apresentarem características sensoriais mais refinadas em relação aos grãos de outras regiões, como sabor e aroma.

Uma das maneiras de garantir a qualidade dos produtos é por meio das indicações geográficas (IGs). As IGs são ferramentas coletivas que certificam a qualidade e diferencial dos produtos ou serviços e são divididas em indicações de procedência (IP) e denominação de origem (DO). Os grãos de café produzidos no Cerrado Mineiro foram pioneiros ao receber certificado de denominação de origem (CERRADO MINEIRO, 2021).

O certificado de denominação de origem consiste no nome geográfico cujas qualidades ou características de um determinado produto estar relacionado ao meio geográfico ao qual ele foi obtido ou produzido. No caso dos grãos de café do Cerrado, as condições de clima, altitude e solo levam a obtenção de um produto com qualidade diferenciada e por consequência maior valor econômico (SEBRAE, 2016).

Diante desse cenário, o desenvolvimento de metodologias que possam caracterizar e identificar esses produtos é cada vez mais necessário. Nesse contexto, as técnicas espectroscópicas vêm sendo amplamente utilizadas em diferentes áreas de pesquisas, em especial, na análise de alimentos (SHAH, 2018) por serem de baixo custo, rápidas e ser necessário mínimo ou nenhum preparo de amostra.

Em geral, para as análises de café, diferentes técnicas analíticas podem ser utilizadas, sendo elas espectroscopia de raios-X por reflexão total (TXRF) , espectroscopia de absorção na região do infravermelho médio (MIR) e ultravioleta-visível (UV-Vis), espectrometria de massas (MS), espectroscopia Raman, entre outras. As técnicas mencionadas acima, possibilitam obter informações a respeito dos compostos moleculares presentes nos grãos de café (MIR, UV-Vis, MS e Raman) e dos elementos inorgânicos (TXRF). Portanto, são alternativas válidas quando se tem interesse em obter o perfil químico (*fingerprint*) dessas amostras.

Em conjunto com as técnicas espectroscópicas, as ferramentas quimiométricas são utilizadas para construir modelos de classificação que possam discriminar ou autenticar

amostras de interesse, destacam-se nesse sentido métodos de modelagem de classe como, modelagem independente e flexível por analogia de classe, SIMCA, (WISE et al., 2006), SIMCA orientado aos dados, DD-SIMCA, (ZONTOV *et al.*, 2017) e mínimos quadrados parciais de uma classe, OCPLS, (XU *et al.*, 2013). Além disso, podem ser aplicados métodos de seleção de variáveis ou fusão de dados com objetivo de melhorar o desempenho dos modelos.

Outra abordagem quimiométrica que é importante de ser destacada é o planejamento de experimentos. O principal objetivo do planejamento de experimentos é otimizar a resposta experimental realizando um número mínimo de ensaios. Na literatura é possível encontrar diferentes trabalhos em que o planejamento de experimentos foi realizado para determinar a escolha de solvente ou condições ideais para otimizar a extração de diferentes compostos, como por exemplo, café (MOREIRA *et al.*, 2014).

Nesse contexto, o presente trabalho visou empregar diferentes técnicas espectroscópicas em conjunto com a quimiometria, para caracterizar os compostos químicos presentes nos grãos de café da Região do Cerrado Mineiro que contribuem para as características diferenciadas destes grãos.

2. OBJETIVO

A proposta deste trabalho foi desenvolver metodologias analíticas eficientes, de baixo custo, rápidas e sustentáveis para caracterizar e discriminar grãos de café verde da espécie arábica com certificado de denominação de origem provenientes da região do Cerrado Mineiro.

2.1 Objetivos específicos

- Otimização das condições ideais para extração dos grãos de café verde a partir do planejamento de misturas e planejamento composto central;

- Emprego das técnicas MIR, PS-MS, TXRF e UV-Vis de modo a caracterizar e identificar as diferentes substâncias presentes nos grãos de café do Cerrado Mineiro e outras regiões;

- Construção de modelos de classificação multivariada, neste caso métodos de modelagem de classe, de modo a discriminar e autenticar os grãos de café do Cerrado em relação aos grãos de café da região do Caparaó, Mogiana e Sul de Minas;

- Uso de método de seleção de variáveis a fim de reduzir ou eliminar as variáveis menos significativas e dessa maneira aumentar a performance dos modelos em classificar as amostras;

- Emprego da estratégia de fusão de dados de nível baixo com o objetivo de aproveitar a sinergia dos dados de diferentes técnicas e, principalmente relacionar as informações obtidas das amostras em termos da composição atômica-molecular.

3. REVISÃO DA LITERATURA

3.1 Café

O café é uma das bebidas mais consumidas no mundo (NEHLIG, 2016). De acordo com a Organização Internacional do Café (*ICO – International Coffee Organization*), o consumo de café cresceu mais que sua produção nos últimos dois anos gerando um superávit no mercado global entre oferta e procura. Economicamente, o café é uma matéria prima de grande importância para o mercado brasileiro, uma vez que o Brasil é o principal produtor e exportador de grãos de café verde no mundo, tendo Minas Gerais como responsável por mais da metade da produção e exportação da matéria-prima (BARBOSA *et al*., 2020).

Segundo a Empresa Brasileira de Agropecuária (EMBRAPA), o Brasil registrou recorde na exportação do café na safra de 2020-2021, com um total de 45,6 milhões de sacas produzidas e obtendo desempenho 13,3% maior que o ano anterior. É importante ressaltar que do total de sacas, 91,3% correspondem ao café verde, sendo 81% referente ao café arábica e 10,3% ao café robusta.

O café pertence à família Rubiacea, sendo as espécies *Coffea arabica* e *Coffea robusta*, as mais cultivadas e as mais importantes para a economia (VEIGA *et al.*, 2019). A espécie arábica apresenta grãos que resultam em bebidas com características sensoriais mais pronunciadas e refinadas em relação a espécie robusta e por consequência possui um maior valor de mercado (EL-ABASSY *et al*., 2011). Como mencionado anteriormente, no Brasil, grande parte da produção e exportação de café verde corresponde a espécie arábica.

Historicamente, o café arábica era cultivado apenas em regiões sem escassez de água nos períodos críticos de colheita, no entanto, com elevado avanço tecnológico e aumento das atividades agrícolas, o cultivo do café espalhou-se para outras áreas, como por exemplo, áreas de Cerrado (VEIGA *et al.*, 2019) e suas condições de cultivo foram otimizadas de modo a adequar condições de solo, manejo, colheita, entre outros fatores (GONÇALVES, 2018).

Para obtenção do grão da espécie arábica com maior qualidade é necessário que a planta do café cresça com um nível de precipitação variando entre 1200 mm e 1800 mm e que o período de chuva seja bem distribuído ao longo de todo processo até a chegada do fruto. Além disso, a temperatura e altitude também são fatores que irão influenciar diretamente na qualidade do grão, é importante que a temperatura seja em torno de 18-23°C e altitude entre 600-1200 m (MESQUITA, 2016).

Com uma produção de café em torno de 12,7% em relação a produção nacional e 25,4% da produção mineira, produtores da região do Cerrado Mineiro, vem cada vez mais se consolidando e sendo reconhecidos no mercado cafeeiro por produzir grãos de café arábica com qualidade diferenciada aos demais. O Cerrado Mineiro foi a primeira região no Brasil a obter certificação de Denominação de Origem para grãos de café (ALMEIDA, 2019). Isso foi possível devido ao fato de a região apresentar condições técnicas ideais que favorecem tanto o plantio quanto a colheita do café, tais como, altitude variando entre 800-1300 m, temperatura média anual de 22°C, média de chuva anual de 1800 mm, inverno seco que favorece o período de colheita, sol intenso e topografia plana que favorece a colheita mecanizada (CERRADO MINEIRO, 2021).

Segundo o INPI (Instituto Nacional de Propriedade Industrial) (2016), o café produzido na região do Cerrado Mineiro resulta de aspectos climáticos únicos, maturidade uniforme, colheita concentrada e flores intensas, ou seja, todos esses fatores influenciam diretamente nas características sensoriais obtidas para os cafés produzidos, como, acidez delicadamente cítrica, aromas intensos variando entre caramelo e nozes, além do sabor adocicado com traços de chocolate e longa duração.

3.2 Composição química do café

O café é uma matriz complexa e sua composição química varia de acordo com sua espécie. Os diferentes compostos presentes no grão de café verde podem influenciar na qualidade sensorial da bebida.

Os grãos de café da espécie *Coffea arabica* possuem uma maior quantidade de ácidos e estes são responsáveis por contribuir para o sabor, aroma e acidez da bebida. Dentre os diferentes tipos de ácidos presentes no café, destacam-se a classe de ácidos clorogênicos, ácido quínico, ácido cítrico e ácido málico (RIBEIRO, 2017).

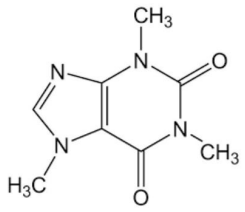
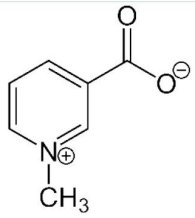
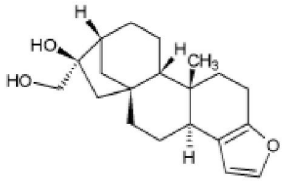
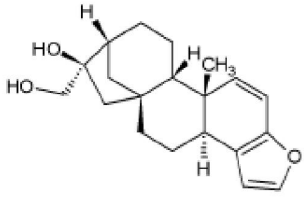
Os ácidos clorogênicos são uma classe de compostos provenientes da esterificação do ácido quínico com outros ácidos, como ácido caféico, ferúlico e p-cumárico. Os principais produtos obtidos dessa reação que são encontrados, majoritariamente, nos grãos de cafés são ácidos cafeoilquínicos e dicafeoilquínicos (TOCI, 2006).

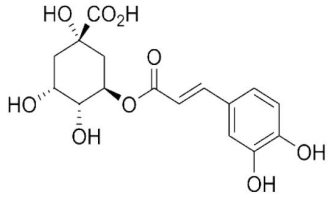
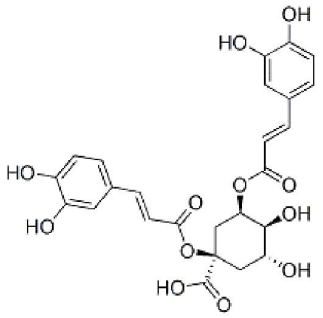
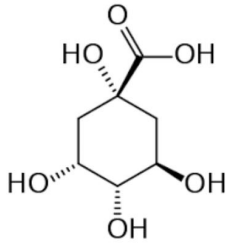
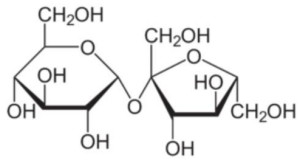
Os compostos cafestol e caveol, assim como os compostos nitrogenados, trigonelina e cafeína, são importantes na caracterização do café e influenciam na qualidade sensorial da bebida. O cafestol e o caveol são diterpenos presentes na fração lipídica das plantas do gênero

Coffea, encontrados principalmente na espécie *Coffea arabica* e podem ser utilizados como indicadores da qualidade do café (ABREU, 2019). A trigonelina é encontrada principalmente em cafés da espécie arábica e é importante na formação do sabor e aroma da bebida de café. Já a cafeína é encontrada em maiores teores nos cafés da espécie robusta e confere a bebida maior amargor (BRAGA, 2019).

Os açúcares também contribuem na qualidade sensorial da bebida, sendo que a sacarose é o açúcar encontrado em maiores quantidade e são importantes na caracterização química dos cafés (BORÉM *et al*., 2016). Na Tabela 1, estão apresentados os principais compostos encontrados no café e sua estrutura química.

Tabela 1 - Estruturas químicas dos principais compostos presentes nos cafés.

Composto	**Fórmula Química**	**Estrutura**
Cafeína	$C_8H_{10}N_4O_2$	
Trigonelina	$C_7H_7NO_2$	
Cafestol	$C_{20}H_{28}O_3$	
Caveol	$C_{20}H_{26}O_3$	

Ácido 5-cafeoilquínico	$C_{16}H_{18}O_9$	
Ácido 1,5-dicafeoilquínico	$C_{25}H_{24}O_{12}$	
Ácido quínico	$C_7H_{12}O_6$	
Sacarose	$C_{12}H_{22}O_{11}$	

3.3 Indicações Geográficas

Com a melhora na qualidade de vida populacional, a procura por produtos e serviços aumentou consideravelmente, principalmente em relação aos produtos que são consumíveis. Paralelamente, têm se os produtores de diferentes matérias primas que buscam cada vez mais assegurar a qualidade dos seus produtos ou serviços. Assim, diante de um cenário em que a sociedade valoriza o aspecto qualitativo de um produto a ser utilizado, a garantia de qualidade se tornou um dos principais pré-requisitos exigidos para se ter um produto com alta valorização comercial (BATISTA, 2012). Nesse contexto, as indicações geográficas vêm de encontro a esses requisitos.

As indicações geográficas (IGs) podem ser definidas como ferramentas coletivas de valorização de produtos tradicionais que estão vinculados a um determinado território, sendo que os principais objetivos são agregar valor econômico e estabelecer um diferencial competitivo ao produto ou serviço, de modo a valorizar e promover a região de produção (SEBRAE, 2016).

No Brasil, as IGs são separadas, de acordo com a Lei de Propriedade Industrial (9.279/96), em duas categorias, indicação de procedência (IP) e denominação de origem (DO). A indicação de procedência consiste no nome geográfico que se tornou conhecido como centro de extração, produção ou fabricação de determinado produto ou prestação de serviço. Já a denominação de origem é o nome geográfico atribuído ao produto ou serviço cujas qualidades ou características que o diferenciam são provenientes em sua maioria do meio geográfico em que se encontra, inclusos fatores naturais e humanos (SEBRAE, 2016).

No âmbito internacional, a regulamentação europeia (No. 2081/92) divide os produtos agrícolas e alimentícios, protegidos por origem, em indicações geográficas e denominação de origem (MARCOZ *et al*., 2016), sendo que a diferença entre elas será definida pela influência em que a área geográfica em questão terá sobre o produto.

Para receber o termo de indicação geográfica, a obtenção do produto deve ser realizada em área específica, não sendo obrigatório todas as etapas de obtenção do produto no mesmo local, uma vez que a característica final do produto não é levada em consideração para obtenção do registro (BECKER, 2009). Para ser elegível como produto com denominação de origem (PDO), o produto precisa ter dois requisitos:

1) As características principais e qualidades do produto devem ser majoritariamente ou particular daquele local geográfico, incluindo clima, solo e conhecimento da área.

2) Todo processamento, desde a matéria prima até o produto, deve ser realizado no local de origem geográfica ao qual o produto é reconhecido.

Atualmente, existem vários tipos de produtos que possuem denominação de origem sendo que o maior número deles se encontram distribuídos na Europa, como por exemplo, queijos produzidos no norte da Itália (ROCCHETTI *et al.*, 2021), vinagres no sul da Espanha (RÍOS-REINA *et al.*, 2020) e azeites no sul da França (MEDINI *et al.*, 2015).

No Brasil, os exemplos são o café na região do Cerrado Mineiro, Mantiqueira de Minas e Caparaó, o queijo na Serra da Canastra, ambos em Minas Gerais e o vinho no Vale dos Vinhedos no Rio Grande do Sul (SILVA *et al.*, 2020). Por possuir um maior valor agregado no mercado, produtos com denominação de origem, têm sido alvos de adulterações que visam

principalmente um maior lucro econômico. Em geral, essas adulterações são feitas com produtos de qualidade inferior ou até mesmo aditivos que não são permitidos por apresentar alguma toxicidade à vida humana (ELLIS *et al.*, 2012).

Diante desses fatores e com objetivo de garantir a qualidade e identidade dos produtos com denominação de origem tem crescido cada vez mais o desenvolvimento de metodologias analíticas que visam discriminar e provar sua autenticidade. Frente a necessidade de novos métodos analíticos, as técnicas espectroscópicas em conjunto com a quimiometria vêm sendo utilizadas e cada vez mais suas aplicações têm sido estudadas, principalmente na área de autenticidade de alimentos (MEDINA *et al.*, 2019).

3.4 Técnicas Analíticas

3.4.1 Fluorescência de raios X por reflexão total (TXRF)

A fluorescência de raios X por reflexão total (TXRF) é uma técnica analítica multielementar, rápida e sensível, razões pela qual é amplamente utilizada nas indústrias (HORNTRICH *et al.*, 2011). As principais vantagens que a técnica apresenta são: requerer mínima quantidade de amostra; possuir baixos limites de detecção; ser fácil de operar, uma vez que não consome nenhum tipo de gás; oferecer resultados com baixo ruído e efeito de matriz e, possibilitar o uso de instrumentos portáteis (DE LA CALLE *et al.*, 2013).

A TXRF é uma modificação geométrica da espectroscopia de fluorescência de raios x (XRF) e foi proposta devido a necessidade de uma técnica com maior sensibilidade e menor efeito de matriz (SZOBOSZLAI *et al.*, 2009). Na fluorescência de raios x, um feixe de raios x primário que foi formado pelo bombardeio de um ânodo com elétrons acelerados é direcionado na amostra (ALLEGRETTA *et al.*, 2019). Em seguida, como resultado da interação entre o feixe primário e os componentes da amostra, é emitida uma radiação de raios x secundária em forma de radiação de fluorescência. Essa radiação possui energia e comprimento de onda específico para cada elemento químico e é proporcional a concentração do elemento presente na amostra (ANTOSZ et al., 2012). Basicamente, as modificações na técnica clássica que resultou na TXRF foram: 1) a formação de uma película fina da amostra na superfície de um plano, que refletisse totalmente a placa transportadora (feita com quartzo, monocristal de silício), 2) o ângulo de incidência dos raios x primários em relação a placa seria menor que 0,1°, ou seja, haveria reflexão total dos fótons dos raios x primários e 3) a distância entre a amostra

e o detector seria reduzida para 1-2 mm (SZOBOSZLAI *et al.*, 2009). A Figura 1 mostra o princípio geral da técnica.

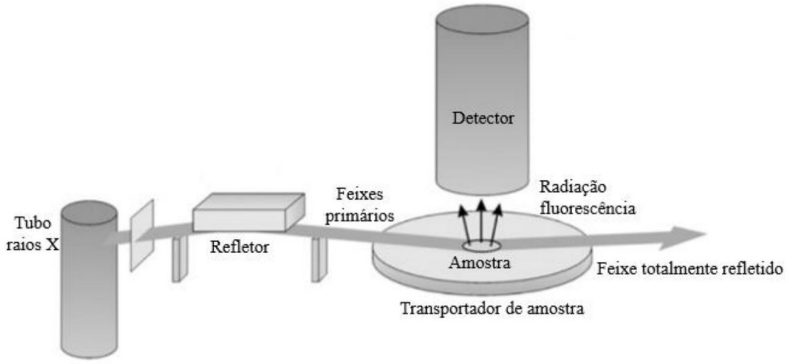


Figura 1 - Princípio geral da técnica de fluorescência de raios x por reflexão total (Adaptada de VON BOHLEN, 2009).

Na Figura 2 abaixo, é possível observar todas as etapas a serem realizadas na análise via TXRF, desde a hidrofobização dos discos até a análise da amostra. É importante frisar que todo o processo de preparo dos discos é de extrema importância para evitar efeito de matriz ou outras interferências.

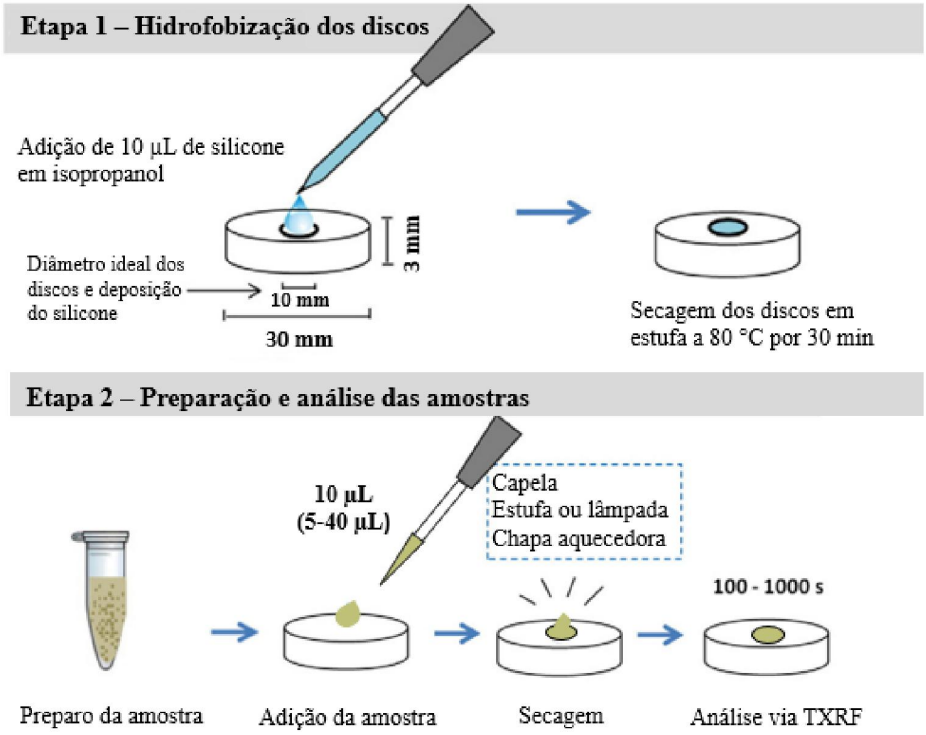


Figura 2 - Etapas a serem realizadas para preparo dos discos e amostras para análise por TXRF (Adaptada de LA CALLE *et al.*, 2013).

O cálculo da concentração analítica dos elementos é realizado por meio das equações 1 e 2 (BRUKER, 2012), a seguir:

$$Ci = \frac{Ia\ x\ Cp}{Ip\ x\ S'a}$$

Equação 1

$$S'a = \frac{Sa}{Sp}$$

Equação 2

Sendo:

Ca é a concentração (mg.kg^{-1}) do analito a na solução;

Ia é a intensidade (cps – contagens por segundo) das linhas K e L características emitidas dos elementos presentes na amostra;

Cp é a concentração (mg.kg^{-1}) do padrão interno;

Ip é a intensidade (cps – contagens por segundo) das linhas K e L características emitidas do padrão interno;

$S'a$ é a sensibilidade relativa do elemento e indica a relação entre a intensidade do pico do analito e a quantidade na amostra;

Sa e Sp referem-se à sensibilidade elementar do sistema (cps/ppm ou mg.kg^{-1}) para o analito *a* e padrão interno, respectivamente.

Na literatura é possível encontrar trabalhos utilizando a técnica de TXRF em conjunto com a quimiometria para prever adulterações em whisky (SHAND *et al.*, 2017), identificação de produtos com origem geográfica como mel (KROPF *et al.*, 2010) e vinho (VITALI ČEPO *et al.*, 2022) e caracterização de compostos presentes em misturas de café arábica e robusta (ASSIS *et al.*, 2020).

3.4.2 Espectrometria de massas (MS)

A espectrometria de massas (*Mass Spectrometry - MS*) é uma das ferramentas analíticas mais poderosas devido à sua alta sensibilidade, seletividade e precisão, e cada vez mais vem sendo utilizada por diferentes áreas da ciência, como: análise forense, autenticação de alimentos, análise *antidoping*, biomedicina e pesquisa farmacêutica (AWAD *et al.*, 2015). A

técnica em si depende da formação de íons na fase gasosa, carregados positiva ou negativamente, que são separados eletromagneticamente com base na sua relação massa-carga (m/z), ou seja, o espectro obtido via análise MS terá como resposta, a razão m/z de um composto (eixo x) versus a quantidade de íons totais (eixo y). A principal relevância do seu uso é a capacidade de fornecer informações tanto de moléculas orgânicas quanto inorgânicas, assim como a respeito da estrutura, composição e pureza (EL-ANEED *et al.*, 2009).

De forma geral, o espectrômetro de massas possui cinco componentes (Figura 3), um dispositivo de entrada (*inlet)*, uma fonte de ionização, um analisador, um detector e um sistema que irá gerar os dados. Basicamente, é introduzida a amostra por meio do dispositivo de entrada na fonte de ionização, esta por sua vez produz os íons que serão separados pelo analisador e, posteriormente, identificados pelo detector (Figura 4). O sistema será controlado por um computador que irá manipular e armazenar os dados por meio de um software para, assim, fornecer o espectro em relação a amostra (LANÇAS, 2019).

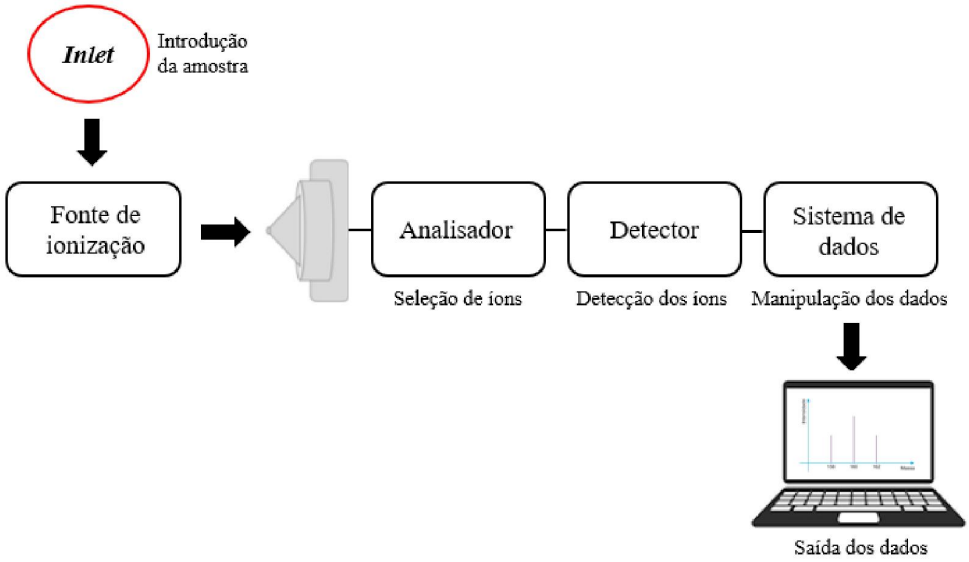


Figura 3 - Esquema ilustrativo de um espectrômetro de massas.

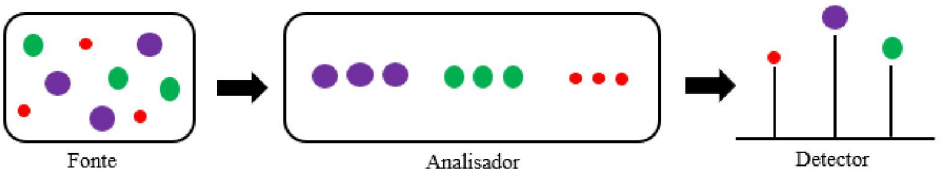


Figura 4 - Esquema simplificado das principais etapas de funcionamento (Adaptado de LANÇAS, 2019).

É importante ressaltar que o esquema mostrado acima é referente somente ao espectrômetro de massas. Geralmente os espectrômetros de massas vêm acoplados ou hifenados a algum cromatógrafo, como grande parte das amostras analisadas são misturas complexas de diferentes compostos, é necessário a separação dos compostos antes deles serem direcionados ao espectrômetro, que por sua vez identifica e quantifica os compostos presentes nas amostras mais facilmente (LANÇAS, 2019).

Cada componente do equipamento é importante para o bom desempenho da análise e é responsável por um processo individual na análise que resultará no espectro final, desse modo, é possível modificar esses componentes de modo a melhorar a performance da técnica e por consequência o resultado. Por exemplo, os analisadores de massa, são responsáveis por separar os íons de acordo com sua m/z, e para isso fazem uso da eletricidade estática ou dinâmica e campos magnéticos. Assim, a forma como esses campos são utilizados resulta em diferentes tipos de analisadores e define se o equipamento será de baixa ou alta resolução. Os analisadores mais comuns são do tipo quadrupolos (Q), aprisionamento de íons (*íon trap* – IT), *orbitrap* (OT) e, tempo de voo (*time of flight* – TOF) (HOFFMANN, 2007).

Outro ponto a ser destacado são as técnicas de ionização da amostra, estas podem ser classificadas de acordo com a energia envolvida, ou seja, pelo grau de ionização e fragmentos obtidos no processo. São divididas em: s*oft* – técnica de ionização branda em que a molécula será ionizada formando íons moleculares em abundância, sem ou com pouca ionização, além de não gerar alta energia interna no íon molecular formado, será mais utilizado na identificação de compostos desconhecidos, onde o interesse é obter a massa molecular, e *hard* – técnica em que será fornecido energia interna suficiente para causar fragmentação no íon molecular, que por sua vez permite a geração de muito íons que podem fornecer informações a respeito da estrutura química dos compostos desconhecidos. No geral, as duas técnicas são utilizadas associadamente para determinação da estrutura química na análise de compostos desconhecidos ou não-alvo (LANÇAS, 2019). Na Tabela 2, encontram-se destacadas as principais técnicas de ionização utilizadas na espectrometria de massas. Dentre as técnicas de ionização apresentadas abaixo, cabe destacar que uma das mais utilizadas em diferentes trabalhos na área de alimentos é a ionização por *electrospray* (KALOGIOURI *et al*., 2018; KALOGIOURI *et al*., 2020; TAHIR *et al*., 2022).

Tabela 2 - Principais técnicas de ionização na espectrometria de massas.

Técnica	Agente ionizante	Aplicação
Ionização com elétrons - EI	Elétrons (30-70 eV)	Determinação de estrutura; GC-MS.
Ionização química - CI	Íons gasosos	Determinação da MM; GC-MS.
Dessorção com laser – LDI	Fótons	Análises de superfície.
LDI auxiliada pela matriz - MALDI	Fótons	Análise de moléculas orgânicas.
Electrospray – ESI	Campo elétrico	LC-MS, CE-MS.
Ionização química sob pressão atmosférica - APCI	Descarga Corona	LC-MS

Na ionização por *electrospray* (ESI), a amostra dissolvida é pressurizada em um capilar sob alta voltagem levando a formação de gotículas (cone de Taylor) carregadas que são dessolvatadas à medida que o solvente evapora. Em seguida, induzido pelos efeitos da atração eletrostática e vácuo, os íons formados fluem para o espectrômetro de massas. Além do mais, dependendo do sinal da tensão aplicada, é possível controlar o modo de operação do electrospray, caso utilize o modo positivo, as gotículas formadas terão cargas positivas e o eletrodo receberá elétrons, ocorrendo um processo de oxidação. O contrário ocorrerá, caso o modo negativo seja escolhido (LANÇAS, 2019).

Apesar de ser um método de ionização capaz de produzir íons negativos e positivos e ainda utilizar de um processo de evaporação dos íons sob pressão atmosférica, o que resulta em uma análise sensível e sem degradação térmica dos analitos, a técnica possui algumas limitações. Dentre as limitações, destacam-se a necessidade de preparo de amostra mais complexo, o capilar utilizado ser suscetível a entupimento e, dificuldade na análise direta em matérias primas como, material biológico ou alimentos, (HU, 2021). Nesse contexto, foram propostos alguns outros modos de ionização que se baseavam no princípio do *electrospray*, mas com uma metodologia mais simplificada, como é o caso, da ionização por *paper spray*.

A ionização por *paper spray* consiste na adição de amostra em um pedaço de papel cortado em formato triangular direcionado a entrada do equipamento, que após a aplicação de uma alta voltagem no papel molhado, leva a formação de íons que serão analisados diretamente no espectrômetro de massas, como ilustrado na Figura 5. As principais vantagens dessa técnica são, baixo custo, rapidez na análise, flexibilidade no desenvolvimento de novas metodologias, e, ainda, o mínimo o preparo de amostras e uso de solvente (LIU *et al.*, 2010).

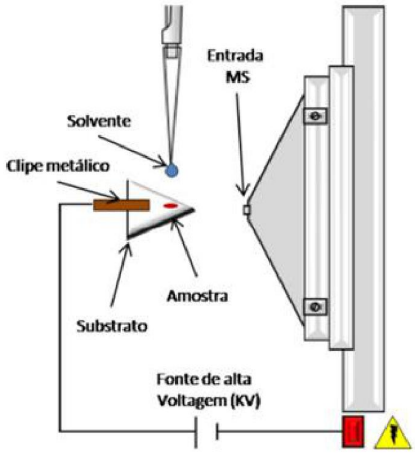


Figura 5 - Esquema do funcionamento do espectrômetro de massas por *paper spray* - PSMS (PEREIRA, 2016).

Em conjunto com as ferramentas quimiométricas, a espectrometria de massas vem sendo utilizada no desenvolvimento de modelos para discriminação de vinhos com origem geográfica (GUO *et al.*, 2021) e caracterização de cafés da espécie arábica (GARRETT *et al.*, 2013).

3.4.3 Espectroscopia de absorção na região do infravermelho médio (MIR)

A espectroscopia no infravermelho é uma das técnicas analíticas mais utilizadas no meio científico, sendo sua aplicação em diferentes áreas de estudos, tais como controle de qualidade, identificação de compostos orgânicos e inorgânicos, análise de misturas complexas, transporte de moléculas bioativas em tecidos vivos, entre outros (BARBOSA, 2013). Isso acontece devido ao fato de a análise não requerer abertura da amostra, minimizando o uso de solventes que podem ser tóxicos, além de ser simples e rápida (BORGES, 2015). A energia na região do infravermelho, está situada na faixa do espectro eletromagnético entre 14290-200 cm^{-1}, sendo a faixa entre 4000-400 cm^{-1} a mais utilizada, principalmente na análise de compostos orgânicos. Essa faixa é denominada infravermelho médio (BARBOSA, 2013).

Semelhante a outros tipos de absorção de energia, quando as moléculas absorvem radiação no infravermelho, elas são excitadas a um estado vibracional de maior energia. A radiação nessa faixa de energia corresponde a região de frequência em que ocorrem as frequências vibracionais de estiramento e deformações das ligações das moléculas covalentes. Cada frequência vibracional equivale a um tipo de ligação, por exemplo, duas ligações iguais em compostos diferentes, estão em ambientes diferentes, e desse modo, os padrões de absorção

no infravermelho para cada um deles serão diferentes. Com isso, pode-se dizer que o espectro no infravermelho possibilita obter a impressão digital (*fingerprint*) das moléculas, sendo uma ferramenta analítica bastante útil (PAVIA, 2010).

Em relação a instrumentação, os espectrômetros atuais são com transformada de Fourier (FTIR – Fourier *transform infrared*), e isso deve-se ao fato de possuírem maiores vantagens em relação aos espectrômetros dispersivos, como maior velocidade e sensibilidade, calibração mais exata, maior precisão, desenho mecânico simples e eliminação de problemas de radiação espúria e emissão IV. Além disso, é possível detectar e medir todos os comprimentos de onda de modo simultâneo, uma vez que é utilizado um interferômetro ao invés de um monocromador para produzir padrões de interferência que contém a informação espectral. De modo geral, para obtenção do espectro de uma amostra, é necessário realizar o *background* (espectro de fundo), e somente após, adquirir o espectro da amostra. O espectro final será registrado em termos da absorbância ou transmitância *versus* número de onda (SKOOG, 2006)

Uma das maneiras de operar o FTIR, é utilizando um acessório de reflectância total atenuada (ATR – *Attenuated total reflection*). Neste método, a amostra sólida ou líquida deve ser colocada próximo a parte ótica do instrumento, onde a luz é totalmente refletida internamente e a amostra interage com a onda evanescente (radiação). Essa interação depende do comprimento do caminho efetivo e é basicamente, a fração de um comprimento de onda. O uso do ATR é ideal para amostras que são altamente absorventes, superfícies, medições finas de filme e soluções (GRDADOLNIK, 2002).

Devido ao fato de ser uma técnica rápida, minimamente destrutiva e que requer pouco preparo de amostra, como mencionado anteriormente, a técnica de FTIR com o uso do acessório de ATR vem cada vez mais sendo utilizada em conjunto com as ferramentas quimiométricas para trabalhos de identificação e caracterização de diferentes matérias-primas como cogumelos, madeira, óleo de mostarda e nano-compósitos. (CUSTERS *et al*., 2015 ; WANG *et al*., 2019 ; SHARMA *et al*., 2020 ; JAMWAL *et al*., 2021 ; MUROGA *et al*., 2021). Além disso, é possível encontrar diferentes trabalhos aplicando a técnica em conjunto com a quimiometria para avaliação da qualidade de grãos de café verde (CRAIG *et al*., 2012), grãos torrados (VARÃO SILVA *et al.*, 2021), e caracterização da bebida a partir das características sensoriais (BELCHIOR *et al.*, 2019) .

3.4.4 Espectroscopia de absorção na região do ultravioleta-visível (UV-Vis)

A espectroscopia na região do ultravioleta-visível é uma técnica analítica muito difundida. Por possuir equipamento de baixo custo e facilidade operacional, ser uma técnica espectroscópica quantitativa e produzir resultados simples de interpretar, a técnica é bastante utilizada, principalmente, em laboratórios de análises e pesquisas nas áreas de física, química, bioquímica e farmacológica. Sua aplicação é amplamente diversa, podendo ser utilizada desde a quantificação direta de pequenas moléculas orgânicas até a investigação de propriedades óptico-eletrônicas de filmes finos (GALO, 2009). O espectro eletromagnético na região do UV-Vis compreende faixa de 800 a 190 nm, sendo 800 a 400 nm, a região do visível e 400 a 190 nm, região do ultravioleta .

Na espectroscopia de absorção na região do UV-Vis, as transições resultantes da absorção da radiação eletromagnética acontecem entre níveis eletrônicos de energia, ou seja, quando uma molécula absorve energia, um elétron se move de um orbital ocupado de maior energia (HOMO) para um orbital desocupado de menor energia (LUMO). Como cada transição eletrônica consiste em muitas linhas próximas entre si, o espectrofotômetro não consegue defini-las, assim, um espectro UV obtido para uma molécula, geralmente, será composto de uma banda larga de absorção na região próxima ao comprimento de onda da transição principal.

Basicamente, um espectrofotômetro é composto de uma fonte de luz de deutério e outra de tungstênio, usadas para emitir radiação eletromagnética na região do ultravioleta e visível, nesta ordem, um monocromador, cuja função é separar o feixe de luz nos comprimentos de onda e um detector que registra a intensidade da luz transmitida. Para fazer a varredura de todos os comprimentos de onda, é necessário um sistema mecânico que possibilita o giro do monocromador e assim fornecer a varredura de toda a faixa espectral. O espectro final é fornecido em termos de absorbância *versus* comprimento de onda, sendo o cálculo da absorbância baseado na Lei de Lambert-Beer, mostrado na equação 3 (SKOOG, 2006; PAVIA, 2010).

$$A = \varepsilon\, b\, c$$

Equação 3

Onde: A é absorbância, ε é a absortividade molar (L mol L^{-1} cm^{-1}), b o caminho óptico (cm) e c a concentração (mol L^{-1}).

A espectroscopia UV-Vis é uma técnica bastante rápida e versátil, possui um custo relativamente baixo e fornece informações relevantes a respeito de moléculas orgânicas

(BARBOZA et al., 2010). Devido a esse fato, é possível encontrar na literatura diferentes trabalhos fazendo uso da técnica em conjunto com a quimiometria, de modo a caracterizar, monitorar e discriminar diferentes matrizes alimentícias que possuem origem geográfica como, chá verde, cúrcuma e bebidas alcoólicas (ABOULWAFA *et al.*, 2019; REN *et al.*, 2021; URÍČKOVÁ, 2015), identificação de adulterações em vinagres (CAVDAROGLU, 2022), cafés (REIS, 2012; SOUTO, 2017), entre outros (LI *et al.*, 2022; TAN, 2015; TARHAN, 2020).

3.5 Quimiometria

Com o grande avanço das técnicas instrumentais e dos processadores, tornou-se possível a medição de muitas variáveis de modo simultâneo, por consequência, fez-se necessário tratamentos de dados mais avançados e complexos, como por exemplo, utilização de estatística multivariada, álgebra matricial e análise numérica (BRUNS, 1985). Nesse contexto, em meados dos anos 70, a quimiometria surgiu como uma subárea da química analítica, em que é possível utilizar de métodos matemáticos e estatísticos avançados de modo a interpretar os dados e processos químicos e assim obter maiores informações a respeito das muitas variáveis estudadas (HOPKE, 2003).

A quimiometria pode ser definida como uma disciplina da química em que por meio da matemática, estatística e lógica formal é possível planejar ou otimizar procedimentos experimentais, extrair o máximo de informação possível por meio das análises de dados e por fim, obter melhor conhecimento a respeito dos sistemas químicos (KOWALSKI, 1975). No entanto, devido a sua aplicação em diferentes áreas de pesquisa, como biológicas, farmacêuticas e engenharia, pode-se definir quimiometria como sendo a administração e processamentos de informações de natureza química (FERREIRA, 2015).

A quimiometria pode ser aplicada de duas maneiras. A primeira, planejamento e otimização de experimentos, em que o objetivo é utilizar dos princípios estatísticos de forma a extrair informações úteis de um sistema realizando um número mínimo de experimentos, em menor tempo e custo (BARROS NETO *et al.*, 2010). A segunda na análise de dados, por exemplo, utilizando calibração multivariada, que pode ser definida como um conjunto de métodos de análise de dados aplicados a extração e interpretação das informações obtidas nas análises multivariadas com intuito de desenvolver modelos de quantificação para o objeto de estudo (KOWALSKI, 1975). Nessa última, encontram-se os métodos de reconhecimento de padrões ou métodos quimiométricos, em que por meio de tendências ou agrupamentos de um

conjunto de dados é possível interpretar os resultados de modo a identificar, caracterizar ou diferenciar amostras a partir dos seus perfis químicos (FERREIRA, 2015).

Os métodos de reconhecimentos de padrões podem ser separados em dois grupos, métodos supervisionados e não supervisionados. Nos métodos não supervisionados, são construídos modelos de modo a visualizar as informações presentes nos dados experimentais sem a obrigatoriedade do conhecimento prévio das características do conjunto de amostras, pode-se dizer que é um método de análise exploratória dos dados, sendo que os mais comumente utilizados são a PCA (*principal component analysis*) e a HCA (*hierarchical cluster analysis*). Já nos métodos supervisionados, é necessário o conhecimento inicial a respeito de uma classe de amostras (ou classes de amostras), uma vez que, durante a análise dos dados serão utilizadas amostras de interesse. As informações obtidas experimentalmente para classe de amostras alvo são definidas como conjunto de treinamento, e são utilizadas para reconhecer um novo conjunto de amostras, denominado conjunto teste. Por sua vez, os métodos supervisionados são divididos em análise discriminante e modelagem de classe (FERREIRA, 2015).

Na análise supervisionada existem diferentes métodos que podem ser utilizados na construção dos modelos de classificação, sendo os mais comuns KNN (*k-nearest neighbours*), LDA e mais recentemente PLS-DA (*partial least squares – discriminant analysis*), na parte de análise discriminante, e SIMCA (*soft independent modelling by class analogy*) na parte de modelagem de classe (JIMÉNEZ-CARVELO *et al.*, 2019; NEVES, 2020). Nos últimos anos, novos métodos foram propostos na área de modelagem de classe, o DD-SIMCA (*data driven soft independent modelling of class analogy*) (ZONTOV *et al.*, 2017) e, OCPLS (*one class partial least squares*) (XU et al., 2013).

Em geral, pode-se dizer que os métodos de análise discriminante são mais utilizados quando se tem interesse em identificar ou caracterizar várias classes em uma matriz de interesse, como por exemplo, casos de adulteração, uma vez que é um método em que é necessário duas ou mais classes presentes no conjunto de treinamento para desenvolvimento do modelo (RÍOS-REINA *et al.*, 2018, 2019). Já os métodos de modelagem de classe são mais aplicados quando se tem interesse em apenas uma classe, como por exemplo, o desenvolvimento de modelos que possam caracterizar e garantir a autenticidade de um determinado produto (CASALE *et al.*, 2012; FORINA *et al.*, 2015).

É importante destacar que a escolha do método modelagem de classe ou análise discriminante é extremamente importante. Métodos de modelagem de classe única tem como

principal objetivo modelar apenas uma única classe de interesse, uma vez que as características das amostras não-alvos não são importantes para serem avaliadas. No caso das análises discriminantes, obrigatoriamente, uma amostra desconhecida será atribuída a uma classe, o que pode levar a um erro de classificação das amostras e modelo final ambíguo. Assim, vale ressaltar que em problemas de autenticação cujo principal objetivo é modelar apenas um grupo de amostras alvo, os métodos de modelagem de classe única são mais indicados (OLIVERI, 2017).

Neste trabalho, foram utilizados os métodos de modelagem de classe, SIMCA, DD-SIMCA e OCPLS para o desenvolvimento de modelos que possam caracterizar e autenticar grãos de café verde com certificação de denominação de origem, provenientes da região do Cerrado Mineiro.

3.5.1 Planejamento de Experimentos

Segundo Miller & Miller (2005), o termo planejamento de experimentos, geralmente é utilizado para descrever as seguintes etapas: a) identificação dos fatores ou variáveis que podem afetar a resposta de um experimento, b) projeção dos experimentos de modo a minimizar os efeitos dos fatores que não são possíveis controlar e c) aplicar análise estatística para avaliar os efeitos dos diferentes fatores envolvidos. Uma vez que existem muitos fatores que podem afetar a resposta experimental, é necessário o uso das ferramentas de planejamento de experimentos, de modo a obter conhecimento mais detalhado a respeito das melhores condições de estudo.

Quando se tem um experimento em que vários fatores são de interesse, utiliza-se um experimento fatorial, ou seja, são realizadas medidas experimentais com todas as combinações dos fatores possíveis de modo a conhecer melhor os efeitos desses fatores na resposta de interesse, assim como a influência das interações entre os fatores (MONTGOMERY, 2009).

No entanto, à medida que vai se aumentando o número de fatores investigados, o número de experimentos a serem realizados aumenta consideravelmente, por exemplo, um experimento com 7 fatores necessita da realização de $2^7 = 128$ ensaios. Assim, baseando-se no princípio da expansão em série de uma função, pode-se dizer que os efeitos principais tendem a ser mais importantes que os efeitos de segunda ordem que, por sua vez, são mais importantes que os de terceira ordem e, assim, por diante. Portanto, conclui-se que há grande possibilidade de as variáveis de ordem superior serem não significativas e irrelevantes para a resposta final. Nesse contexto, pode ser utilizado o planejamento fatorial fracionário, em que é possível obter

informações das variáveis mais relevantes realizando um menor números de experimentos (BARROS NETO *et al*., 2010).

Uma alternativa a ser utilizada no planejamento de experimento, é o planejamento fatorial com ponto central. Na maioria das vezes, a utilização desse planejamento acontece quando não é possível realizar replicatas em todos os pontos do planejamento, assim, para contornar esse problema e obter uma boa estimativa do erro do modelo e falta de ajuste, são realizados experimentos no centro do planejamento, ou seja, no valor médio dos níveis alto e baixo . No entanto, caso o planejamento com ponto central não seja adequado, é possível adicionar pontos axiais ao experimento e realizar um planejamento composto central para avaliar a existência de termos quadráticos no modelo de regressão.

Uma vez que, o modelo foi determinado e foram obtidas as condições ideais para otimização dos seus experimentos, é necessário avaliar a qualidade de ajuste do modelo, realizado por meio da análise de variância (ANOVA). Basicamente, um bom modelo será aquele em que a maior parte da variação em torno da média será explicada pela equação de regressão e a menor parte sendo resíduos. Além disso, é importante que grande parte dos resíduos sejam atribuídos ao erro puro, ou seja, o erro experimental, e não à falta de ajuste do modelo. Em resumo, um bom modelo será aquele em que é obtido regressão significativa e falta de ajuste não significativa. Outra forma de conferir a qualidade do modelo obtido é avaliando o coeficiente de determinação R^2, uma vez que ele representa a parcela da variação que é explicada pela falta de ajuste. Quanto mais próximo de 1, maior será o ajuste do modelo em relação às respostas experimentais (TEÓFILO, 2006).

Após avaliada a qualidade do modelo, é possível otimizar as condições experimentais por meio da análise de superfície de resposta, ou seja, determinar na superfície de resposta os valores das variáveis em que é possível obter a melhor resposta para o experimento. A superfície de resposta é construída com base nos modelos matemáticos empíricos utilizando funções polinomiais lineares ou quadráticas de modo a descrever o comportamento do sistema, sendo possível deslocar e modelar a resposta até a otimização (TEÓFILO, 2006).

Além dos planejamentos fatoriais, também é possível construir um planejamento de misturas, cujo objetivo é avaliar um sistema multicomponente e suas devidas combinações de modo a otimizar a resposta de interesse. Contudo, diferente do planejamento fatorial que é composto por variáveis independentes, no planejamento de misturas tem-se variáveis dependentes (REIS, 1996). Desse modo, a resposta medida em um planejamento envolvendo

misturas irá depender das proporções de cada componente, sendo que a combinação dos *i* componentes não pode exceder 100%.

Para a escolha de um planejamento ideal de misturas deve ser considerado o número de variáveis e interações, sendo os tipos mais comuns, planejamento em rede simplex e planejamento centroide-simplex. No planejamento em rede simplex, o modelo matemático construído será calculado apenas com base nos componentes puros e misturas binárias. Já no centroide-simplex, devido à presença de um ponto central, é possível avaliar a interação de 3 componentes presentes na mistura (NOVAES *et al.*, 2018). Semelhante ao planejamento fatorial, a otimização das condições experimentais será realizada pela avaliação da qualidade do modelo e análise da superfície de resposta.

Neste trabalho foi empregado um planejamento de misturas centroide simplex para otimizar o solvente de extração. Após a escolha do solvente, foi realizado um planejamento composto central para otimizar as condições de extração, sendo tempo, temperatura e modo de contato os fatores estudados.

3.5.2 Análise de componentes principais (PCA)

A análise de componentes principais (PCA) pode ser considerada um método de redução da dimensionalidade, em que é possível projetar os dados multivariados em um espaço de menor dimensão, reduzindo o espaço original dos dados e separando as informações mais relevantes presentes em novas variáveis que facilitam a análise visual dos dados. Assim, a partir desta metodologia é possível visualizar e identificar as relações existentes entre as amostras e variáveis, e detectar a presença de amostras anômalas, uma vez que estas apresentam um comportamento diferente das demais amostras e serão identificadas na projeção dos dados (FERREIRA, 2015).

Estas novas variáveis que agora representam o novo espaço dimensional são denominadas componentes principais (PC). É importante frisar que, uma característica importante em relação a estas componentes é o fato de elas serem não correlacionadas e ortogonais entre si, ou seja, as informações presentes em uma componente não estão contidas na outra, eliminando assim a redundância de informações de um conjunto de dados. Basicamente, cada umas dessas variáveis ou PC's serão capazes de descrever a maior quantidade de informação possível em relação aos dados originais. Por exemplo, a primeira componente principal, PC1, será definida pela direção que descreve a máxima variância dos

dados iniciais, já a segunda componente principal, PC2, terá a direção de máxima variância dos dados no subespaço ortogonal à PC1, e assim acontece para todas as outras componentes escolhidas para explicar o conjunto de dados (FERREIRA, 2015).

A Figura 6, exemplifica a descrição acima, isto é, a primeira componente principal descreve o máximo espalhamento das amostras e variação dos dados, definidos pelo ajuste da reta no espaço e a segunda componente principal, perpendicular a primeira, explica a variância comum em sua direção, mas em quantidade menor que a primeira.

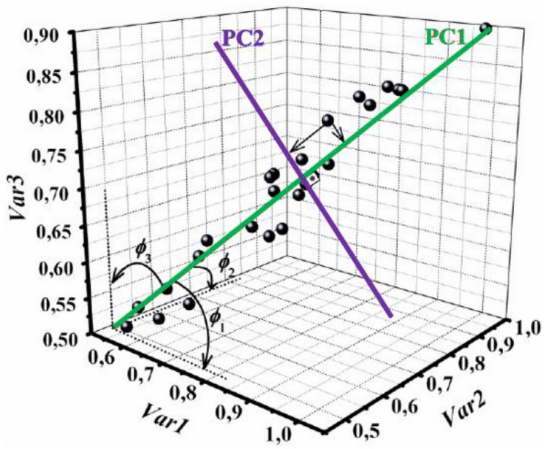


Figura 6 - Representação das componentes principais dispostas em um espaço com três variáveis (TEÓFILO, 2013).

Do ponto de vista matemático, a matriz $\mathbf{X}_{ij}$ que contém as informações dos dados será decomposta em três novas, uma matriz de escores **T**, uma matriz ortonormal de pesos (*loadings*) **P**, e a última matriz **E** contendo os resíduos, como mostrado na Equação 4 e na Figura 7.

$$\boldsymbol{X_{ij} = TP^t + E}$$

Equação 4

Sendo que **T** é a matriz de escores que relaciona as informações referentes as características das amostras (*i*), **P** é a matriz de loadings contendo as relações presentes nas variáveis (*j*) e, **E** contém os resíduos, variância contida nos dados que não são explicadas pelas componentes principais selecionadas.

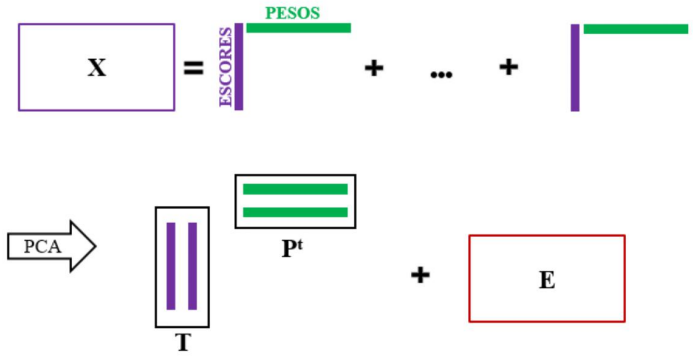


Figura 7 - Representação da decomposição matricial dos dados para um modelo PCA.

A PCA é um dos principais métodos quimiométricos e é a base de muitos métodos de classificação e calibração multivariada. Uma análise exploratória empregando a PCA é geralmente empregada como primeira análise dos dados, antes da construção de modelos supervisionados.

Uma etapa importante antes da construção de um modelo PCA ou qualquer outro modelo quimiométrico que deve ser mencionada é o pré-processamento dos dados. O principal objetivo dessa etapa é minimizar as variações indesejadas adquiridas durante a análise de dados que não são eliminadas naturalmente, e que podem influenciar na resposta final. Os pré-processamentos podem ser aplicados para minimizar as variações sistemáticas e aleatórias presentes nas amostras e reduzir variações contidas nas variáveis (FERREIRA, 2015).

Os pré-processamentos direcionados as amostras, geralmente são técnicas de alisamento ou correção de linha de base de modo a corrigir erros aleatórios (experimentais) ou minimizar variações sistemáticas quando necessário, sendo as principais, alisamento pelo método de Savitzky-Golay, derivadas, correção multiplicativa de espalhamento (MSC), padronização normal de sinal (SNV). Em relação as variáveis, os pré-processamentos mais comuns são centragem na média e autoescalamento, sendo o primeiro aplicado às variáveis contínuas, como por exemplo, dados espectrais, e o segundo aplicado às variáveis discretas, que possuem magnitudes e unidades diferentes e podem influenciar na distribuição dos dados (FERREIRA, 2015).

3.5.3 Modelagem flexível e independente por analogia de classes (SIMCA)

O método SIMCA é um dos primeiros métodos de modelagem de classe, e o mais importante e utilizado nas mais diversas áreas de pesquisas. Neste método, o principal objetivo é construir uma “regra” de classificação para um conjunto *m* de amostras conhecidas a medida

em que vai se obtendo maiores informações a respeito das variáveis responsáveis pelas características das amostras (VANDEN BRANDEN, 2005). Para isso, o método se baseia no modelo PCA, que por definição, é a descrição de um conjunto de dados com limites estatísticos e direcionados para a máxima variância dentro de um espaço de dados multivariados (OLIVERI, 2012).

Primeiramente, para a construção do modelo é realizado um modelo PCA para uma classe de amostras, em que o número de componentes principais sobre a classe centroide é definido, isto é, as projeções das amostras *i* pertencentes ao conjunto de treinamento são estabelecidas. Uma vez que as PC's são ortogonais, os limites para a hipercaixa do modelo SIMCA será na forma de um cilindro para a escolha de apenas uma componente, terá a forma de um retângulo quando forem utilizadas duas componentes para construção dos limites da classe, como mostrado na Figura 8 ou de um paralelepípedo quando forem escolhidas três ou mais componentes principais. As componentes não significativas do modelo também são avaliadas no espaço multidimensional possibilitando analisar informações a respeito da distribuição aleatórias das amostras (OLIVERI, 2012).

Em seguida, os limites da classe de interesse são estabelecidos por meio dos valores obtidos de T^2 de Hotelling e resíduos Q a um nível de significância definido previamente. Em geral é utilizado 95% (α=0,05), em que T^2 e Q residual fornecem informações a respeito das características das amostras que estão dentro e fora do modelo, respectivamente (MÁRQUEZ *et a*l., 2016). Além disso, a etapa de detecção de *outliers* é baseada nos altos valores obtidos para T^2 de Hotteling e Q residual, ou seja, amostras que apresentam altos valores de T^2 e/ou Q residual podem ser consideradas amostras anômalas e assim, serem retiradas do modelo.

Os cálculos para esses dois limites estatísticos são mostrados nas Equações 5 e 6.

$$T^2{}_{k,m,\alpha} = \frac{k(m-1)}{m-k} F_{k,m-k,\alpha}$$

Equação 5

onde *m* é o número de amostras totais utilizadas na construção do modelo PCA, *k* é o número de componentes principais escolhidas e α é o nível de significância definido.

$$Q_\alpha = \Theta_1 \left[\frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}}$$

Equação 6

em que,

$$\Theta_i = \sum_{j=k+1}^{n} \lambda_j^i \; para \; i = 1,2,3$$

Equação 7

$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^2}$$

Equação 8

Onde: c_α é o desvio padrão correspondente à parte superior (1- α) e λ_j^i são os autovalores obtidos na matriz **X** de covariância (WISE et al., 2006).

A distância entre uma amostra *i* à classe modelo *j* (d*ij*) para a atribuição de uma amostra na classe de interesse é definida baseada nos valores estatísticos reduzidos de T^2 e Q residual (WISE et al., 2006), como mostrado na Equação 9. O limite para uma amostra ser classificada como pertencente ao modelo é um semicírculo com raio 1, ou seja, d deve ser igual ou menor que 1.

$$d_{ij} = \sqrt{(Q_{r,i})^2 + (T^2{}_{r,i})^2}$$

Equação 9

em que,

$$Q_r = \frac{Q}{Q_{0,95}}$$

Equação 10

$$T^2{}_r = \frac{T^2}{T^2{}_{0,95}}$$

Equação 11

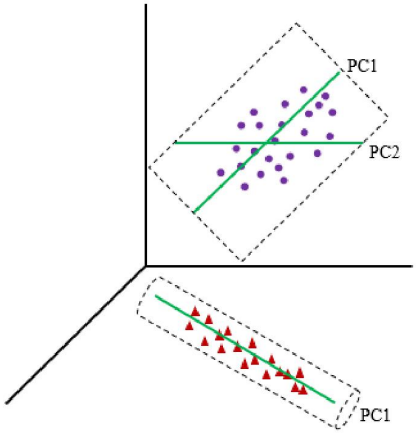


Figura 8 - Limites das hipercaixas do modelo SIMCA para uma e duas componentes principais (Adaptada de Ferreira, 2015).

Como mencionado anteriormente, um dos pontos importantes do SIMCA é o fato de conseguir obter informações a respeito das variáveis que são responsáveis e influenciam nas características das amostras. No entanto, essas características não são exploradas na literatura na área de autenticação de alimentos, os modelos são discutidos apenas em termos da classificação e não em termos das características das variáveis. Nesse contexto, uma das formas de determinar quais são as variáveis com maior influência e responsáveis para a construção dos limites da classe de interesse é por meio do cálculo do poder de modelagem da variável *j*. Basicamente, o cálculo é realizado comparando o desvio padrão residual (s_j) com o desvio padrão do conjunto de treinamento ($s_{j,y}$), como mostrado na Equação 12.

$$\psi_j = 1 - \frac{s_j}{s_{j,y}}$$

Equação 12

em que,

$$s_j = \sqrt{\frac{1}{Q}\sum_{q=1}^{Q}\frac{M}{M - A_q}\sum_{i=1}^{Q}\frac{\varepsilon_{ij}^2}{N_q - A_q - 1}}$$

Equação 13

$$\varepsilon_{ij} = Y_{ij} - T\,P^t$$

Equação 14

$$s_{j,y} = \sqrt{\sum_{q=1}^{Q}\sum_{i=1}^{N_q}\frac{(y_{ij} - \bar{y}_j)^2}{\sum_{q=1}^{Q} N_q - 1}}$$

Equação 15

$$\bar{y}_j = \sum_{q=1}^{Q}\sum_{i=1}^{N_q}\frac{y_{ij}}{\sum_{q=1}^{Q} N_q}$$

Equação 16

Sendo:

Q – número de classes presentes no modelo

N_q – número total de amostras para a classe de interesse

M – número total de variáveis

A_q – número de componentes principais escolhidas para o modelo

$\mathbf{y}_{ij}$ – vetor de dados (*i* amostra *vs*. *j* variáveis)

ε_{ij} – resíduos obtidos no modelo PCA

T – matriz de escores

P – matriz de *loadings* (pesos)

Uma variável terá maior influência sob um modelo à medida que seu poder de modelagem aumenta, ou seja, quanto mais próximo de 1, maior será a contribuição de uma variável na construção dos limites estatísticos para uma determinada classe de interesse (WOLD, 1977).

3.5.4 Modelagem flexível e independente por analogia de classes orientada aos dados (DD-SIMCA)

O DD-SIMCA é um método recente na literatura, apresentado em 2017, e é uma modificação do método SIMCA em que a principal mudança está relacionada à construção da área de aceitação de confiança que delineia a classe alvo das outras classes. A primeira etapa a ser realizada para construção do modelo é a decomposição da matriz de dados utilizando a PCA, conforme Equação 1. Em seguida, com os resultados obtidos na decomposição da PCA, são calculadas as distâncias de escores (SD), h_i, e as distâncias ortogonais (OD), v_i, para as amostras presentes no conjunto de treinamento, como mostrado nas Equações 17 e 18 (ZONTOV *et al.*, 2017) .

$$h_i = t_i^t (T^t T)^{-1} t_i = \sum_{a=1}^{A} \frac{t^2{}_{ia}}{\lambda_a}$$

Equação 17

$$v_i = \sum_{j=1}^{j} e_{ij}^2$$

Equação 18

Onde: λ_a, a=1, ..., A, são elementos diagonais da matriz $T^tT = \Lambda = diag\ (\lambda 1, ..., \lambda_A)$.

A distância total para cada amostra é calculada conforme a Equação 19, em que h_o e v_o são parâmetros de escala e N_h e N_v, números de graus de liberdade (ZONTOV *et al.*, 2017).

$$c = N_h \frac{h}{h_0} + N_v \frac{v}{v_0}$$

Equação 19

A terceira etapa define a área de aceitação para a classe de interesse, levando em consideração erros do tipo I (α), Equação 20, isto é, a rejeição incorreta da hipótese nula de que a amostra é autêntica (classe alvo).

$$c \leq c_{crit}(\alpha) = X^{-2}(1 - \alpha, N_h + N_v)$$

Equação 20

Nesse caso, (1- α) são os quantis da distribuição de chi-quadrado para (N_h + N_v) graus de liberdade (ZONTOV *et al.*, 2017).

Após as três etapas descritas acima, o modelo construído será capaz de classificar as novas amostras, e sua área de aceitação é representada por um gráfico da OD *vs*. SD, definido para um valor α, isto é, taxa de aceitação de falsos negativos. Em resumo, cada amostra do conjunto de treinamento é caracterizada por sua posição na área de aceitação e pode ser classificada ou não como sendo pertencente à classe de interesse (ZONTOV *et al.*, 2017).

Além do limite de decisão estabelecido para as amostras do conjunto de treinamento, também é construído para um valor y, um segundo nível de decisão para identificação de *outliers*, ou seja, é considerada a probabilidade de que pelo menos uma amostra do conjunto de dados ser erroneamente um *outlier*. A área de *outliers* depende do tamanho do conjunto de treinamento, quanto maior for a quantidade de amostras, maior será a área definida para *outliers* (ZONTOV *et al.*, 2017)

Por fim, estabelecidas as áreas de aceitações para o conjunto de treinamento e *outliers*, a última etapa é a classificação do conjunto teste a partir do modelo construído. Os resultados obtidos nesta etapa são apresentados na área de aceitação. Além disso, nesta etapa é possível calcular erros do tipo II (β), ou seja, é calculada a taxa de aceitação de falsos positivos (ZONTOV *et al.*, 2017).

3.5.5 Mínimos quadrados parciais de uma classe (OCPLS)

Semelhante ao DD-SIMCA, o OCPLS, proposto em 2013, também é um método de modelagem de classe recente na literatura, no entanto, para ele é ajustado um modelo de mínimos quadrados parciais (PLS) que correlacionam as medidas experimentais com um vetor composto de valores 1 (XU *et al.*, 2013). Além disso, diferente do SIMCA que considera somente a variância das variáveis independentes, no OCPLS, a variância do conjunto de dados é definida agora pelas variáveis latentes (VL's) que consideram as variâncias explicadas tanto das variáveis dependentes quanto das independentes. O número de VL's a serem escolhidas para um modelo são estimadas por validação cruzada (XU *et al.*, 2014b).

Para a construção de um modelo OCPLS são calculadas duas medidas de distância, a primeira é o valor T² de Hotelling baseada na distância dos escores (SD) e a segunda, os resíduos absolutos centralizados (ACR), representadas nas Equações 21 e 22, respectivamente. O ACR pode assumir distribuição normal com média igual a 0.

$$T^2 = \sum_{i=1}^{K} \frac{(t_i - \bar{t}_i)^2}{s_{t,i}^2}$$

Equação 21

em que $\overline{t_i}$ e $s_{t,i}^2$ são a média e variância das amostras da i-ésima VL, nesta ordem, e *K* é o número de VL's significantes.

$$ACR = \left|1 - \hat{y}_j - \widehat{\mu_e}\right|$$

Equação 22

onde $\hat{y}_j$é a resposta ajustada do objeto *j* e $\widehat{\mu_e}$ é a média dos erros de treinamento.

O desvio padrão do modelo residual pode ser estimado por validação cruzada (CV), conforme Equação 23.

$$\hat{\sigma}_e = \sqrt{\sum_{i=1}^{N} \frac{(1 - \hat{y}_i - \widehat{\mu_e})^2}{N - 1}}$$

Equação 23

em que *N* é o total de objetos retirados durante a validação cruzada e ŷ é a resposta prevista do *i*-ésimo objeto retirado (XU et al., 2014).

Para um nível de confiança, α, os limites de confiança superiores (UCLs) para T^2 e ACR podem ser calculados, como mostrado nas Equações 24 e 25.

$$T^2{}_{UCL} = \frac{(n^2-1)K}{n(n-K)} F_{\alpha(K,n-K)}$$

Equação 24

$$ACR_{UCL} = Z_{\frac{\alpha}{2}} \hat{\sigma}_e$$

Equação 25

em que $F_{\alpha(K,n-K)}$ é o ponto crítico superior da distribuição F com (*K, n-K*) graus de liberdade e $Z_{\frac{\alpha}{2}}$ é o ponto crítico superior da distribuição normal padrão.

A distância dos escores (SD) é a distância do objeto ao centro da classe ocupado pelas VLs significativas e ACR é a dispersão dos resíduos projetados nos vetores dos coeficientes de regressão. Com os valores de SD e ACR são possíveis atribuir as amostras em quatro grupos distintos: 1) amostra regular – pequeno SD e ACR, 2) pontos de alavancagem (*leverage*) bons – grande SD e pequeno ACR, 3) outliers de classes – pequeno SD e grande ACR e 4) pontos de alavancagem ruins – grande SD e grande ACR. Sendo que as atribuições 2, 3 e 4 podem ser considerados diferentes tipos de *outliers*. Os limites para os valores de SD e ACR são calculados para níveis de confiança definidos previamente (XU et al., 2014).

3.5.6 Conjunto de treinamento e teste

Uma das etapas mais importante na construção de modelos supervisionados é a escolha das amostras para o conjunto de treinamento, de modo a garantir a maior representatividade e variabilidade dos dados na construção de modelos supervisionados (classificação/calibração). Sob esse ponto de vista, o algoritmo de Kennard-Stone (KS) (KENNARD; STONE, 1969) é usualmente aplicado como ferramenta de seleção de amostras para o conjunto de treinamento (MIAW, 2018).

A seleção das amostras por esse algoritmo baseia-se na distância Euclidiana de cada par (p, q) de amostras, conforme mostrado na Equação 26, ou seja, as duas primeiras amostras selecionadas são aquelas que apresentam a maior e a menor distância em relação ao ponto médio central, a terceira amostra será aquela que apresenta maior distância em relação às duas primeiras, e assim acontece até preencher o número total de amostras pré-estabelecido. Em geral, 2/3 do total de amostras são definidos para o conjunto de treinamento.

$$d_x(p,q) = \sum_{j=1}^{J} [x_p(j) - x_q(j)]^2 \; p,q \in [1,M]$$

Equação 26

em que *J* representa o número de covariáveis e M é o número de amostras. As amostras remanescentes são atribuídas ao conjunto de validação/teste (SOUSA *et al*., 2011; FERREIRA *et al*., 2022)

3.5.7 Seleção de variáveis

A seleção de variáveis é utilizada com objetivo de identificar as variáveis que são, para um conjunto específico, mais importantes e contribuem para fornecer um modelo com melhor desempenho, além de possibilitar a interpretação química dos dados mais facilmente. Existem diferentes métodos para seleção de variáveis e, basicamente, diferem entre si pela forma como buscam um subconjunto e determinam as variáveis mais relevantes (TEÓFILO, 2013). Para o desenvolvimento deste trabalho foi utilizado o algoritmo de seleção dos preditores ordenados (OPS) como método de seleção de variáveis (TEÓFILO, *et al*., 2009).

O OPS faz uso dos vetores informativos como *VIP scores* e coeficientes de regressão, a respeito das variáveis mais importantes na matriz de dados **X** e seleciona as variáveis mais preditivas para a construção do novo modelo. Na Figura 9 estão apresentadas as etapas da seleção de variáveis utilizando o algoritmo.

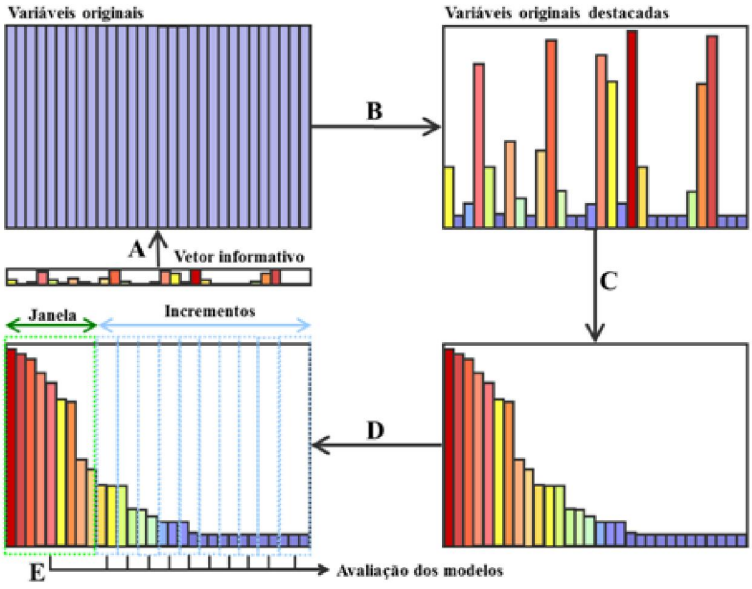


Figura 9 - Esquema das etapas da seleção de variáveis utilizando o OPS (TEÓFILO, 2013).

A etapa A consiste na seleção dos vetores informativos sendo obtidos a partir de cálculos realizados com as variáveis dependentes e independentes. Na etapa B, a partir de vetor prognóstico obtido, as variáveis independentes originais são organizadas em ordem decrescente de valor absoluto, sendo que os maiores valores nos vetores informativos indicam a posição das variáveis originais independentes mais importantes. Em seguida, a etapa C consiste na construção e avaliação dos modelos por validação cruzada (método *leave-N-out*) e seleção de janelas em que serão adicionados pequenos incrementos de novas variáveis. E, por fim, a última etapa, D, os conjuntos de variáveis avaliados (janela inicial e janela + incrementos) são comparados por meio de parâmetros de qualidade, sendo que o conjunto de variáveis que apresenta melhores parâmetros é o que contém as variáveis com melhor capacidade preditiva sendo, portanto, o conjunto selecionado (TEÓFILO, 2013).

3.5.8 Fusão de dados

A abordagem de fusão de dados pode ser definida como sendo a fusão de blocos de dados provenientes de diferentes técnicas analíticas instrumentais ou variáveis físico-químicas em um único bloco de dados de modo a construir apenas um modelo. O principal objetivo da fusão é eliminar as informações redundantes e explorar a sinergia dos conjuntos de dados, ou seja, fazer uso das informações complementares dos diferentes instrumentos e construir um modelo com melhor desempenho (ASSIS *et al*., 2020). As estratégias utilizadas para realizar a fusão de dados são divididas em três níveis diferentes, nível baixo, médio e alto (BIANCOLILLO *et al*., 2014), como mostrado na Figura 10.

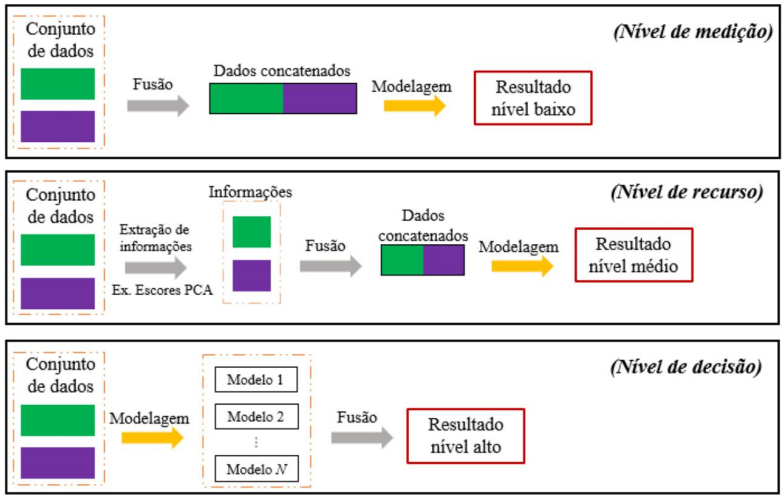


Figura 10 - Representação das estratégias de fusão de dados (Adaptada de COCCHI *et. al*., 2019).

Na fusão de nível baixo, os diferentes blocos de dados originais são concatenados após a etapa de pré-processamento, cujo principal objetivo é modelar os dados para melhorar a capacidade preditiva do modelo. A maior vantagem dessa estratégia é a possibilidade de interpretar os resultados em termos das variáveis originais (COCCHI *et. al.*, 2019). Na literatura é possível encontrar diferentes trabalhos em que foi utilizado a fusão de nível baixo no desenvolvimento de modelos para autenticação de alimentos e bebidas (BORRÀS *et al.*, 2015). Isso pode ser justificado pelo fato da maioria das técnicas a serem utilizadas para análises de alimentos, tais como, NIR, MIR e UV-Vis apresentarem informações complementares possibilitando a construção de um modelo com maior capacidade preditiva e de simples interpretação (BORRÀS *et al.*, 2015). A fusão de nível baixo também pode ser chamada de nível de medição.

Para a fusão de nível médio ou nível de informação, é necessária uma etapa de modelagem antes da fusão dos dados. Essa etapa tem por objetivo a extração das informações mais importantes dos conjuntos de dados de modo separado, como por exemplo, a decomposição da matriz por PCA (COCCHI *et. al.*, 2019). Após essa etapa de extração de "recursos", os dados são fundidos e o modelo construído. É mais aplicado quando se tem um conjunto de variáveis muito grande, desse modo, ao realizar a etapa de extração de informações é possível construir um modelo com menor número de variáveis e, ainda assim, manter as informações importantes a respeito do conjunto de dados (BORRÀS *et al.*, 2015).

Por último, na fusão de nível alto ou nível de decisão, os dados só serão concatenados após ser obtido um modelo individual para cada bloco de dados, geralmente um modelo supervisionado. O foco dessa estratégia está apenas no resultado, a interpretação das variáveis originais não é investigada (COCCHI, 2019).

Neste trabalho foi aplicado a fusão de nível baixo com intuito de avaliar a sinergia dos conjuntos de dados obtidos por TXRF, FTIR, UV-Vis e PS-MS, assim como, o desenvolvimento de modelos com melhor desempenho. É importante frisar que após a fusão das matrizes originais é necessário realizar o autoescalamento dos dados de modo a fornecer o mesmo "peso" para as variáveis dos diferentes blocos e assim garantir a homogeneidade do modelo.

3.5.9 Figuras de mérito

As figuras de mérito são calculadas para avaliar o desempenho do modelo de classificação construído em termos da avaliação preditiva. Elas são descritas por meio dos parâmetros sensibilidade (*SEN*), especificidade (*ESP*) e eficiência (*EFC*), em que a sensibilidade expressa a taxa de acerto do modelo em classificar corretamente amostras pertencentes a classe de interesse, e a especificidade, a capacidade do modelo em rejeitar corretamente amostras não pertencentes a classe interesse. A fim de estabelecer uma medida global entre os dois parâmetros é calculado a eficiência do modelo através da média geométrica entre eles (OLIVERI *et al.*, 2011). Os cálculos para as figuras de méritos são mostrados nas Equações 27-29.

$$SEN(\%) = \frac{VP}{VP + FN}\ 100$$

Equação 27

$$ESP\ (\%) = \frac{VN}{VN + FP} 100$$

Equação 28

$$EFC\ (\%) = \sqrt{SEN\ x\ ESP}$$

Equação 29

onde:

VP, verdadeiro positivo, é o número de amostras da classe-alvo classificadas corretamente dentro da classe;

FN, falso negativo, é o número de amostras da classe-alvo classificadas erroneamente fora da classe de interesse;

VN, verdadeiro negativo, é o número de amostras da classe não-alvo classificadas corretamente fora da classe de interesse;

FP, falso positivo, é o número de amostras da classe não-alvo classificadas erroneamente na classe de interesse.

4. METODOLOGIA

4.1 Amostras

Os grãos de café verde utilizados no trabalho foram cedidos pela Federação de Cafeicultores do Cerrado, sendo proveniente das cidades de Araguari, Araxá, Campos Altos, Carmo do Paranaíba, Coromandel, Ibia, Indianópolis, João Pinheiro, Monte Carmelo, Patos de Minas, Patrocínio, Perdizes, Pratinha, Presidente Olegário, Rio Paranaíba, Romaria, Santa Rosa da Serra, Serra do Salitre, Tapira, Unaí e Varjão de Minas, totalizando 100 amostras. Na Figura 11 está apresentada a região do Cerrado Mineiro onde os cafés possuem indicação geográfica. Além disso, foram também cedidas 30 amostras de três regiões produtoras de café, sendo elas provenientes das regiões do Caparaó, Mogiana e Sul de Minas.

Em relação ao preparo, as amostras, inicialmente, foram trituradas/moídas em um moedor de café doméstico (Modelo Hamilton Beach 150W) com auxílio de nitrogênio líquido, e posteriormente, armazenadas em tubos Falcon de 15mL e microtubos. Todas as amostras foram mantidas sob refrigeração até o momento de uso.

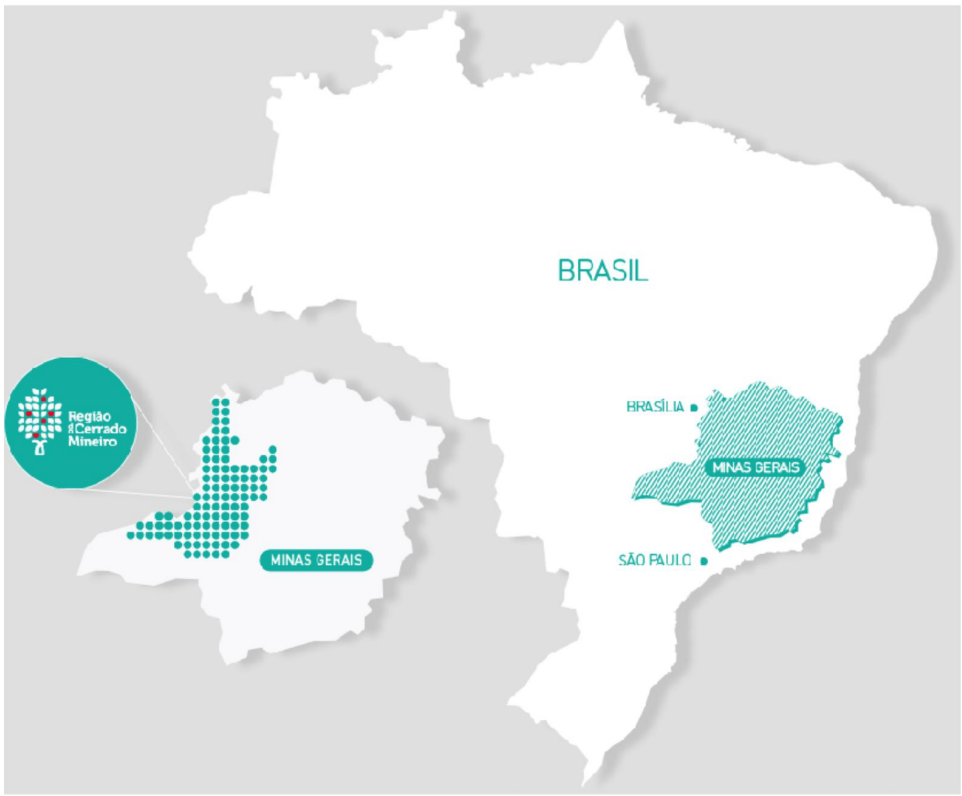


Figura 11 - Localização da indicação geográfica dos cafés da Região do Cerrado Mineiro (Fonte: Federação dos Cafeicultores do Cerrado Mineiro).

4.2 Procedimentos de Extração

Na etapa de extração das amostras, foi realizado o planejamento de misturas e planejamento fatorial para avaliar o efeito do solvente na extração dos compostos orgânicos presentes no café e otimizar os parâmetros temperatura, tempo e modo de contato no processo de extração.

4.2.1 Planejamento de misturas Centróide Simplex

Foram adicionados 0,5 g dos grãos de café triturados previamente, em 10 mL do solvente puro ou mistura. A escolha dos solventes água, etanol e metanol foi baseada na literatura (MOREIRA *et al.*, 2014). A princípio, essas misturas foram colocadas em banho de ultrassom durante 20 minutos, à 40 °C. Os experimentos foram realizados em duplicata, conforme o planejamento de misturas centroide simplex, mostrado na Figura 12, para três componentes, sendo um total de 14 experimentos apresentados na Tabela 3.

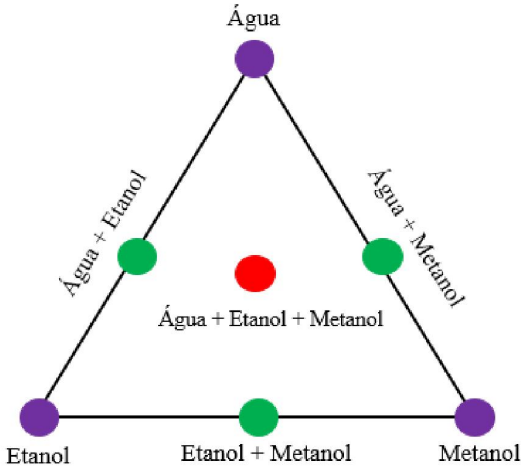


Figura 12 - Planejamento de misturas Centróide Simplex para três componentes.

Tabela 3 - Proporções dos solventes utilizados no processo de extração dos compostos.

Extratos	Água	Etanol	Metanol
1	1,00	0,00	0,00
2	1,00	0,00	0,00
3	0,00	1,00	0,00
4	0,00	1,00	0,00
5	0,00	0,00	1,00
6	0,00	0,00	1,00
7	0,50	0,50	0,00
8	0,50	0,50	0,00
9	0,50	0,00	0,50
10	0,50	0,00	0,50
11	0,00	0,50	0,50
12	0,00	0,50	0,50
13	0,34	0,33	0,33
14	0,34	0,33	0,33

4.2.2 Planejamento composto central

Uma vez definido a melhor condição de solvente, foi realizado o planejamento composto central, no intuito também de otimizar a extração dos componentes presentes nos grãos de café, em relação a temperatura, tempo e modo de contato entre o grão de café e o solvente. Como é um planejamento em que são avaliados 3 fatores, foram necessários a realização de no mínimo 8 experimentos. Foram realizados experimentos no ponto central em triplicata para as variáveis temperatura e tempo, para o modo de contato estático e ultrassom. Todos os experimentos foram realizados de forma aleatória e o modelo matemático obtido foi avaliado por meio da ANOVA. A Tabela 4, mostra as condições iniciais avaliadas para otimização da extração, já a Tabela 5 apresenta a matriz de planejamento com as variáveis codificadas.

Tabela 4 - Parâmetros avaliados na otimização da extração.

Parâmetros	Nível inferior (-)	Nível superior (+)	Ponto Central	
Temperatura (°C)	25	60	40	40
Tempo (min)	10	30	20	20
Modo de contato	Estático	Ultrassom	Estático	Ultrassom

Tabela 5 - Matriz de planejamento com as variáveis codificadas.

Experimentos	Temperatura	Tempo	Modo de contato estático	Modo de contato ultrassom
1	-1	-1	-1	+1
2	+1	-1	-1	+1
3	-1	+1	-1	+1
4	+1	+1	-1	+1
5	0	0	-1	+1
6	0	0	-1	+1
7	0	0	-1	+1
8	0	-1	-1	+1
9	0	1	-1	+1
10	-1	0	-1	+1
11	1	0	-1	+1

A construção do modelo matemático, tanto para o planejamento de misturas quanto para o planejamento composto central, foi realizada por meio das equações 30 e 31, respectivamente, onde y é a resposta prevista em função das variáveis avaliadas.

$$y= b_1x_1+b_2x_2+b_3x_3+b_{12}x_{12}+b_{13}x_{13}+b_{23}x_{23}+b_{123}x_{123}$$

Equação 30

$$y = b_0 + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2$$

Equação 31

As respostas obtidas nos planejamentos foram as somas das absorbâncias dos compostos extraídos dos grãos de café, obtidas na análise por espectroscopia de absorção na região do UV-Vis. Os dados do planejamento foram analisados no software *Design Expert* (v. 11). Os extratos foram empregados nas análises de absorção na região do UV-Vis e na espectrometria de massas.

4.3 Análises

4.3.1 Fluorescência de raios X por reflexão total (TXRF)

Para as análises via TXRF foi utilizado um espectrômetro de fluorescência de raios X por reflexão total, modelo S2 PICOFOX (Bruker, Alemanha), mostrado na Figura 13. O equipamento contém um tubo de Molibdênio (Mo) como fonte de excitação, um monocromador

multicamada e um detector SDD (*Silicon Drift Detector*). O padrão interno utilizado foi solução padrão de Gálio (Ga) 1000 mg L^{-1} e discos de quartzo de 30 mm de diâmetro e espessura de 3,0 ± 0,1 mm como porta amostra. Antes dos discos serem utilizados, eles foram limpos utilizando detergente alcalino 5% v/v (Sigma-Aldrich, EUA) e acetona pura PA (Vetec, Brasil). Com os discos limpos e secos, foram adicionados 10 µL de solução de silicone em isopropanol (Serva, Alemanha).

Para o preparo de amostra, foram pesados 30 mg dos grãos de café moídos anteriormente e adicionados 790 µL de água deionizada, 200 µL de Triton (solução emulsificante) 5% v/v e 10 µL de solução de Ga 1000 mg L^{-1}, resultando em uma solução final de 1,0 mL. Os microtubos foram agitados em vórtex para melhor homogeneização. Após agitação, rapidamente, foi retirada uma alíquota de 10 µL da solução e depositada no centro dos discos, em seguida, os discos foram levados à estufa por 10 min a 80 °C. Após esse tempo e com os discos já em temperatura ambiente (Figura 14), foram realizadas as análises no espectrômetro com tempo de leitura de 500 s para cada amostra.

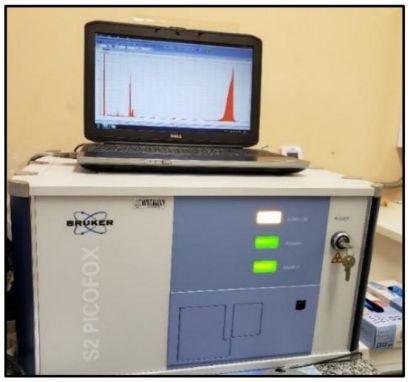


Figura 13 - Equipamento utilizado nas análises via TXRF.

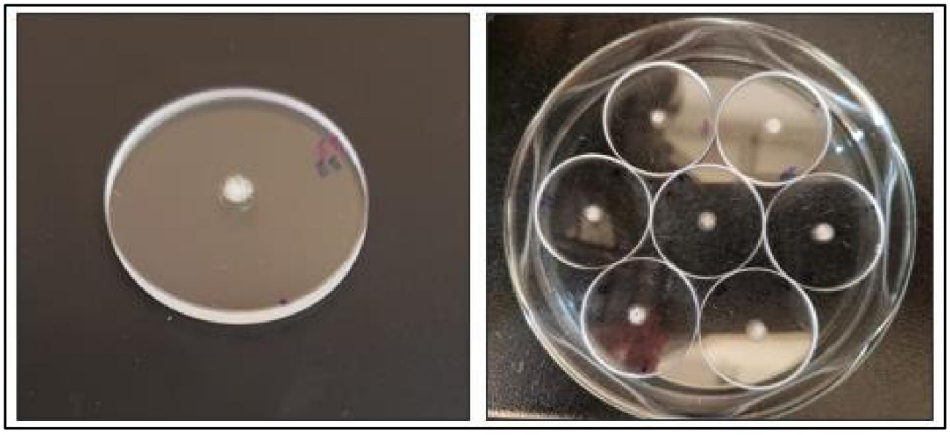


Figura 14 - Discos prontos para serem analisados.

4.3.2 Espectroscopia na região do infravermelho médio (FTIR)

As análises foram realizadas utilizando um espectrômetro na região do infravermelho médio com transformada de Fourier, modelo FTIR Frontier (Perkin Elmer, Massachussetts, EUA), com acessório de reflectância total atenuada (ATR) equipado com cristal de diamante, mostrado na Figura 15a. Os grãos de café verde moídos anteriormente foram adicionados, em pouca quantidade, sobre o cristal de ATR com auxílio de uma espátula até o preenchimento total da abertura do cristal, conforme Figura 15b. Em seguida, os espectros foram adquiridos na faixa de 4000 cm^{-1} a 650 cm^{-1} com resolução de 4 cm^{-1} e 32 varreduras.

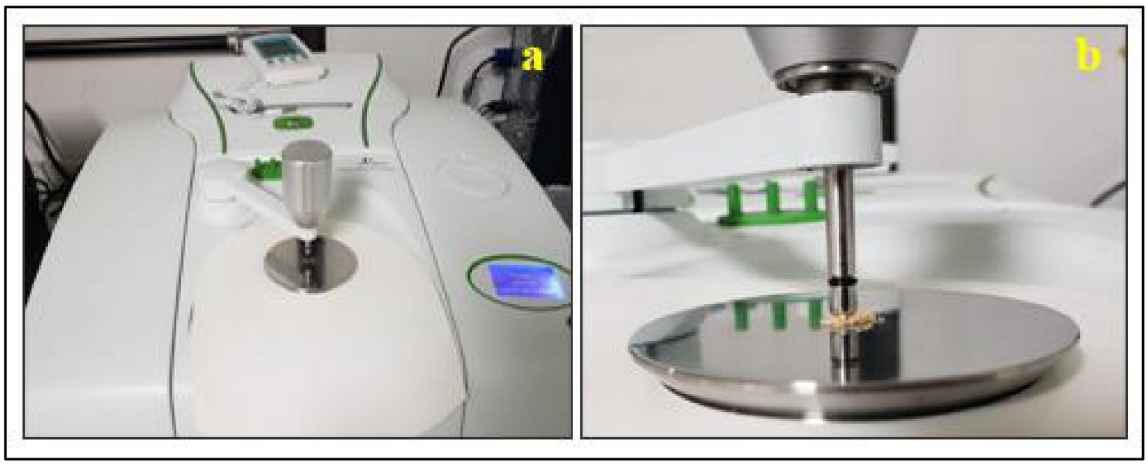


Figura 15 - (a) Equipamento ATR-FTIR e (b) amostra adicionada no cristal pronta para análise.

4.3.3 Espectrofotometria na região do ultravioleta e visível (UV-Vis)

Para as análises de absorção na região do UV-Vis foram empregados os extratos obtidos nas condições estabelecidas pelo planejamento de misturas e composto central. Sendo assim, a extração foi realizada pesando 0,5 g dos grãos de café triturado e adicionando-os em 10 mL de água, essa mistura foi colocada em aquecimento à 60 °C, e em banho ultrassônico por 30 minutos, como apresentado na Figura 16a. As extrações foram realizadas em duplicata. Posteriormente, as amostras foram filtradas e armazenadas em tubos Falcon de 15 mL, conforme mostrado na Figura 16b, e mantidas sob refrigeração.

Para as medidas de absorção, foram retiradas alíquotas de 50 μL dos extratos, e estas foram avolumadas com água até um total de 10 mL. Em seguida, aproximadamente 1 mL dessa solução foi colocada em uma cubeta de quartzo com caminho ótico de 1 cm e analisadas em

um espectrofotômetro UV-Vis, modelo Cary 60 (Agilent), na faixa de trabalho entre 800 nm e 200 nm, Figura 17.

Figura 16 - (a) Extração dos compostos presentes em grãos de café sob condições estabelecidas no planejamento de experimentos e (b) amostras filtradas prontas para análise.

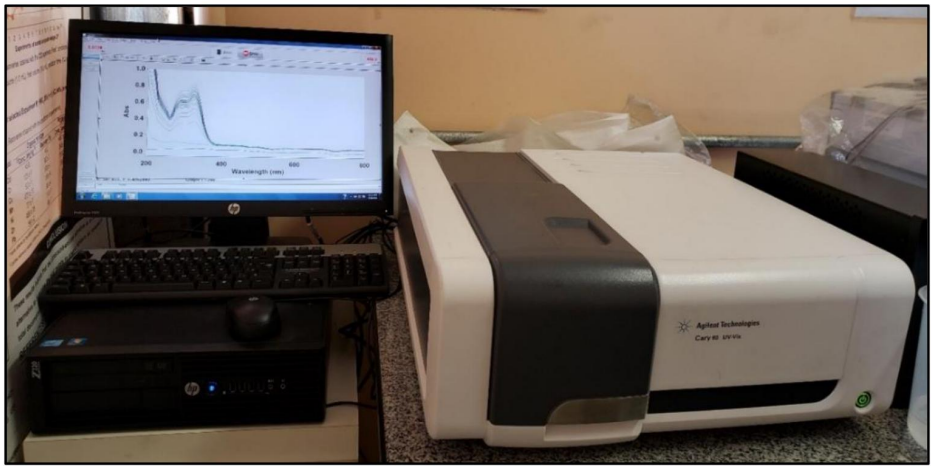


Figura 17 - Espectrofotômetro na região do UV-Vis utilizado nas análises.

4.3.4 Espectrometria de massas por *paper-spray* (PS-MS)

Para realização das análises, foram retirados 100 µL do extrato preparado conforme etapa anterior, e adicionados em 900 µL de metanol grau HPLC (J. T. Baker, EUA). As medidas foram executadas em um espectrômetro de massas Thermo Fisher LCQ-Fleet com analisador de massas de baixa resolução do tipo *Ion Trap* (San Jose, California, EUA. A fonte de ionização utilizada foi o *paper spray*. Para isso, foi construído uma base (Figura 17) utilizando um clipe metálico do tipo jacaré e um suporte universal que permite movimentação nos planos x, y e z, de modo a ajustar o posicionamento do papel em direção a entrada do espectrômetro.

As análises foram realizadas em triplicata, adicionando 25 µL de amostra (extrato) no papel cromatográfico (número 1 da Whatman), cortados na forma de triângulos equiláteros, com dimensões de 1,0 cm. A aquisição dos espectros foi feita no modo positivo e negativo do espectrômetro e em uma faixa de m/z variando entre 100 e 1000 Da, com temperatura do capilar à 275 °C, potencial de 5 kV e com uma distância, entre a ponta do papel e a entrada do espectrômetro, fixada a 0,5 cm. Na Figura 18, é possível visualizar a imagem do equipamento e o suporte onde foram realizadas as análises

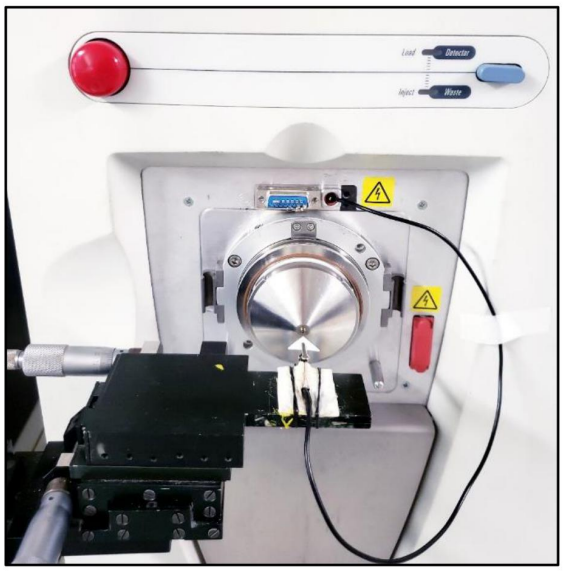


Figura 18 - Equipamento e suporte onde foram realizadas as análises de massas via *paper spray*.

4.4 Tratamento dos dados

O processo de tratamento dos dados, esquematizado na Figura 19, seguiu as seguintes etapas:

1) construção da matrizes de dados; 2) pré-processamento dos dados de acordo com as características do conjunto; 3) análise exploratória dos dados por PCA; 4) separação das amostras em conjuntos de treinamento e teste; 5) detecção de outliers, sendo que a retirada de amostras atípicas foi baseada nos altos valores de T^2 e Q residual; 6) construção dos modelos quimiométricos; 7) cálculo das figuras de mérito; 8) aplicação do método de seleção de variáveis OPS; 9) construção de novos modelos somente com as variáveis selecionadas; e10) cálculo das novas figuras de mérito.

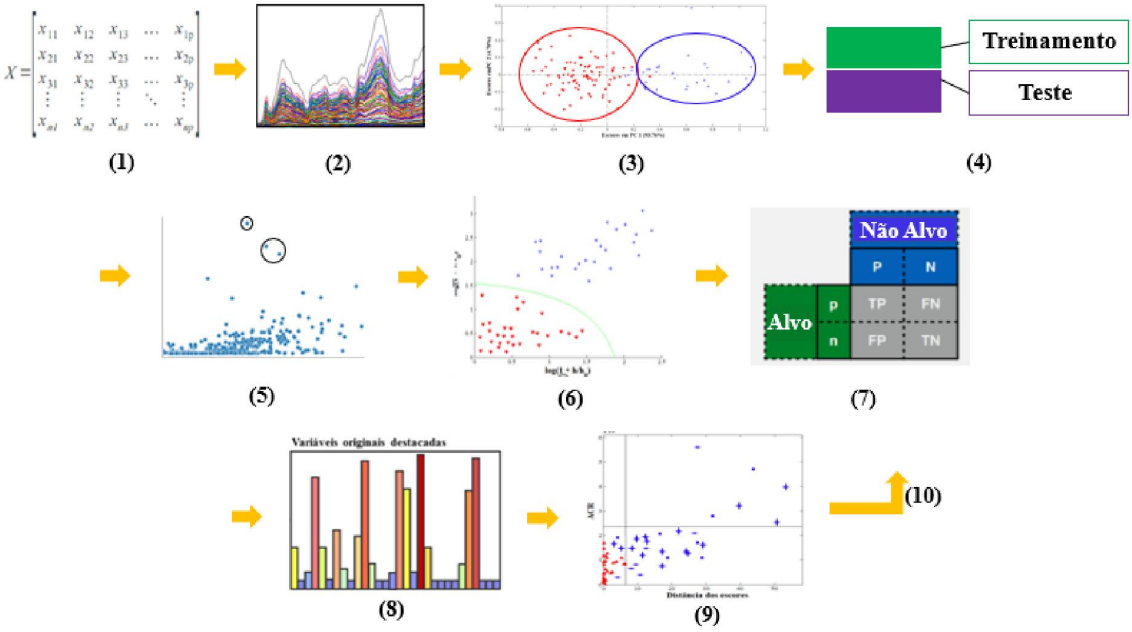


Figura 19 - Representação das etapas realizadas na construção dos modelos

O tratamento dos dados e a construção dos modelos quimiométricos foram realizados utilizando o software MATLAB, versão 7.10 (The MathWorks, Natick, MA, EUA) com pacote PLS Toolbox, versão 5.2.2 (Eigenvector Research, Manson, WA, EUA) para os modelos PCA e SIMCA. A construção dos modelos DD-SIMCA foi realizada a partir da rotina fornecida por Zontov e colaboradores (2017), disponível em https://github.com/yzontov/dd-simca e os modelos OCPLS foram construídos partir da rotina cedida por Lu Xu e colaboradores (2013; 2014).

Os dados obtidos por TXRF foram autoescalados e os dados por PS-MS foi realizado a centragem na média. Em relação as matrizes com os dados de FTIR e UV-Vis, foram testados diferentes pré-processamentos, como a) SNV seguido da primeira derivada e alisamento Savitsky-Golay; b) *baseline*; c) MSC e d) SNV de modo a obter o melhor o resultado. Para a fusão de dados, foi realizado o pré-processamento ideal para cada técnica de modo individual e em seguida, as matrizes foram concatenadas e autoescaladas.

A construção do *dataset* foi feita com os dados obtidos para cada técnica em blocos individuais, em que as linhas da matriz **X** correspondiam às amostras (N=130) e as colunas, às variáveis. Os dados foram separados em dois conjuntos distintos, o primeiro denominado conjunto treinamento/calibração, contendo 70 amostras provenientes do Cerrado Mineiro e o segundo denominado conjunto teste/validação, contendo 60 amostras, sendo 30 amostras do Cerrado Mineiro e 30 amostras de regiões próximas. É importante ressaltar que as amostras do

Cerrado presentes no conjunto de treinamento foram selecionadas utilizando o algoritmo Kennard-Stone, as demais foram colocadas no conjunto teste.

A detecção de outliers foi realizada analisando os parâmetros estatísticos, T^2 e resíduos Q. Amostras que contém altos valores de T^2 e Q residual são amostras com comportamento atípicos e diferem-se da grande maioria das amostras presentes no conjunto de dados, logo podem ser consideradas outliers e assim, serem retiradas do modelo. É importante ressaltar que a retirada de amostras anômalas deve obedecer ao limite de 22% do valor total de amostras presentes no conjunto de treinamento de modo a evitar o sobreajuste do modelo.

Em todos os modelos construídos, foi aplicado o método de seleção de variáveis OPS visando obter modelos com melhor performance. A rotina para execução do OPS foi cedida por Teófilo e colaboradores, e pode ser encontrada no link https://lqta.iqm.unicamp.br/portugues/Downloads.html#ops.

O desempenho do modelo foi avaliado em termos das figuras de mérito sensibilidade, especificidade e eficiência, calculados conforme descrito na seção 3.4.9.

5. RESULTADOS

5.1 Otimização das condições de extração

Para as análises realizadas por espectrometria de massas e de absorção na região do UV-Vis, foi necessário realizar a extração dos componentes presentes nos grãos de café verde. Com o intuito de otimizar as condições para o uso do solvente, foram testados água, etanol e metanol para avaliar qual resultaria em um melhor rendimento de extrato obtido. A princípio, tempo e temperatura foram fixados em 20 min e 40 °C, respectivamente, e a extração foi realizada sob banho ultrassônico. Como resposta, foi avaliada a soma das absorbâncias dos espectros de absorção dos extratos de café referentes as principais substâncias presentes no café obtidos na análise por espectrofotometria no UV-Vis. Na Tabela 6, estão apresentados os valores encontrados para a soma das absorbâncias, resposta considerada no planejamento.

Tabela 6 - Resultado do planejamento de misturas centroide simplex para 3 variáveis.

Ensaio	Água	Etanol	Metanol	R_1
1	1,00	0,00	0,00	96,89
2	1,00	0,00	0,00	87,90
3	0,00	1,00	0,00	0,00
4	0,00	1,00	0,00	0,54
5	0,00	0,00	1,00	5,65
6	0,00	0,00	1,00	10,42
7	0,50	0,50	0,00	41,45
8	0,50	0,50	0,00	42,40
9	0,50	0,00	0,50	52,61
10	0,50	0,00	0,50	47,08
11	0,00	0,50	0,50	0,00
12	0,00	0,50	0,50	0,00
13	0,33	0,34	0,33	38,14
14	0,33	0,34	0,33	35,33

O modelo matemático obtido para esse planejamento está representado pela equação 32 abaixo:

$$y = 92{,}40x_1 + 8{,}04x_3 + 202{,}74x_{123}$$

Equação 32

Em que x_1, x_3 e x_{123} representam os solventes água, metanol e suas interações, nesta ordem. Na Figura 20 estão apresentados os gráficos de contorno e superfície de resposta obtida para o planejamento. É possível observar que os maiores valores de resposta são obtidos quando se usa somente água ou uma mistura de água: etanol e água: metanol, sendo a água em proporção maior que 50% v/v para realizar as extrações.

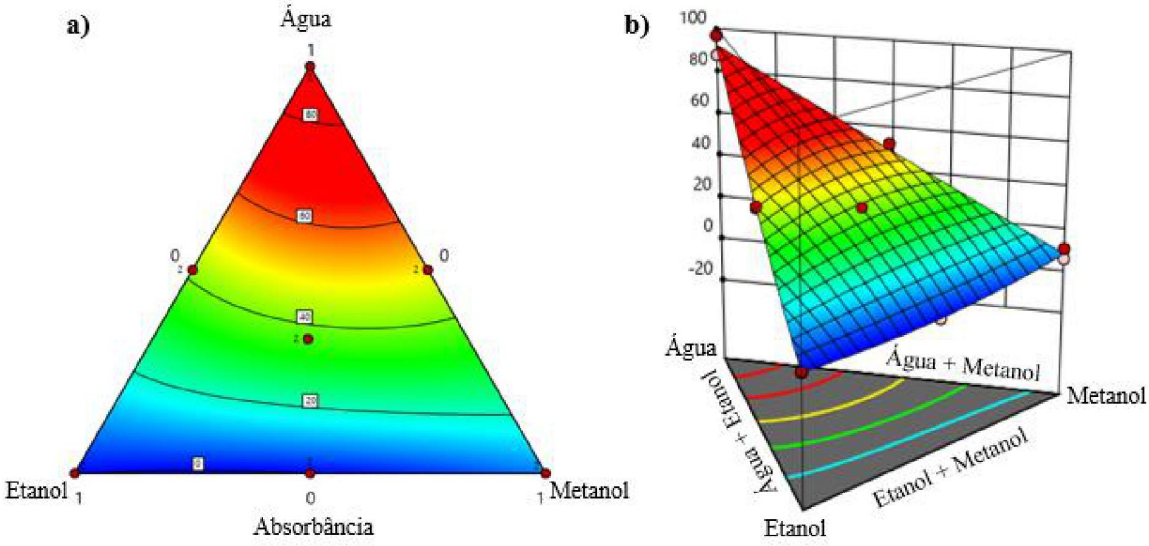


Figura 20 – Gráficos obtidos para o planejamento de misturas, (a) contorno e (b) superfície de resposta.

Uma vez que foi definido que o solvente ideal para a extração seria água, foi realizado o planejamento composto central para otimização dos parâmetros tempo, temperatura e modo de contato, assim como, avaliar o efeito de interações entre eles. Na Tabela 7 encontram-se os resultados obtidos para o planejamento. No entanto, como um dos fatores era categórico e não, numérico, foram construídos dois planejamentos, um considerando as absorbâncias obtidas para o modo de contato estático (R_1) e o outro considerando as absorbâncias obtidas para o modo de contato em banho de ultrassom (R_2).

Tabela 7 - Resultado do planejamento composto central para condições ideais de extração.

Ensaio	Temperatura (°C)	Tempo (min)	R_1	R_2
1	25	10	66,68	74,60
2	60	10	70,30	88,73
3	25	30	67,43	79,98
4	60	30	76,09	94,55
5	40	20	43,74	72,49
6	40	20	38,32	67,72
7	40	20	40,72	69,41
8	40	10	43,70	71,38
9	40	30	55,71	78,59
10	25	20	47,13	55,69
11	60	20	74,58	81,32

O modelo matemático obtido para o planejamento composto central está apresentado nas equações 33 e 34 abaixo, sendo a equação 33 para o planejamento em que a extração foi realizada em modo estático e a equação 34, extração realizada em ultrassom.

$$y = 44{,}44 + 6{,}62x_1 + 22{,}60x_1^2$$

Equação 33

$$y = 69{,}32 + 9{,}05x_1 + 11{,}98\,x_2^2$$

Equação 34

Onde x_1 é o efeito da temperatura e x_2 o efeito do tempo. A equação 33 mostra que para o modelo obtido com o modo de contato estático, o tempo não teve efeito significativo para o modelo e, portanto, nesse caso, não influenciou nos valores de absorbância obtidos. Já o modelo obtido quando a extração foi realizada em ultrassom, equação 34, tanto a temperatura quanto o tempo foram significativos, mas o efeito de interação entres os dois, não.

Ao avaliar a superfície de resposta na Figura 21, para ambos os modelos, observa-se que de fato não há interação significativa entre os efeitos temperatura-tempo, uma vez que o resultado obtido com a mudança da temperatura não depende do nível da variável tempo. Entretanto, quando avaliado qual modelo obteve melhores respostas, pode-se observar que a extração realizada sob banho de ultrassom forneceu maiores respostas. Dessa maneira, as

melhores condições de temperatura e tempo foram avaliadas utilizando a superfície de resposta em que o modo de contato era o ultrassom.

Assim sendo, ao analisar o gráfico de superfície de resposta encontrado na Figura 21-b, é possível perceber que os resultados mais altos de absorbância foram obtidos quando a temperatura e o tempo estão em níveis superiores. Portanto, as condições estabelecidas para a extração dos grãos de café verde foram, temperatura igual a 60 °C durante 30 min em banho ultrassônico. Os resultados obtidos pela ANOVA do planejamento de misturas e do planejamento fatorial estão apresentados nas Tabelas 1A-3A do anexo.

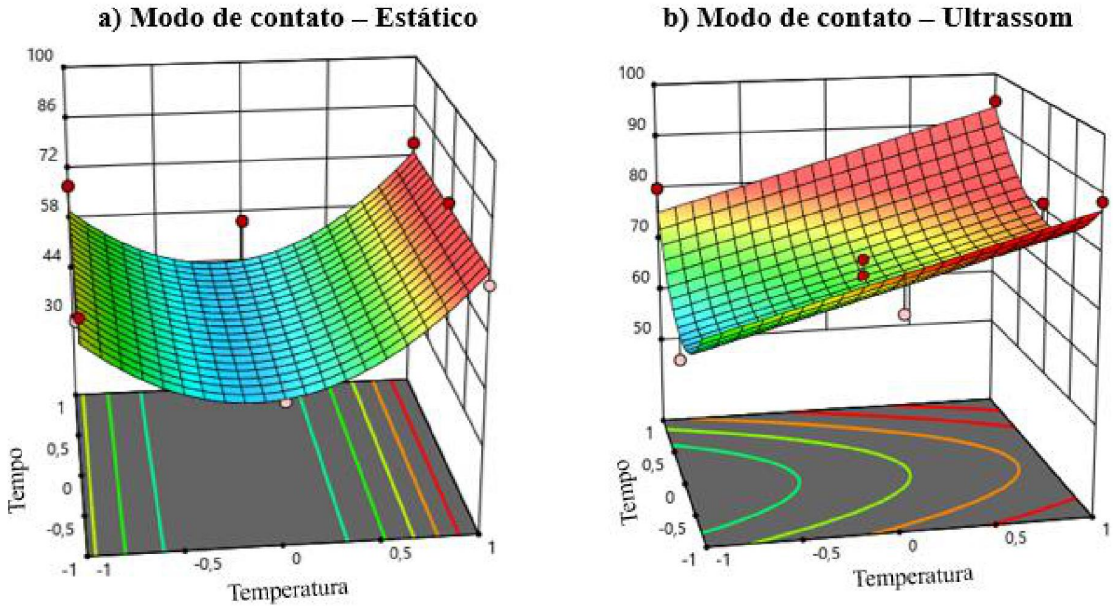


Figura 21 - Superfícies de resposta obtidas para o planejamento composto central para otimização da extração nos modos de (a) contato estático e (b) sob banho ultrassônico.

5.2 Construção dos modelos de classificação – classe única

5.2.1 TXRF

Para desenvolvimento dos modelos utilizando os dados obtidos por TXRF, foi construída uma matriz 130x15, sendo 130 amostras e 15 a quantidade de elementos analisados pela técnica. Os elementos analisados foram P, S, Cl, K, Ca, Ti, Cr, Mn, Fe, Ni, Cu, Zn, Br, Rb e Sr. A quantidade de cada elemento foi expressa em $mg.kg^{-1}$. Apesar de ser uma técnica espectroscópica, os dados obtidos não foram avaliados por meios dos seus espectros e sim pelas concentrações obtidas, portanto, não foi realizado nenhum tipo de pré-processamento em

relação a linha de base, apenas o autoescalamento dos dados. Os maiores valores de concentração foram obtidos para os elementos K, Ca, P, S, e Cl o que corrobora com dados obtidos em estudos anteriores (ASSIS, 2018).

Foram construídos modelos de classificação utilizando os métodos SIMCA, DD-SIMCA e OCPLS e calculadas as figuras de mérito (sensibilidade, especificidade e eficiência) para o conjunto de treinamento (70 amostras) e teste (60 amostras). Na etapa de detecção de outliers, cinco amostras apresentaram altos valores de T^2 e Q residual, sendo consideradas anômalas e retiradas do conjunto de treinamento para a construção do modelo SIMCA. Para os modelos DD-SIMCA e OCPLS não foram detectados outliers. É importante frisar que como no conjunto de treinamento estão presentes somente amostras da classe alvo, o desempenho do modelo foi avaliado apenas em termos da sensibilidade. Os resultados obtidos estão apresentados na Tabela 8.

Tabela 8 - Figuras de mérito para os métodos de modelagem de classe única a partir das análises por TXRF, sendo (a) dados com as 15 variáveis e (b) com as 10 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	***Teste***		
Métodos	**PC's/LV's**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	8	1,00	0,89	0,20	0,42
DD-SIMCA	8	0,97	0,96	0,17	0,40
OCPLS	3	0,95	1,00	0,13	0,37
b) Seleção de variáveis					
SIMCA	4	1,00	0,96	0,17	0,40
DD-SIMCA	4	0,99	1,00	0,07	0,29
OCPLS	4	0,93	1,00	0,10	0,32

Os dados acima mostram que os modelos obtidos com os dados originais de TXRF, Tabela 8-a, apesar de conseguir classificar corretamente as amostras provenientes do Cerrado Mineiro, obtendo alta sensibilidade para o conjunto de treinamento e teste, não obteve bons resultados em diferenciar e classificar as amostras que não eram do Cerrado Mineiro. Para os três métodos de modelagem de classes utilizados o modelo não foi capaz de prever e classificar corretamente as amostras provenientes de outras regiões como sendo classe não-alvo. Desse modo, pode-se concluir que com os dados obtidos somente pela técnica em questão, não foi possível obter modelos robustos e satisfatórios.

Com o intuito de melhorar o desempenho do modelo, foi realizada a seleção de variáveis no conjunto de dados, os resultados estão apresentados na Tabela 8-b. Os elementos que tiveram maior influência para o modelo foram P, S, Cl, K, Ca, Fe, Cu, Br, Rb e Sr. Entretanto, mesmo com o uso de seleção de variáveis, o modelo individual com a técnica de TXRF não apresentou bom desempenho em classificar as amostras não-alvo corretamente para os métodos escolhidos neste trabalho.

5.2.2 FTIR

Os espectros apresentados na Figura 22 foram obtidos na região do infravermelho médio na faixa de 4000-650 cm^{-1}. Para a construção dos modelos, foi selecionado o intervalo em que normalmente aparecem as absorções de ligações presentes nos diferentes componentes do café, esse intervalo é denominado região de *fingerprint* e se encontra na faixa entre 1800-700 cm^{-1}, Figura 22-b.

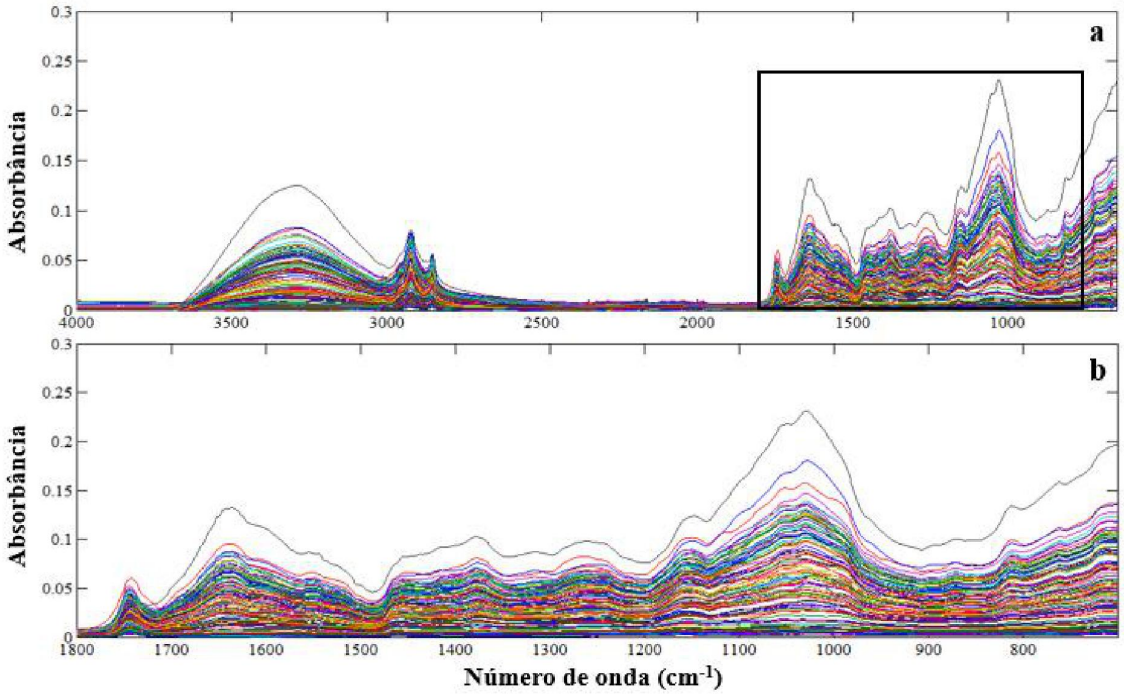


Figura 22 - (a) Espectros brutos obtidos na região do infravermelho médio e (b) faixa espectral utilizada para construção dos modelos de classificação.

Observando a Figura 22b, nota-se que os espectros são bem parecidos sendo que as diferenças encontradas estão relacionadas apenas as intensidades das bandas em algumas regiões, como por exemplo, a região entre 1800-1600 cm^{-1}, atribuídos a presença de ácidos clorogênicos com pico na região próxima a 1750 cm^{-1} , referente a ligação carbonila, e em 1636

cm^{-1}, relacionado à ligação C=C (MONJE et al., 2018). As bandas observadas entre 1100-1000 cm^{-1}, podem estar relacionadas à deformação de compostos heterocíclicos (AZEVEDO, 2015).

Uma vez definido que a construção dos modelos seria realizada com a região de *fingerprint*, a matriz de dados foi de 130x1100, sendo 130, as linhas correspondentes a quantidade de amostras e 1100 as colunas relacionadas as absorbâncias por números de onda.

Antes das construções dos modelos de classificação foram testados diferentes pré-processamentos de modo a corrigir linha de base e diminuir ruídos e assim melhorar o perfil espectral do conjunto de dados. O pré-processamento SNV (*Standard Normal Variate*) foi o escolhido para dar sequência no desenvolvimento dos modelos, seguido da centragem dos dados na média. Na etapa de retirada de outliers, quatorze amostras foram consideradas atípicas e retiradas do conjunto de treinamento para construção dos modelos. Os resultados para os modelos de classe única estão apresentados na Tabela 9.

Tabela 9 - Figuras de mérito para os métodos de modelagem de classe única a partir das análises por FTIR, sendo (a) dados com as 1100 variáveis e (b) com as 79 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	***Teste***		
Métodos	**PC's/LV's**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	4	0,93	0,56	0,63	0,59
DD-SIMCA	4	0,99	0,50	0,37	0,43
OCPLS	6	0,96	0,54	0,40	0,46
b) Seleção de variáveis					
SIMCA	4	0,94	0,32	0,57	0,43
DD-SIMCA	4	0,94	0,46	0,47	0,47
OCPLS	6	0,93	0,50	0,50	0,50

Analisando a Tabela 9-a é possível notar que para o conjunto de treinamento foi obtido alta sensibilidade, isto é, o modelo foi capaz de atribuir corretamente as amostras do Cerrado na classe-alvo. Já para o conjunto teste, a eficiência do modelo foi baixa para os três modelos construídos, ou seja, as informações fornecidas pela técnica não foram suficientes para discriminar e diferenciar amostras alvo das não-alvo.

Após aplicação do método de seleção de variáveis OPS, foram selecionadas 79 variáveis e os resultados dos modelos são mostrados na Tabela 9-b. Pode ser observado que os resultados

obtidos após aplicação do método de seleção foram semelhantes aos encontrados com os dados originais e não melhorou o desempenho do modelo.

5.2.3 PS-MS

Os espectros de massas dos extratos de café foram obtidos no modo positivo e negativo do equipamento em um intervalo de 100 a 1000 m/z e o resultado foi expresso em termos da abundância relativa (%) dos íons presentes na amostra. A Figura 23 mostra o espectro médio de massas para os grãos de café do Cerrado obtido no modo positivo do equipamento. Nota-se que, os sinais mais intensos foram as m/z 176, 317 e 381, estes sinais foram relatados em estudos anteriores e atribuídos à trigonelina $[M\text{-}K]^+$, ácido linolênico $[M\text{-}K]^+$ e sacarose $[M\text{-}K]^+$, respectivamente. Outros íons, como m/z 104, 138 e 195 foram identificados como sendo colina $[M\text{-}H]^+$, trigonelina $[M\text{-}H]^+$ e cafeína $[M\text{-}H]^+$, nesta ordem, sendo todas essas substâncias importantes na composição do café (GARRETT, *et al*., 2013). Além disso, é possível observar a baixa intensidade no sinal da cafeína, o que já era esperado, uma vez que em grãos de café da espécie arábica, o teor de cafeína é menor (ASSIS, 2018). Outro ponto importante de frisar é que a trigonelina, um dos sinais mais intensos obtidos, é um dos compostos responsáveis por dar sabor e aroma com notas de amargor característicos para a bebida de café (MONTEIRO, 2005).

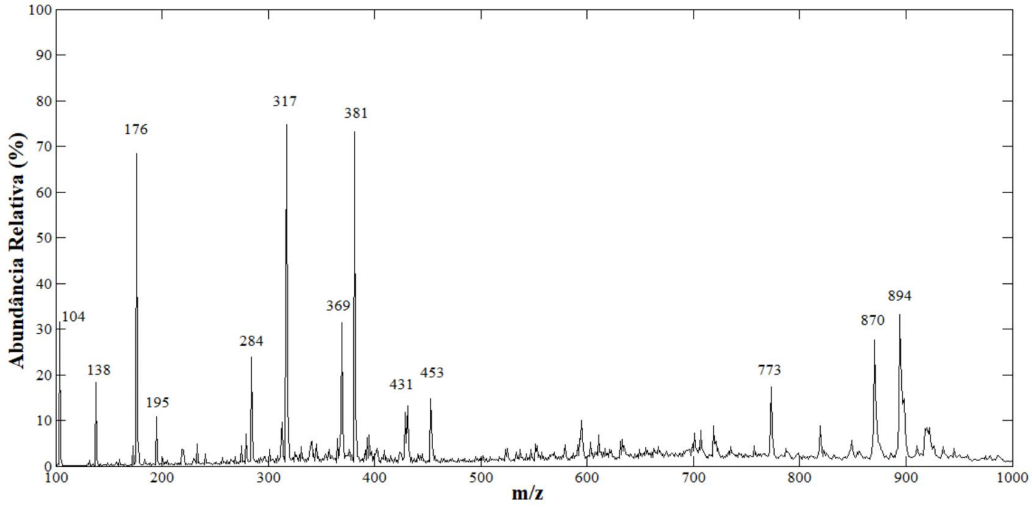


Figura 23 - Espectro médio PS-MS (+) dos extratos dos grãos de café verde do Cerrado.

A matriz de dados construída para esta técnica foi de 130x901 e os dados foram centrados na média. Não foram identificadas amostras atípicas presentes no conjunto de dados.

No geral, os modelos construídos para os dados de PS-MS(+), não tiveram alto desempenho. Na Tabela 10-a são apresentados os resultados do desempenho do modelo para os dados originais e após a seleção de variáveis com o método OPS. Observa-se que a melhor eficiência encontrada foi para o método SIMCA, sendo de 63% para os dados originais e 77% para os dados com seleção de variáveis, sendo que foram definidas 188 variáveis como as mais importantes para o conjunto de dados.

Ao analisar a Tabela 10-b para o método SIMCA, nota-se que houve uma diminuição da sensibilidade no conjunto de treinamento, ou seja, houve um aumento da taxa de falsos negativos, isto quer dizer que o novo modelo não foi capaz de classificar todas as amostras do Cerrado como sendo da classe-alvo. No entanto, para o conjunto teste, houve um aumento na especificidade do modelo, as amostras não-alvo foram classificadas corretamente como sendo não-alvo, e dessa forma, houve um ligeiro aumento no desempenho do modelo obtendo 77% de eficiência. Para os outros métodos DD-SIMCA e OCPLS, mesmo aplicando-se a seleção de variáveis não foi possível obter modelos satisfatórios.

Tabela 10 - Figuras de mérito para os métodos de modelagem de classe única a partir das análises por PS-MS(+), sendo (a) dados com as 901 variáveis e (b) com as 188 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	***Teste***		
Métodos	**PC's/LV's**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	6	1,00	1,00	0,40	0,63
DD-SIMCA	6	0,99	1,00	0,30	0,55
OCPLS	7	0,97	1,00	0,27	0,52
b) Seleção de variáveis					
SIMCA	4	0,95	1,00	0,60	0,77
DD-SIMCA	4	0,97	1,00	0,23	0,48
OCPLS	4	0,93	1,00	0,17	0,41

O resultado do espectro médio obtido para os grãos de café do Cerrado no modo negativo do espectrômetro de massas está apresentado na Figura 24. Observa-se que o sinal m/z mais intenso foi para o íon m/z 353, identificado como ácido cafeoilquínico $[M-H]^-$. Os outros íons encontrados, também são compostos presentes no café, sendo o sinal m/z 191 atribuído a ácido quínico $[M-H]^-$, m/z 515 a ácido di-cafeoilquínico $[M-H]^-$ e o íon m/z 707 como sendo também ácido cafeoilquínico, no entanto, como aduto $[2M-H]^-$ (GARRETT *et al.*, 2012). Todos

os ácidos mencionados acima fazem parte dos ácidos clorogênicos e são encontrados em grandes quantidades no café (MONTEIRO, 2005).

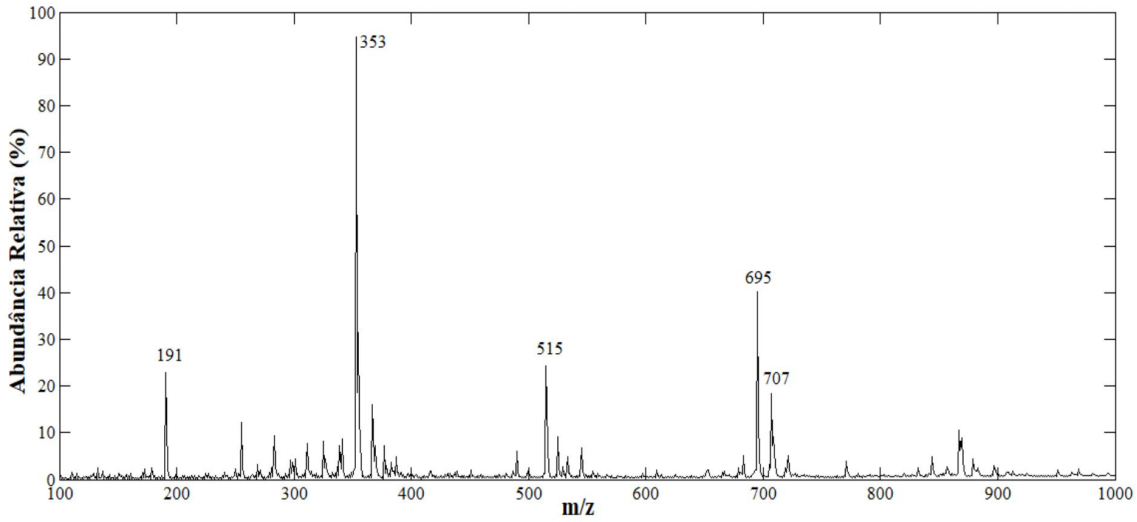


Figura 24 - Espectro médio PS-MS (-) do extrato dos grãos de café verde do Cerrado.

Na Tabela 11 estão apresentados os valores obtidos para as figuras de mérito para essa matriz. O desempenho dos modelos seguiu comportamento semelhante aos modelos obtidos para os dados no modo positivo, o método SIMCA apresentou melhor eficiência para os dados originais e para os dados após a seleção de variáveis. Neste modelo, os valores obtidos para a eficiência do modelo SIMCA foram de 52% e 66% para os dados originais e dados após aplicação da seleção de variáveis, respectivamente.

Tabela 11 - Figuras de mérito para os métodos de modelagem de classe única a partir das análises por PS-MS(-), sendo (a) dados com as 901 variáveis e (b) com as 204 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	***Teste***		
Métodos	**PC's/LV's**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	6	1,00	1,00	0,27	0,52
DD-SIMCA	6	1,00	1,00	0,23	0,48
OCPLS	5	0,97	1,00	0,07	0,26
b) Seleção de variáveis					
SIMCA	4	0,97	1,00	0,43	0,66
DD-SIMCA	4	0,97	1,00	0,20	0,45
OCPLS	5	0,99	1,00	0,23	0,48

5.2.4 UV-Vis

Os espectros para os extratos dos grãos de café verde foram adquiridos na faixa de 800-200 nm. Para a construção dos modelos foi escolhida a faixa entre 450-230 nm, região que apresenta informações importantes sobre a composição das amostras. Na Figura 25 estão apresentados os espectros obtidos para os extratos. É possível notar que há a presença de duas bandas características nas regiões de 280 nm e 320 nm, que são atribuídas as substâncias trigonelina e ácidos clorogênicos, nesta ordem (MOREIRA *et al.*, 2014). A trigonelina e os ácidos clorogênicos são compostos bastante importantes na composição do café, uma vez que são responsáveis por fornecer características sensoriais mais refinadas, influenciando assim na qualidade final da bebida (MONTEIRO, 2005).

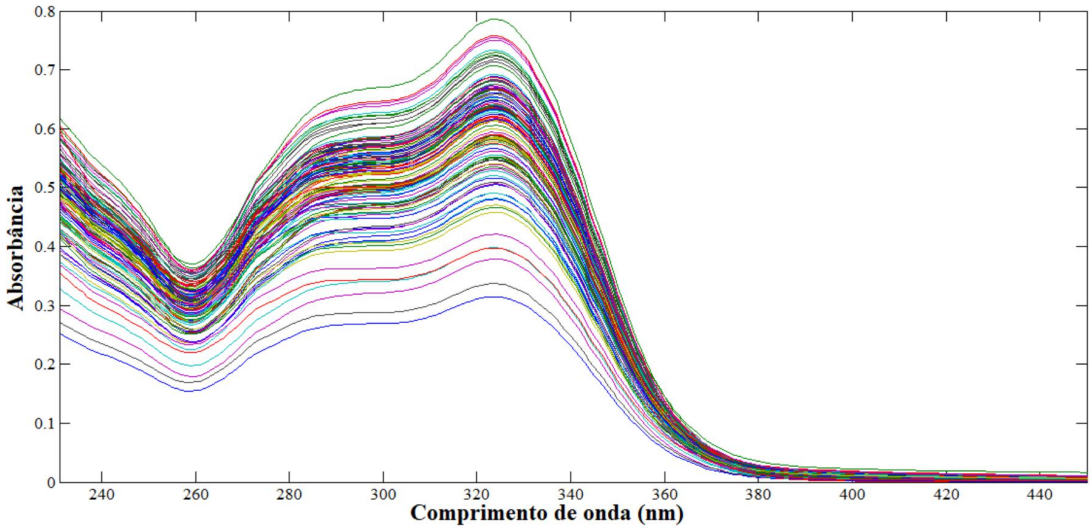


Figura 25 - Espectros de absorção na região do UV-Vis.

Semelhante as outras técnicas, foram construídos modelos SIMCA, DD-SIMCA e OCPLS com os dados originais e com aplicação do método de seleção de variáveis. A matriz de dados, após a escolha da faixa espectral, apresentou dimensão de 130x220. O pré-processamento escolhido para esse conjunto de dados foi o SNV, seguido da centragem dos dados na média.

Observando os resultados obtidos, mostrados na Tabela 12, é possível notar que os melhores modelos construídos a partir dos blocos individuais para cada técnica foram para os dados obtidos por meio da espectroscopia na região do UV-Vis. Foi possível obter aproximadamente 90% de eficiência para o conjunto teste do modelo SIMCA, e mais de 90%

para DD-SIMCA e OCPLS, estes dois últimos ao utilizar o método de seleção de variáveis OPS, Tabela 12-b.

Tabela 12 - Figuras de mérito para os métodos de modelagem de classe única a partir das análises por UV-Vis, sendo (a) dados com as 220 variáveis e (b) com as 34 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	***Teste***		
Métodos	**PCs/VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	2	1,00	0,89	0,90	0,89
DD-SIMCA	3	1,00	0,96	0,63	0,78
OCPLS	3	0,96	0,96	0,43	0,65
b) Seleção de variáveis					
SIMCA	2	1,00	0,96	0,70	0,82
DD-SIMCA	3	1,00	1,00	0,87	0,93
OCPLS	2	0,94	1,00	0,87	0,93

Como os modelos construídos com a técnica de absorção na região do UV-Vis apresentaram melhores resultados, os modelos serão detalhados a seguir.

Modelo SIMCA

A construção do modelo SIMCA é realizada por meio da decomposição matricial baseado no modelo de PCA para uma determinada classe, em que é possível obter informações a respeito das amostras por meio do gráfico de escores e a respeito das variáveis por meio do gráfico de *loadings*. Nesse contexto, foi realizada a análise exploratória dos dados por PCA de modo a visualizar o perfil e a estrutura dos dados. Nas Figuras 26 e 27 estão apresentados o gráfico dos escores obtido para o modelo PCA para duas componentes principais que explicam 98,46% da variância total dos dados e o gráfico de *loadings* para a PC1, uma vez que esta componente é responsável por explicar praticamente toda a variância presente no conjunto de dados, explicando 93,76% da variância cumulativa.

No gráfico de PC1 *vs.* PC2, é possível observar que existe uma tendência de agrupamento das amostras. No lado negativo e região central da PC1 estão distribuídas as amostras provenientes do Cerrado Mineiro (▼), nesse caso, a classe-alvo. Já no lado positivo

de PC1 estão agrupadas amostras oriundas de outras regiões, classe não-alvo (∗). Nota-se claramente o agrupamento entre as duas classes de amostras, e, assim, os resultados obtidos pela PCA já indicam que modelos baseados no SIMCA apresentarão desempenho de classificação satisfatórios.

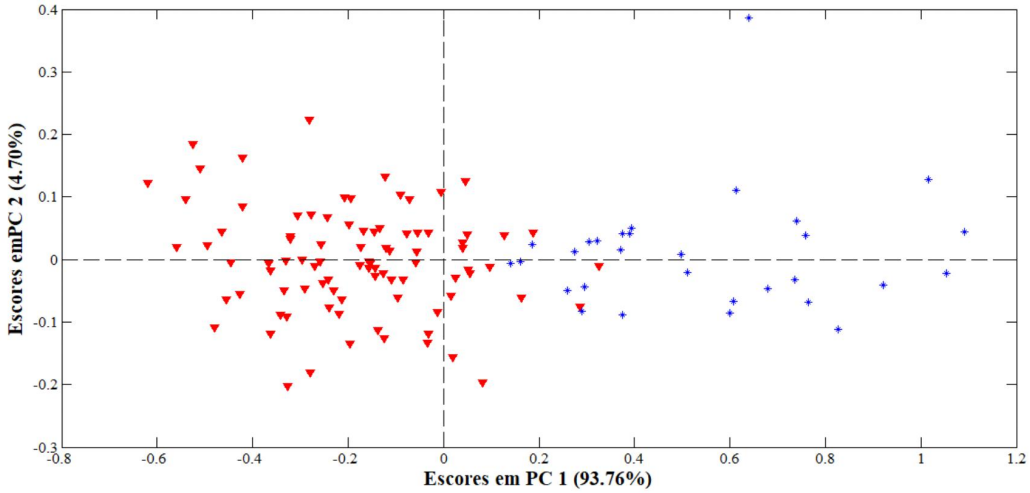


Figura 26 – Análise das componentes principais para duas PCs com os dados obtidos por UV-Vis para amostras provenientes do Cerrado Mineiro (▼) e amostras de outras regiões (∗).

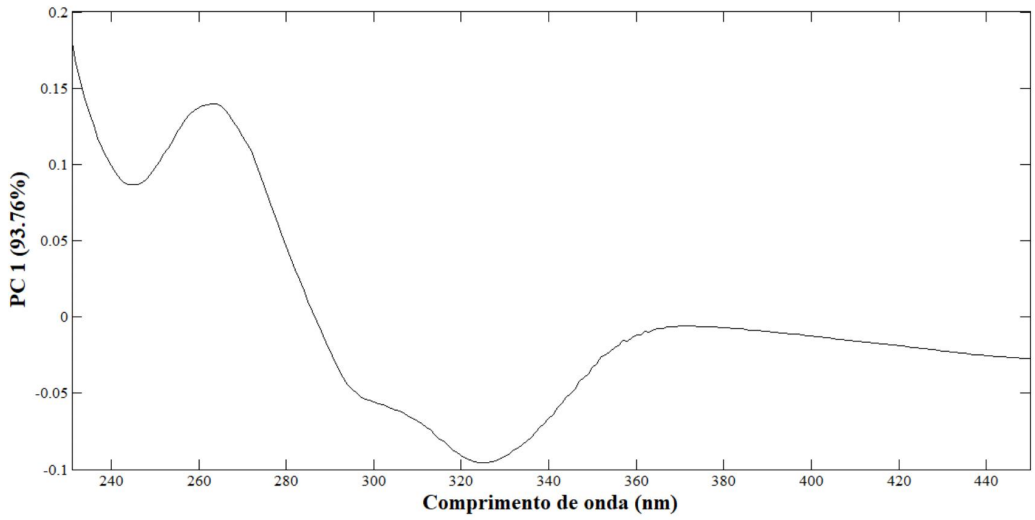


Figura 27 - *Loadings* em PC1 para o modelo PCA construído.

O gráfico de pesos na PC1 vs. comprimento de onda traz informações a respeito das variáveis que foram importantes e influenciaram na tendência de agrupamento das amostras. Observa-se na figura 27 que existem duas bandas características que foram significativas para o comportamento das amostras. No lado positivo de PC1, as variáveis que influenciam as amostras de outras regiões se encontram na faixa espectral entre 260-272 nm, e no lado negativo

de PC1 as variáveis mais importantes para as amostras do Cerrado estão em torno da região entre 320-326 nm. Como mencionado anteriormente, estas duas regiões são associadas às transições eletrônicas das substâncias trigonelina (~272 nm) e ácidos clorogênicos (~326 nm) (MOREIRA et al., 2014).

Para a construção dos limites de aceitação para a classe de interesse, o SIMCA faz uso dos valores de T² de Hotelling e Q residual. Na figura 28 é mostrado o gráfico de T² de Hotelling *vs.* Q residual com limites estabelecidos em 95% de confiança. Nota-se que existem algumas amostras com altos valores de T² de Hotelling, tal situação pode ser explicada pelo fato de as amostras serem provenientes de diferentes cidades do Cerrado Mineiro e, como a região do Cerrado é extensa, uma maior variabilidade de amostras pode ser obtida. Para a construção do modelo SIMCA foi realizada a detecção de outliers baseado nos limites construídos e estabelecendo o limite de 22%, em que foram retiradas seis amostras do conjunto de treinamento.

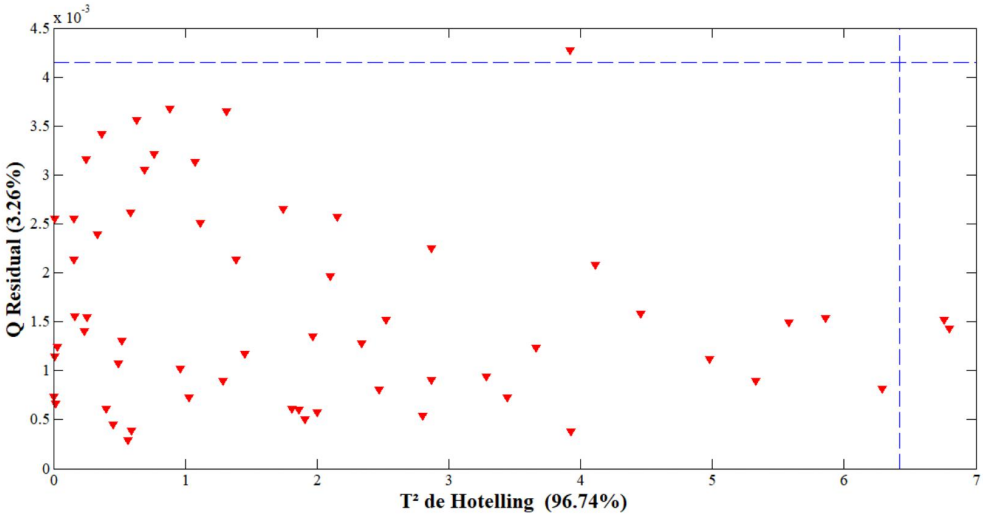


Figura 28 - Limites estatísticos T² e Q residual estabelecidos para as amostras presentes no conjunto de treinamento a 95% de confiança após a detecção de *outliers*.

A Figura 29 mostra a classificação das amostras para o modelo SIMCA, em que a classificação das amostras está dividida no conjunto de treinamento e teste. Observa-se que para o conjunto de treinamento todas as amostras foram classificadas dentro da classe de interesse. Já no conjunto teste quatro amostras pertencentes a classe-alvo foram classificadas erroneamente fora da classe (falso negativo) e por consequência houve uma diminuição na sensibilidade do modelo. Em relação as amostras de outras regiões, pertencentes à classe não-alvo, apenas três amostras foram erroneamente classificadas como sendo pertence a classe de

interesse (falso positivo), mostrando assim que o modelo possui especificidade. O bom desempenho destes parâmetros fornece um modelo com 89% de eficiência.

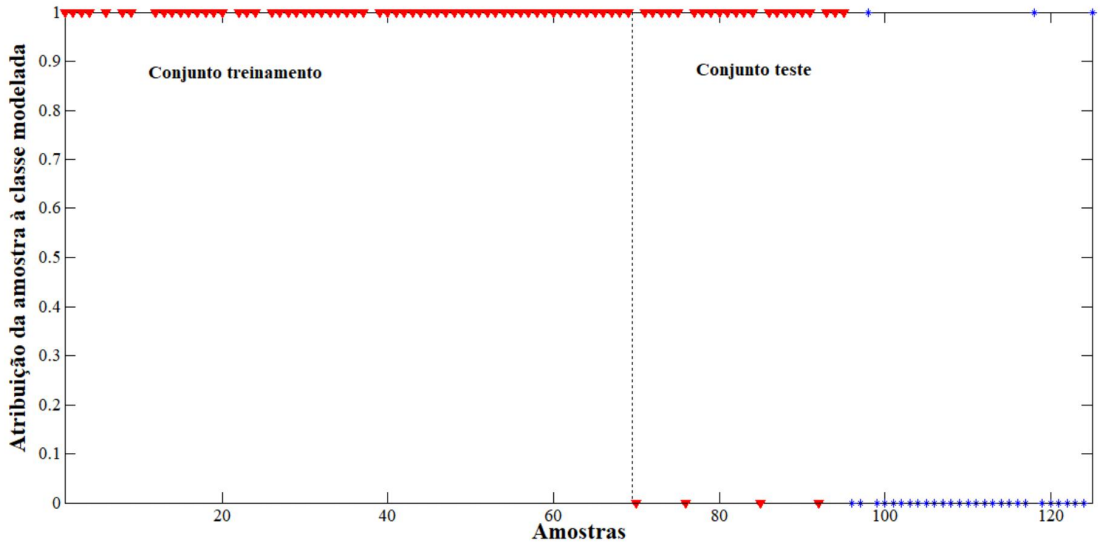


Figura 29 - Modelo SIMCA para os espectros na região do UV-Vis, sendo (▼) grãos arábica do Cerrado Mineiro e (∗) grãos arábica de outras regiões.

Como mencionado anteriormente, uma das vantagens da construção de modelos utilizando o SIMCA é o fato de ser possível avaliar o poder de influência das variáveis no desempenho do modelo. A partir das equações 9-13 foi calculado o poder de modelagem, ψ, para as 220 variáveis contidas neste modelo, isto é, a influência da variável no estabelecimento dos limites definidos para a classe de interesse. Os resultados das 15 variáveis mais importantes para o modelo são mostrados na Figura 30.

Considerando que quanto mais próximo de 1 for ψ, maior será a contribuição da variável para o modelo, conclui-se observando a Figura 30 que as regiões entre 271-274 nm e 320-326 nm, atribuídas anteriormente, foram de fato importantes na discriminação das amostras de café, influenciando assim no desempenho final do modelo. Além disso, outra região também contribuiu para o modelo, a faixa de 236 a 240 nm, as substâncias que podem ser encontradas nessa região são ácido cítrico, ácido caféico, ácido fenil acético, cafeína e ácido quínico (SOUTO, 2017), sendo estes dois últimos também identificados na análise por PS-MS descrita anteriormente.

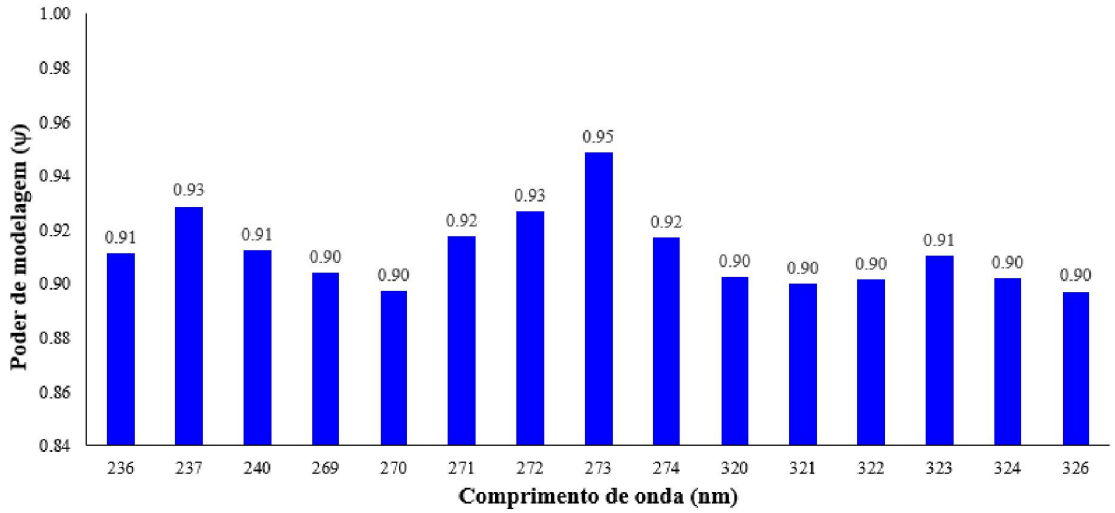


Figura 30 - Contribuição das variáveis contidas no modelo (poder de modelagem).

A partir do gráfico de *loadings* da PCA e do poder de modelagem, é possível notar que as substâncias que tiveram maior contribuição para discriminar os grãos de café da região do Cerrado Mineiro em relação aos grãos de café das regiões do Caparaó, Mogiana e Sul de Minas foram trigonelina e ácidos clorogênicos. Sendo que os ácidos clorogênicos foram responsáveis por agrupar as amostras do Cerrado e a trigonelina as amostras das outras regiões, como mostrado na Figura 26 e 27.

Conforme apresentado na Tabela 12, os modelos construídos com os métodos DD-SIMCA e OCPLS apresentaram melhor desempenho que o SIMCA quando aplicado o método de seleção de variáveis. Desse modo, os modelos DD-SIMCA e OCPLS foram construídos para os dados com as 34 variáveis selecionadas pelo algoritmo OPS.

Modelo DD-SIMCA

Semelhante ao SIMCA, para a construção do modelo DD-SIMCA foram escolhidas três PCs para modelar as amostras do Cerrado em que a área de aceitação foi baseada na distribuição qui-quadrado e o erro tipo I (α) definido em α=0,05. O nível de significância para a detecção de *outliers* foi definido como α=0,01. Não foram identificadas amostras anômalas no conjunto de dados. A Figura 31 mostra o gráfico da região de aceitação para o conjunto de treinamento (70 amostras) e para o conjunto teste (60 amostras), sendo que a linha verde estabelece a fronteira/limite das amostras pertencentes à classe-alvo e classe não-alvo e o pontilhado preto o limite da área de aceitação.

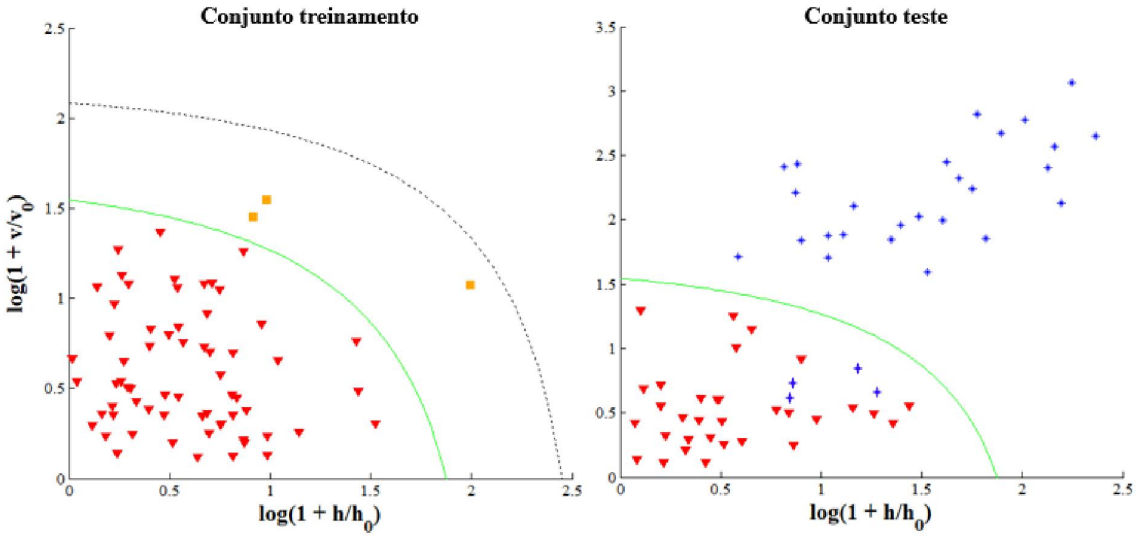


Figura 31 – Gráficos de aceitação obtidos para o modelo DD-SIMCA para os conjuntos de treinamento e teste, sendo h_i e v_i, nessa ordem, os valores da distância dos escores e da distância ortogonal para a amostra $i = 1, ..., n$, sendo (▼) grãos da espécie arábica do Cerrado Mineiro e (∗) grãos da espécie arábica de outras regiões.

Nota-se que houve 100% de sensibilidade para o conjunto de treinamento, uma vez que amostras em laranja ficaram dentro da região limite de aceitação, e para o conjunto teste houve 100% de sensibilidade e 87% especificidade já que houve quatro falsos positivos. A eficiência de modelo final foi de 93%.

Modelo OCPLS

No modelo OCPLS, o número de variáveis latentes foi determinado pela validação cruzada aplicada no conjunto de treinamento em que foi fornecido um vetor contendo os desvios padrão no modelo residual com diferentes VLs. O número ideal de VLs para este modelo foi baseada no menor desvio padrão dos resíduos absolutos centralizados do modelo (ACR), que neste caso, foi igual a duas. Os limites estabelecidos para a construção do modelo foram a um nível de confiança de 95% (α=0,05) e a detecção de *outliers* foi baseada nos altos valores da distância de scores (SD) e ACR. Não foi identificada nenhuma amostra como sendo *outlier*.

A Figura 32 mostra os limites (em preto) para a classe-alvo no conjunto de treinamento, (amostras fora do 3° quadrante são classificadas como amostras fora da classe-alvo), foram obtidos quatro falsos negativos e sensibilidade igual 94%. Para o conjunto teste, a sensibilidade foi igual a 100% e a especificidade igual 87%, já que houve 4 falsos positivos. A eficiência do modelo foi igual a 93%.

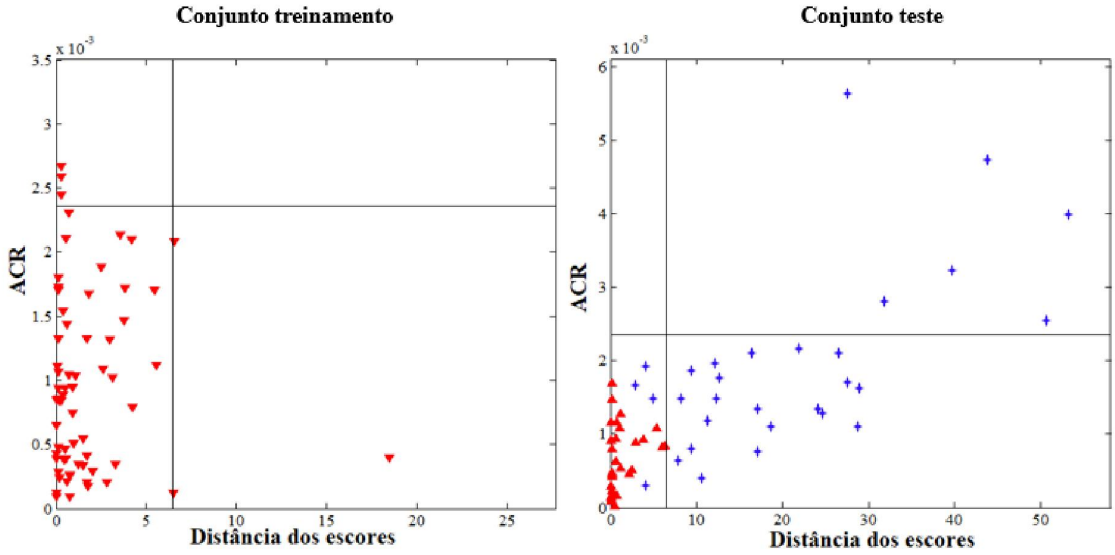


Figura 32 - Gráficos obtidos pelo modelo OCPLS da distância dos escores (SD) *vs.* resíduos absolutos centralizados (ACR) para os conjuntos de treinamento e teste, sendo (▼) grãos da espécie arábica do Cerrado Mineiro e (∗) grãos da espécie arábica de outras regiões.

Os resultados encontrados para os modelos DD-SIMCA e OCPLS mostram a importância do método de seleção de variáveis, uma vez que foi possível melhorar o desempenho dos modelos consideravelmente, corroborando com outros estudos encontrados na literatura que mostram a importância do método de seleção de variáveis na construção dos modelos quimiométricos (MIAW *et al*., 2018; RIBEIRO *et al*., 2021; GOMES *et al*., 2022).

5.2.5 Considerações parciais

Com as técnicas espectroscópicas PS-MS e UV-Vis utilizadas neste trabalho foi possível identificar as principais substâncias orgânicas presentes no café como trigonelina, ácidos clorogênicos, ácido quínico e sacarose, sendo que a maioria desses compostos influenciam diretamente na qualidade sensorial do produto.

Utilizando a técnica de TXRF foi obtido informações a respeito da composição elementar dos analitos inorgânicos presentes nos grãos, sendo que potássio, cálcio, fósforo, enxofre e cloro foram encontrados em maiores teores.

Em relação a discriminação das amostras do Cerrado e não Cerrado foi possível obter modelos com alta performance a partir dos dados obtidos pela técnica de UV-Vis em conjunto com os métodos de modelagem de classe SIMCA, DD-SIMCA e OCPLS, sendo obtido eficiência superior a 90%.

Por fim, ao calcular o poder de modelagem das variáveis para os dados de UV-Vis obteve-se informações a respeito das variáveis que mais contribuíram para discriminação das amostras entre as duas classes. Além disso, a aplicação do método de seleção de variáveis se mostrou uma estratégia válida para obter informações a respeito das variáveis com maior capacidade preditiva e melhorar o desempenho dos modelos construídos.

5.3 Fusão de dados

A partir dos resultados apresentados anteriormente, foi possível observar que somente os modelos construídos para os dados obtidos por espectroscopia na região do UV-Vis apresentaram desempenho adequado na discriminação das amostras de café do Cerrado Mineiro. Em relação as outras técnicas, algumas apresentaram alta sensibilidade e baixa especificidade (PS-MS e TXRF) ou baixa sensibilidade e baixa especificidade (FTIR) para o conjunto teste do modelo. Nesse contexto, visando avaliar a sinergia das informações obtidas por diferentes técnicas foram construídos modelos de classificação utilizando a fusão de dados em nível baixo. A escolha do nível baixo foi feita baseando-se no fato que nesse nível não há perda de informações relacionadas às variáveis importantes e que influenciam na performance do modelo (BORRÀS et al., 2015). O objetivo do uso da fusão de dados não é para obter apenas melhores modelos em relação às técnicas individuais, mas contribuir com a interpretabilidade dos modelos construídos, buscando correlação entre as técnicas, principalmente correlações atômico-moleculares. Como os resultados para os dados da espectroscopia na região do UV-Vis foram satisfatórios, esta técnica não foi utilizada na fusão de dados com as outras.

Os modelos quimiométricos foram construídos de uma forma integrada com os dados obtidos por FTIR, PS-MS e TXRF. Foram concatenados os seguintes blocos de dados:1) PS-MS (+) e PS-MS (-); 2) FTIR e PS-MS; 3) FTIR e TXRF; 4) PS-MS e TXRF; 5) FTIR, PS-MS e TXRF. Os dados foram concatenados já devidamente pré-processados, e em seguida foram autoescalados. Os modelos foram construídos a partir dos dados originais com todas as variáveis e depois para os dados apenas com as variáveis selecionadas pelo método OPS, sendo estas selecionadas após o autoescalamento dos dados concatenados. Nesta etapa, o método SIMCA foi o único capaz de fornecer modelos satisfatórios para as matrizes concatenadas, sendo estes resultados apresentados e discutidos nas seções 5.3.2 a 5.3.4.

Estudos anteriores mostram que modelos construídos pelo método SIMCA apresentaram bom desempenho em discriminar produtos com denominação de origem, como é

o caso do trabalho desenvolvido por Ríos-Reina e colaboradores (2019), onde obtiveram 100% de classificação correta para os conjuntos de validação de vinagres de vinhos a partir de dados obtidos pela técnica de UV-Vis. Já modelos construídos com o método OCPLS para identificação de adulterações em mel (SOUZA, 2021) e discriminação de vinhos com origens geográficas (GOMES, 2021) apresentaram baixa especificidade.

5.3.1 PS-MS (+) e PS-MS (-)

A construção dos modelos com os dados de PS-MS (+) e PS-MS (-) foi realizada com intuito de avaliar se a performance do modelo seria semelhante as obtidas nos dados individuais. A matriz final de dados foi de 130x1802. Os resultados obtidos são mostrados na Tabela 13.

Tabela 13 - Figuras de mérito calculadas para o modelo de fusão de dados PS-MS (+) e PS-MS (-), sendo (a) dados com as 1802 variáveis e (b) com as 294 variáveis selecionadas pelo OPS.

a) Dados originais

		Treinamento	*Teste*		
Métodos	**PCs/ VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	5	0,98	1,00	0,33	0,58
DD-SIMCA	5	0,99	1,00	0,30	0,55
OCPLS	5	0,97	1,00	0,07	0,26
b) Seleção de variáveis					
SIMCA	6	1,00	0,97	0,43	0,65
DD-SIMCA	6	0,99	1,00	0,23	0,48
OCPLS	6	0,99	1,00	0,13	0,37

Comparando os resultados aos obtidos nas matrizes individuais nota-se que a fusão de dados não influenciou na performance dos modelos, uma vez que os resultados obtidos para a sensibilidade no conjunto de treinamento e eficiência no conjunto teste foram semelhantes para todos os modelos construídos com os dados originais e com a seleção de variáveis. Os modelos construídos apresentaram especificidade menor que 50%.

Dessa maneira, para os demais modelos construídos a partir da fusão de dados, com os dados obtidos por PS-MS optou-se por utilizar a matriz de dados PS-MS (+) e PS-MS (-) concatenadas, denominada como PS-MS, de modo a otimizar o número de modelos a serem construídos.

5.3.2 FTIR e PS-MS

Para o desenvolvimento dos modelos de fusão de dados com FTIR e PS-MS foi realizado pré-processamento SNV e centragem na média para os dados FTIR e apenas centragem na média para os dados de PS-MS. Após as matrizes serem concatenadas foi realizado o autoescalamento dos dados. As etapas para a construção dos modelos foram realizadas de acordo com as etapas descritas anteriormente na seção 3.4. Na etapa de identificação de *outliers*, quatorze amostras foram retiradas do modelo por apresentarem altos valores de T^2 e Q residual.

Na Tabela 14 estão apresentados os valores das figuras de mérito para o modelo SIMCA a partir dos dados originais e após seleção de variáveis utilizando o *OPS*.

Tabela 14 - Figuras de mérito calculadas para o modelo de fusão de dados FTIR e PS-MS, sendo (a) dados com as 2902 variáveis e (b) com as 225 variáveis selecionadas pelo OPS.

		Treinamento	***Teste***		
Dados	**PCs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
Originais	8	0,87	1,00	0,80	0,89
Seleção de variáveis	6	0,85	1,00	0,83	0,91

Para ambos os conjuntos de dados, originais e com variáveis selecionadas, foram obtidos modelos com performance superior aos modelos obtidos com as matrizes individuais. A matriz original com os dados concatenados foi de 130x2902 enquanto a matriz com as variáveis selecionadas foi 130x225. Das 225 variáveis selecionadas, 15 foram dos espectros FTIR e as demais dos espectros PS-MS. As regiões selecionadas no espectro FTIR foram, 1497-1500 cm^{-1}, 1684-1694 cm^{-1} e 1703-1708 cm^{-1} atribuídas a ésteres e ácidos. Em relação as variáveis mais importantes selecionadas para os espectros PS-MS destacam-se os íons m/z 440 (-), m/z 530 (-), m/z 590 (-), m/z 891 (+) e m/z 910 (+).

O método de seleção de variáveis não melhorou o desempenho do modelo quando se avalia a capacidade preditiva nos conjuntos de treinamento e teste, no entanto, aplicando a seleção de variáveis a matriz foi reduzida de 130x2902 a 130x225, o que facilita a interpretação dos dados. Além disso, nos dados originais são necessários 8 PCs para a construção do modelo, já para as variáveis selecionadas a quantidade foi reduzida a 6 PCs. A Figura 33 mostra a classificação das amostras para o conjunto treinamento e teste. No conjunto de treinamento foi obtido 85% de sensibilidade, com a ocorrência de 8 falsos negativos. Para o conjunto teste foi

obtido 100% de sensibilidade e 83% de especificidade. O desempenho final do modelo foi igual a 91%.

Pode-se concluir que para o conjunto de espectros FTIR e PS-MS, a fusão de dados foi uma estratégia válida em que foi possível melhorar o desempenho do modelo SIMCA ao considerar o sinergismo das duas técnicas, uma vez que os resultados obtidos superaram os dados obtidos anteriormente para as matrizes individuais.

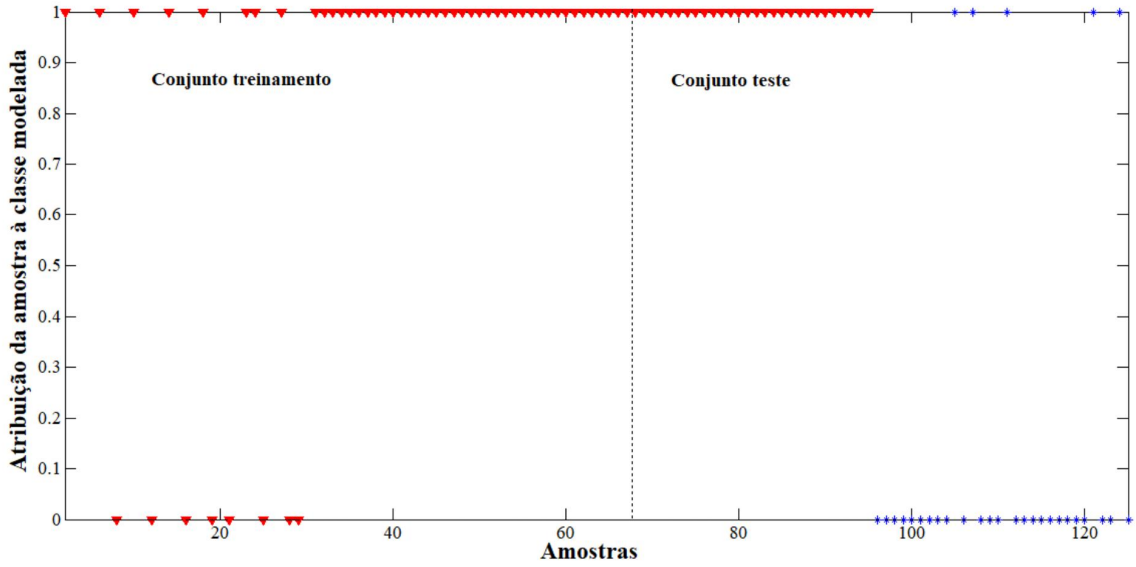


Figura 33 - Modelo SIMCA construído a 95% de confiança para os dados FTIR e PS-MS para as 225 variáveis selecionadas, sendo (▼) as amostras do Cerrado Mineiro e (*) amostras de outras regiões.

5.3.3 PS-MS e TXRF

O terceiro modelo com fusão de dados foi construído com os dados obtidos por PS-MS e TXRF. Para a construção da matriz, os espectros de massas foram centrados na média e os dados de TXRF autoescalados, em seguida foi realizado a fusão das matrizes e aplicado o autoescalamento na matriz concatenada. A matriz final foi de 130x1817. Foram identificadas 14 amostras com altos valores de T^2 e Q residual e estas foram retiradas do conjunto de dados.

A construção dos modelos de classificação foi realizada para os dados concatenados e após a seleção de variáveis executada pelo OPS. Na Tabela 15 estão apresentadas as figuras de mérito obtidos para ambos os modelos. Observa-se que para o mesmo número de PCs, o modelo com seleção de variáveis apresentou um desempenho ligeiramente melhor. O algoritmo OPS selecionou 180 variáveis como as mais importantes, sendo 9 variáveis de TXRF e 171 variáveis de PS-MS. Os elementos selecionados como mais relevantes para o modelo foram P, S, Cl, Ti,

Fe, Cu, Zn Br, Rb. De acordo com a literatura, esses elementos caracterizam grande parte da composição do solo do Cerrado Mineiro (LIMA *et al.*, 2019). As variáveis mais importantes para os dados de PS-MS foram iguais as obtidas anteriormente.

Tabela 15 - Figuras de mérito calculadas para o modelo de fusão de dados PS-MS e TXRF, sendo (a) dados com as 1817 variáveis e (b) com as 180 variáveis selecionadas pelo OPS.

		Treinamento	***Teste***		
Dados	**PCs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
Originais	6	0,85	0,97	0,63	0,78
Seleção de variáveis	6	0,89	1,00	0,70	0,84

A Figura 34 mostra a classificação das amostras para o modelo SIMCA construídos com as 180 variáveis. Percebe-se que houve algumas amostras no conjunto de treinamento que foram falsos negativos o que interferiu no desempenho da sensibilidade do modelo. Já para o conjunto teste foi obtido 100% de sensibilidade na classificação das amostras alvo e 70% de especificidade, uma vez que o modelo classificou erroneamente algumas amostras fora da classe de interesse como sendo amostras alvo. Mais uma vez a estratégia de fusão de dados mostrou a sinergia entre as técnicas, apresentando modelos com melhores desempenho que as técnicas de forma individual.

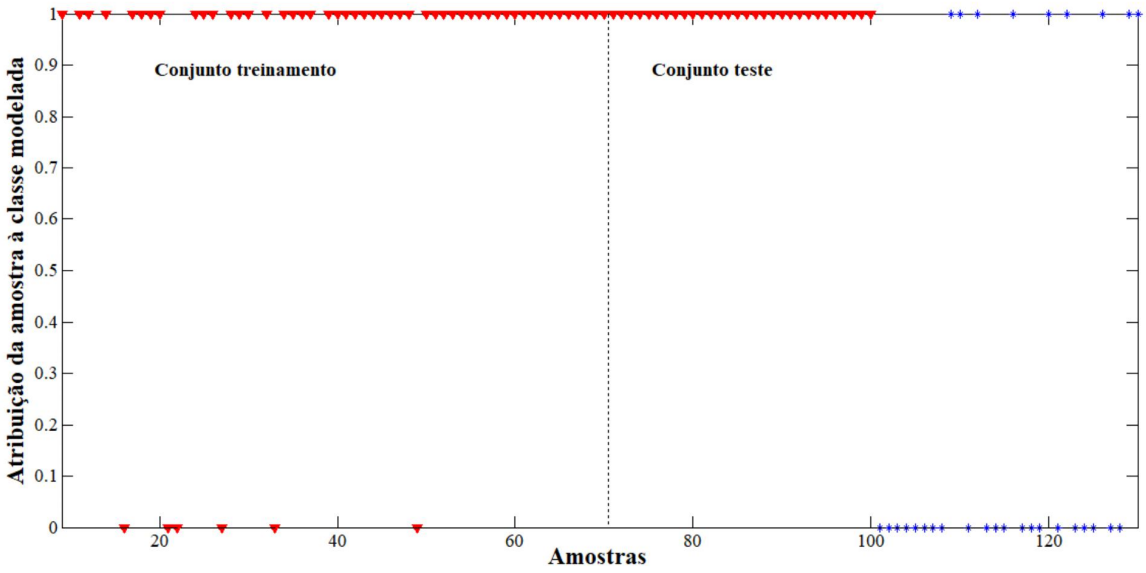


Figura 34 - Modelo SIMCA construído a 95% de confiança para os dados PS-MS e TXRF para as 180 variáveis selecionadas, sendo (▼) as amostras do Cerrado Mineiro e (∗) amostras de outras regiões.

5.3.4 FTIR, PS-MS e TXRF

Para a construção da matriz de fusão de dados para os dados FTIR, PS-MS e TXRF foram realizados os pré-processamentos adequados para cada matriz individual já descritos anteriormente e em seguida o autoescalamento dos dados. Semelhante as outras matrizes de fusão, foram retiradas quatorze amostras consideradas *outliers*.

O principal interesse em desenvolver modelos para os dados com estas três técnicas em conjunto é o fato delas possuírem informações aditivas em relação aos compostos orgânicos (FTIR e PS-MS) e inorgânicos (TXRF) presentes nos grãos de café.

Na Tabela 16 estão apresentados os resultados obtidos para as figuras de mérito dos modelos SIMCA para os dados concatenados e após aplicação do método de seleção de variáveis. Nota-se que a seleção de variáveis melhorou principalmente a especificidade do modelo, aumentando de 63% para 80% e por consequência melhorando a eficiência final do modelo. Em relação as variáveis selecionadas, foram escolhidas 6 variáveis para os dados de TXRF (P, Cl, Ti, Cu, Zn e Rb) e 204 variáveis para os dados de PS-MS, sendo que os principais íons escolhidos foram mencionados anteriormente na seção 5.3.2. O OPS não selecionou nenhuma variável da técnica de FTIR como sendo importante para o modelo.

Tabela 16 - Figuras de mérito calculadas para o modelo de fusão de dados FTIR, PS-MS e TXRF sendo (a) dados com as 2917 variáveis e (b) com as 210 variáveis selecionadas pelo OPS.

		Treinamento	***Teste***		
Dados	**PCs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
Originais	7	0,80	1,00	0,63	0,80
Seleção de variáveis	6	0,83	1,00	0,80	0,89

Na Figura 35 é possível visualizar a classificação das amostras para o modelo SIMCA construídos com as 210 variáveis, observa-se que houve 11 falsos negativos no conjunto de treinamento e 6 falsos positivos no conjunto teste. A eficiência do modelo foi de 89%, o que pode ser considerado satisfatório para modelos de classificação.

É importante ressaltar que as três técnicas individualmente não foram capazes de fornecer modelos robustos, entretanto, ao realizar a fusão de dados isso foi possível, mostrando assim que a utilização da estratégia de fusão de dados é uma alternativa válida e deve ser considerada, principalmente quando se tem dados oriundos de diferentes técnicas ou natureza, sendo possível aproveitar as informações complementares de cada uma.

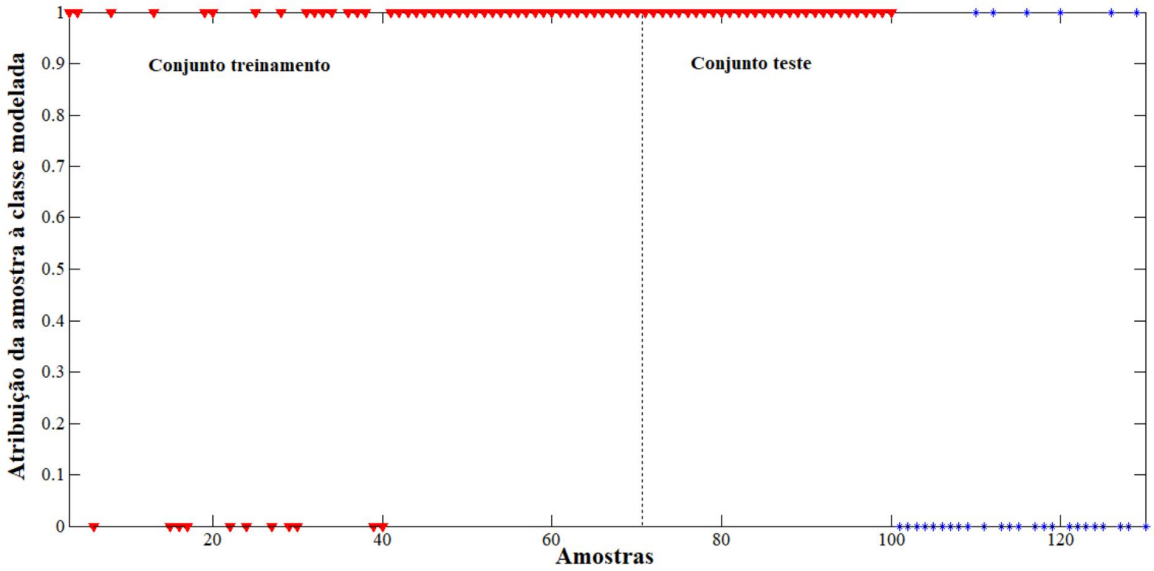


Figura 35 - Modelo SIMCA construído a 95% de confiança para os dados FTIR, PS-MS e TXRF para as 210 variáveis selecionadas, sendo (▼) as amostras do Cerrado Mineiro e (*) amostras de outras regiões.

A abordagem de fusão de dados permitiu obter informações atômico-moleculares dos grãos de café, variáveis como o zinco (Zn) e os ácidos clorogênicos foram selecionadas como contribuintes na discriminação dos cafés do Cerrado. Na literatura é relatado que solos supridos com quantidades ideais de zinco influenciavam nos altos teores de ácidos clorogênicos (MARTINEZ *et al.*, 2014). Essa informação corrobora com a interpretação obtida neste trabalho.

Como mencionado anteriormente, os açúcares são compostos importantes na qualidade da bebida de café e seu metabolismo varia de acordo com a disponibilidade de potássio (K). De acordo com Martinez e colaboradores (2014), o papel do potássio no metabolismo dos carboidratos é essencial e influencia diretamente nos teores de açucares presentes nos grãos de café. Neste trabalho, foi possível observar que o potássio foi uma variável importante na caracterização dos cafés do Cerrado, sendo encontrado em alta concentração nos grãos de café verde nas análises por TXRF. Além disso, na análise realizada por PS-MS foi possível observar que um dos compostos obtidos em maior abundância foi a sacarose (m/z 381). Tais resultados vão de encontro ao obtido por Martinez e colaboradores anteriormente.

5.3.5 Demais modelos

Os demais modelos construídos para DD-SIMCA e OCPLS com a fusão de dados não superaram os resultados obtidos nos modelos individuais, nem quando aplicado método de seleção de variáveis, e, portanto, não foram apresentados na discussão do trabalho. No entanto, as figuras mérito obtidas para ambos e para fusão FTIR e TXRF estão apresentadas no anexo.

5.3.6 Considerações parciais

A estratégia de fusão de dados permitiu explorar as informações aditivas dos dados obtidos pelas técnicas individuais, mostrando-se uma abordagem válida que melhorou o desempenho do modelo e a interpretabilidade dos resultados. Além disso, foi possível obter informações atômico-moleculares das amostras, possibilitando caracterizar os grãos de café em relação ao perfil orgânico e inorgânicos.

Neste trabalho, os modelos de fusão de dados construídos forneceram melhores resultados quando comparados aos modelos individuais para cada técnica. Ao utilizar o método de seleção de variáveis, foi possível melhorar ainda mais o desempenho dos modelos. Desse modo, pode-se concluir que ambas as estratégias são importantes e devem ser consideradas para construção de modelos.

6. CONSIDERAÇÕES FINAIS

A partir do planejamento de experimentos foi possível otimizar o melhor solvente e as condições ideais para extração dos compostos presentes no café, de modo a minimizar tempo, o uso de solvente e geração de resíduos a partir do planejamento de misturas e do planejamento composto central. Os extratos obtidos, após definidas as condições ideais, foram utilizados nas análises por espectrometria de massas e espectroscopia de absorção na região do UV-Vis. Os resultados mostram que o planejamento de experimentos é uma etapa bastante importante e deve ser utilizado de modo a garantir a qualidade no desenvolvimento do trabalho.

As informações obtidas pelas técnicas instrumentais utilizadas neste trabalho possibilitaram obter informações a respeito da composição química dos grãos de café, assim como caracterizar os grãos de cafés do Cerrado Mineiro, e assim por serem técnicas de análise rápidas, de baixo custo, com mínimo preparo de amostra e sem necessidade do uso de solventes tóxicos, podem ser utilizadas em conjunto com a análise quimiométrica para prever e autenticar amostras-alvo.

Em especial, os modelos construídos com os métodos de modelagem de classe única SIMCA, DD-SIMCA e OCPLS a partir dos dados obtidos por espectroscopia na região do UV-Vis forneceram resultados satisfatórios em que foi possível identificar que os ácidos clorogênicos e a trigonelina foram responsáveis por discriminar os grãos de café da região do Cerrado Mineiro das regiões do Caparaó, Mogiana e Sul de Minas. Além disso, não são encontrados muitos trabalhos na literatura utilizando a técnica de UV-Vis para autenticação/classificação de amostras com certificação de origem, logo, os resultados obtidos neste trabalho mostram que esta técnica deve ser avaliada como alternativa para estudos nesta área.

Em relação aos compostos inorgânicos, foi possível identificar elementos como Ti, Fe, Cu e Zn que são encontrados em maiores quantidades no solo do Cerrado e podem ser considerados como característicos dos grãos dessa região. Além disso, pode ser destacado que a importância dos ácidos clorogênicos, substâncias presentes nos grãos de café do Cerrado, para os modelos de classificação podem estar relacionados com quantidades de zinco presentes no solo.

Com a utilização do método de seleção de variáveis OPS foi possível obter informações a respeito das variáveis mais importantes para os modelos, o que por sua vez facilita a interpretação em relação a informação química presente nas amostras. Dessa maneira, são

métodos que valem a pena serem explorados e testados nas matrizes de dados, além de facilitar a interpretação dos dados, ainda possibilita melhorar a performance do modelo já que elimina as informações irrelevantes contidas nos dados. Modelos construídos com o método SIMCA forneceram melhores resultados quando comparados com os outros métodos de modelagem de classe.

A estratégia de fusão de dados possibilitou explorar as informações complementares e adicionais presentes nos dados individuais e forneceu modelos com alto desempenho, diferentes dos modelos individuais, mostrando assim que é uma abordagem interessante e que o sinergismo dos dados influencia e contribui para a qualidade dos modelos construídos.

Por fim, com os resultados obtidos neste trabalho foi possível caracterizar os grãos de café da região do Cerrado Mineiro e diferenciá-los das demais regiões, mostrando assim a importância da análise quimiométrica para o desenvolvimento de experimentos e para construção de modelos com dados multivariados visando autenticação de alimentos ou outros produtos.

7. REFERÊNCIAS

ABOULWAFA, M. M. *et al*. Authentication and discrimination of green tea samples using UV–vis, FTIR and HPLC techniques coupled with chemometrics analysis, *Journal of* **Pharmaceutical and Biomedical Analysis**, v. 164, p. 653–658, 2019.

ABREU, F. G. et al. Raman spectroscopy: A new strategy for monitoring the quality of green coffee beans during storage, **Food Chemistry**, v. 283, p. 241-248, 2019.

ALLEGRETTA, I. *et al*. A fast method for the chemical analysis of clays by total-reflection x-ray fluorescence spectroscopy (TXRF), **Applied Clay Science**, v. 180, 2019.

ALMEIDA, L. F., TARABAL, J. Cerrado Mineiro Region Denomination of Origin Mark: Internationalization Strategy, **Proceedings in Food System Dynamics**, p. 133–144, 2019.

ANTOSZ, F. J. *et al*. The use of total reflectance X-ray fluorescence (TXRF) for the determination of metals in the pharmaceutical industry, **Journal of Pharmaceutical and Biomedical Analysis**, v. 62, p. 17–22, 2012.

ASSIS, C., **Aplicação de técnicas espectroscópicas, métodos quimiométricos, fusão de dados e seleção de variáveis no controle de qualidade de blends das espécies de café arabica e robusta**, Tese de doutorado, UFMG, Belo Horizonte, 2018.

ASSIS, C. *et al*. A data fusion model merging information from near infrared spectroscopy and X-ray fluorescence. Searching for atomic-molecular correlations to predict and characterize the composition of coffee blends, **Food Chemistry**, v. 325, p. 126953, 2020.

AWAD, H. *et al*. Mass spectrometry, review of the basics: Ionization, **Applied Spectroscopy Reviews**, v. 50, n. 2, p. 158–175, 2015.

AZEVEDO, A. L. **De biomateriais a supramoléculas teranósticas: a inovação terapêutica destinada a proteger, transportar e entregar moléculas bioativas de forma planejada**, Dissertação de mestrado, Recife, PE, 2015.

Barbosa, L. C. **A. Espectroscopia no Infravermelho na caracterização de compostos orgânicos**. Viçosa: Editora UFV, 2013.

BARBOSA, S. O. L. *et al*. A participação de Minas Gerais e do Brasil na cadeia produtiva global do café, **Economia & Região**, v. 9, n. 1, p. 147, 2020.

BARBOZA, M. F. *et al.* Desenvolvimento e validação de um método analítico simples e rápido por espectroscopia uv para quantificação de aciclovir em matrizes hidrofílicas de liberação prolongada, **Química Nova**, v.33, n.3, p, 747-749, 2010.

BATISTA, L. A. **A indicação geográfica como indutora da organização dos pequenos produtores: O caso "Café das montanhas do Sul de Minas Gerais".** Universidade Federal Fluminense, p. 114, 2012.

BARROS NETO, B. *et al.* **Como fazer experimentos: aplicações na ciência e na indústria**. 4° edição, São Paulo: Bookman, 2010.

BECKER, T. European Food Quality Policy: The Importance of Geographical Indications, Organic Certification and Food Quality Assurance Schemes in European Countries and Trade Policy, **The Estey Centre Journal of International Law and Trade Policy**, v.10, n.1, p. 111-130, 2009.

BELCHIOR, V. *et al.* Attenuated Total Reflectance Fourier Transform Spectroscopy (ATR-FTIR) and chemometrics for discrimination of espresso coffees with different sensory characteristics**, Food Chemistry**, v. 273, p. 178–185, 1 fev. 2019.

BIANCOLILLO, A. *et al*. Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication, **Analytica Chimica Acta**, v. 820, p. 23–31, 2014.

BORRÀS, E. *et al.* Data fusion methodologies for food and beverage authentication and quality assessment - A review, **Analytica Chimica Acta**, v. 891, p. 1–14, 2015.

BORÉM, M. F. The relationship between organic acids, sucrose and the quality of specialty coffees, **African Journal of Agricultural Research**, p. 709-717, 2016.

BORGES, A. Y. **Aplicação de técnicas multivariadas em espectros de infravermelho para a determinação de teores totais de carbono, oxigênio e hidrogênio em amostras de biomassas e biocarvões**, Dissertação de mestrado, Goiânia, 2015.

BRAGA, L. M. **Comparativo de métodos sensoriais descritivos na avaliação de café torrado e moído**, Dissertação de mestrado, Campo Mourão, PR, 2019.

BRUKER. **Picofox: User Manual**, p.118, 2012.

BRUNS, E. R., FAIGLE, G. F. J. Quimiometria*,* **Química Nova**, 1985.

CASALE, M. *et al*. Characterisation of PDO olive oil Chianti Classico by non-selective (UV-visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques, **Analytica Chimica Acta**, v. 712, p. 56–63, 2012.

CAVDAROGLU, C., OZEN, B. Detection of vinegar adulteration with spirit vinegar and acetic acid using UV–visible and Fourier transform infrared spectroscopy, **Food Chemistry**, v. 379, p. 132150, 2022.

CERRADO MINEIRO, Federação dos Cafeicultores do Cerrado. Disponível em: https://www.cerradomineiro.org/. Acesso em: dezembro de 2021.

COCCHI, M., Data Handling in Science and Technology – **Data fusion methodology and applications**, Editora Elsevier, 2019.

CRAIG, P. A. *et al*. Evaluation of the potential of FTIR and chemometrics for separation between defective and non-defective coffees, **Food Chemistry**, v. 132, n. 3, p. 1368–1374, 2012.

CUSTERS, D. *et al.* ATR-FTIR spectroscopy and chemometrics: An interesting tool to discriminate and characterize counterfeit medicines, **Journal of Pharmaceutical and Biomedical Analysis**, v. 112, p. 181–189, 10 ago. 2015.

DE LA CALLE, I. *et al*. Sample pretreatment strategies for total reflection X-ray fluorescence analysis: A tutorial review, **Spectrochimica Acta - Part B Atomic Spectroscopy**, v. 90, p. 23–54, 2013.

EMBRAPA, Empresa Brasileira de Pesquisa Agropecuária. Disponível em: https://www.embrapa.br/cafe. Acesso em: dezembro de 2021.

EL-ABASSY, R. M. *et al*. Discrimination between Arabica and Robusta green coffee using visible micro-Raman spectroscopy and chemometric analysis, **Food Chemistry**, 126(3), p.1443–1448, 2011.

EL-ANEED, A. *et al.* Mass spectrometry, review of the basics: Electrospray, MALDI, and commonly used mass analyzers, **Applied Spectroscopy Reviews**, v. 44, n. 3, p. 210–230, 2009.

ELLIS, I. D. *et al*. Fingerprinting food: Current technologies for the detection of food adulteration and contamination, **Chemical Society Reviews**, v. 41, n. 17, p. 5706–5727, 2012.

FERREIRA, A. R. *et al.* Kennard-Stone method outperforms the Random Sampling in the selection of calibration samples in SNPs and NIR data, **Ciência Rural**, v. 52, n. 5, 2022.

FERREIRA, M. M. C. **Quimiometria - Conceitos, Métodos e Aplicações**. *In*: UNICAMP (Ed.), Campinas, 2015.

FORINA, M. *et al.* Artificial nose, NIR and UV-visible spectroscopy for the characterisation of the PDO Chianti Classico olive oil, **Talanta**, v. 144, p. 1070–1078, 2015.

ICO, International Coffee Organization. Disponível em: https://www.ico.org/. Acesso em: dezembro de 2021.

GALO, L. A., COLOMBO, F. M. Espectrofotometria de longo caminho óptico em espectrofotômetro de duplo-feixe convencional: uma alternativa simples para investigações de amostras com densidade óptica muito baixa, **Química Nova**, v. 32, n.2, p. 488-492, 2009.

GARRETT, R *et al*. Arabica and Robusta coffees: Identification of major polar compounds and quantification of blends by direct-infusion electrospray ionization-mass spectrometry, **Journal of Agricultural and Food Chemistry**, v. 60, n. 17, p. 4253–4258, 2012.

GARRETT, R. *et al*. Coffee origin discrimination by paper spray mass spectrometry and direct coffee spray analysis, **Analytical Methods**, v. 5, n. 21, p. 5944–5948, 2013.

GARRETT, R. *et al.* Discrimination of arabica coffee cultivars by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and chemometrics, **LWT - Food Science and Technology**, v. 50, n. 2, p. 496–502, 2013.

GOMES, A. *et al*. Slovak Tokaj wines classification with respect to geographical origin by means of one class approaches, **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy,** v. 257, p. 119770, 2021.

GOMES, A. *et al*. Variable selection in the chemometric treatment of food data: A tutorial review, **Food Chemistry**, v. 370, 2022.

GONÇALVES, B. D. M., **Produção e consumo de café: Uma análise do custo de oportunidade de produção de cafés especiais e convencionais**, Dissertação de Mestrado, São Paulo, 2018.

GRDADOLNIK, J. ATR-FTIR Spectroscopy: Its advantages and limitations, **Acta Chim. Slov.**, v. 49, p. 631-642, 2002.

GUO, T. *et al*. Non-target geographic region discrimination of Cabernet Sauvignon wine by direct analysis in real time mass spectrometry with chemometrics methods, **International Journal of Mass Spectrometry**, v. 464, p. 116577, 2021.

HOFFMANN, E., STROOBANT, V. **Mass Spectrometry: Principles and Applications**, Editora Wiley, 3° edição, 2007.

HOPKE, K. P. The evolution of chemometrics, **Analytica Chimica Acta**, v. 500, n. 1–2, p. 365–377, 2003.

HORNTRICH, C. *et al*. Considerations on the ideal sample shape for Total Reflection X-ray Fluorescence Analysis, **Spectrochimica Acta - Part B Atomic Spectroscopy**, v. 66, n. 11–12, p. 815–821, 2011.

HU, B., YAO, P. Z. Electrospray ionization mass spectrometry with wooden tips: A review, **Analytica Chimica Acta**, p. 339136, 2021.

JAMWAL, R *et al*. Rapid and non-destructive approach for the detection of fried mustard oil adulteration in pure mustard oil via ATR-FTIR spectroscopy-chemometrics, **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 244, p. 118822, 2021.

JIMÉNEZ-CARVELO, M. A. *et al*. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review, **Food Research International**, v. 122, p. 25–39, 2019.

KALOGIOURI, P. N. *et al*. Application of an advanced and wide scope non-target screening workflow with LC-ESI-QTOF-MS and chemometrics for the classification of the Greek olive oil varieties, **Food Chemistry**, v. 256, p. 53–61, 2018.

KALOGIOURI, P. N. *et al*. Application of High-Resolution Mass Spectrometric methods coupled with chemometric techniques in olive oil authenticity studies - A review, **Analytica Chimica Acta**, v. 1134, p. 150–173, 2020.

KENNARD, R W; STONE, L A. Computer Aided Design of Experiments, **Technometrics,** v. 11, n.1, 1969.

KOWALSKI, B. R. Chemometrics: Views and Propositions, **Journal of Chemical Information and Computer Science**s, v. 15, n. 4, p. 201–203, 1975.

KROPF, U. *et al*. Determination of the geographical origin of Slovenian black locust, lime and chestnut honey, **Food Chemistry**, v. 121, n. 3, p. 839–846, 2010.

LANÇAS, M. F. **Espectrometria de massas: fundamentos, instrumentação e aplicações**. Campinas: Átomo, 2019.

LI, L. *et al*. Rapid monitoring of black tea fermentation quality based on a solution-phase sensor array combined with UV-visible spectroscopy, **Food Chemistry**, v. 377, p. 131974, 2022.

LIU, J. *et al.* Development, Characterization, and Application of Paper Spray Ionization, **Analytical Chemistry**, v. 82, n. 6, p. 2463-2471, 2010.

LIMA, M. T. *et al.* Elemental analysis of Cerrado agricultural soils via portable X-ray fluorescence spectrometry : Inferences for soil fertility assessment, **Geoderma**, v. 353, p. 264–272, 2019.

MARCOZ, M. E. *et al*. The value of region of origin, producer, and protected designation of origin label for visitors and locals: the case of Fontina Cheese in Italy, **International Journal of Tourism Research**, v. 18, n. 3, p. 236–250, 2016.

MARTINEZ, P. E. H. *et al.* Nutrição mineral do cafeeiro e qualidade da bebida, **Revista Ceres**, v. 61, p. 838–848, 2014.

MÁRQUEZ, C. *et al*. FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud, **Talanta**, v. 161, p. 80-86, 2016.

MEDINA, S. *et al*. Current trends and recent advances on food authenticity technologies and chemometric approaches, **Trends in Food Science and Technology**, v. 85, n. January, p. 163–176, 2019.

MEDINI, S. et al. Methodological development for 87Sr/86Sr measurement in olive oil and preliminary discussion of its use for geographical traceability of PDO Nîmes (France), **Food Chemistry**, v. 171, p. 78–83, 2015.

MESQUITA, M. C. *et al*. **Manual do café: manejo de cafezais em produção**, EMATER-MG, 2016.

MIAW, W. S. C. **Detecção de fraudes em néctares de frutas: análises espectroscópicas aliadas a métodos de calibração e classificação multivariada**, Tese de doutorado, Belo Horizonte, UFMG, 2018.

MIAW, W. S. C. *et al*. Variable selection for multivariate classification aiming to detect individual adulterants and their blends in grape nectars, **Talanta**, v. 190, p. 55-61, 2018.

MILLER, N. J., MILLER, C. J. **Statistics and Chemometrics for Analytical Chemistry**, Editora Pearson Education, 5° edição, 2005.

MONJE, B. F. A. *et al.* ATR-FTIR for discrimination of espresso and americano coffee pods, **Coffee Science**, v. 13, n.4, p. 550-558, 2018.

MONTEIRO, C. M., TRUGO, C. L. Determinação de compostos bioativos em amostras comerciais de café torrado, **Química Nova**, v. 28, n. 4, p. 637–641, 2005.

MONTGOMERY, C. D., RUNGER, C. G. **Estatística aplicada e probabilidade para engenheiros**, 4° edição, Editora LTC, 2009.

MOREIRA, I. *et al*. Efeito do solvente na extração de ácidos clorogênicos, cafeína e trigonelina em coffea arabica, **Quimica Nova**, v. 37, n. 1, p. 39–43, 2014.

MUROGA, S. *et al.* Novel Approaches to In-Situ ATR-FTIR Spectroscopy and Spectroscopic Imaging for Real-Time Simultaneous Monitoring Curing Reaction and Diffusion of the Curing Agent at Rubber Nanocomposite Surface, **Polymers**, v.13, p.1-12, 2021.

NEHLIG, A. Effects of coffee/caffeine on brain health and disease: What should I tell my patients? **Practical Neurology**, v.16, p. 89-95, 2016.

NEVES, G. M., POPPI, J. R. Authentication and identification of adulterants in virgin coconut oil using ATR/FTIR in tandem with DD-SIMCA one class modeling, **Talanta**, v. 219, 2020.

NOVAES, G. C. *et al.* Otimização de Métodos Analíticos Usando Metodologia de Superfícies de Respostas - Parte II : Variáveis de Mistura, **Revista Virtual de Química**, v. 10, n. 2, p. 393–420, 2018.

OLIVERI, P. *et al*. Comparison between classical and innovative class-modelling techniques for the characterisation of a PDO olive oil, **Analytical and Bioanalytical Chemistry**, v. 399, n. 6, p. 2105–2113, 2011.

OLIVERI, P., DOWNEY, G. Multivariate class modeling for the verification of food-authenticity claims, **TrAC - Trends in Analytical Chemistry**, v. 35, p. 74–86, 2012.

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

PAVIA, D. L. *et al*., **Introdução à Espectroscopia**, Editora Cengage Learning, 1°edição, 2010.

PEREIRA, V. H. **Espectrometria de massas com ionização por paper spray combinada a métodos quimiométricos para identificação de falsificações em cervejas**, Dissertação de mestrado, Belo Horizonte, MG, UFMG, 2016.

REIS, C., ANDRADE, C. J. Planejamento experimental para misturas usando cromatografia em papel, **Química Nova**, v.19, n.3, p. 313-319, 1996.

REIS, N. **Detecção de adulteração de café torrado e moído com cascas de café e milho por espectroscopia no infravermelho**, Dissertação de mestrado, Belo Horizonte, UFMG, 2012.

REN, X. *et al.* UV spectroscopy and HPLC combined with chemometrics for rapid discrimination and quantification of Curcumae Rhizoma from three botanical origins, **Journal of Pharmaceutical and Biomedical Analysis**, v. 202, p. 114145, 2021.

RIBEIRO, E. D. **Descritores químicos e sensoriais para discriminação da qualidade da bebida de café arábica de diferentes genótipos e métodos de processamento**, Tese de doutorado, Lavras, MG, 2017.

RIBEIRO, S. J. *et al*. Prediction of a wide range of compounds concentration in raw coffee beans using NIRS, PLS and variable selection, **Food Control**, v. 125, 2021.

RÍOS-REINA, R. *et al*. NIR spectroscopy and chemometrics for the typification of Spanish wine vinegars with a protected designation of origin, **Food Control**, v. 89, p. 108–116, 2018.

RÍOS-REINA, R. *et al*. Data fusion approaches in spectroscopic characterization and classification of PDO wine vinegars, **Talanta**, v. 198, p. 560–572, 2019.

RÍOS-REINA, R. *et al*. Characterization of the aroma profile and key odorants of the Spanish PDO wine vinegars, **Food Chemistry**, v. 311, p. 126012, 2020.

ROCCHETTI, G. *et al*. A combined metabolomics and peptidomics approach to discriminate anomalous rind inclusion levels in Parmigiano Reggiano PDO grated hard cheese from different ripening stages, **Food Research International**, v. 149, p. 110654, 2021.

SEBRAE, INPI, **Indicações geográficas brasileira: Café**, 2ª edição, 2016.

SHAH, D, *et al*. A spectroscopic chemometric modeling approach based on statistics pattern analysis, **IFAC – Papers online**, v.51 (18), p. 369-374, 2018.

SHAND, A. C. *et al.* Multivariate analysis of Scotch whisky by total reflection x-ray fluorescence and chemometric methods: A potential tool in the identification of counterfeits, **Analytica Chimica Acta**, v. 976, p. 14–24, 2017.

SHARMA, V. *et al*. On the rapid and non-destructive approach for wood identification using ATR-FTIR spectroscopy and chemometric methods, **Vibrational Spectroscopy**, v. 110, p. 103097, 2020

SILVA, B. M. *et al*. Indicações geográficas: um panorama de estudos recentes geographical indications: an overview of recent studies, **INGI - Indicação Geográfica e Inovação**, v. 4, p. 780–801, 2020.

SKOOG, D. A., WEST, D. M., HOLLER, F. J., CROUCH, S. R., **Fundamentos de Química Analítica**. São Paulo: Thomson, 2006.

SOUSA, C. L. *et al.* Desenvolvimento de modelos de calibração NIRS para minimização das análises de madeiras de *Eucalyptus spp*., **Ciência Florestal**, v. 21, n.3, p. 591-599, 2011.

SOUTO, P. C. T. U**. Metodologia baseada em imagem digital, espectros UV-Vis e quimiometria para screening de adulteração de café por cascas e paus**, Tese de doutorado, João Pessoa, PB, 2017.

SOUZA, R. R. *et al*. Honey authentication in terms of its adulteration with sugar syrups using UV–Vis spectroscopy and one-class classifiers*,* **Food Chemistry**, v. 365, p. 130467, 2021.

SZOBOSZLAI, N. *et al*. Recent trends in total reflection X-ray fluorescence spectrometry for biological applications, **Analytica Chimica Acta**, v. 633, n. 1, p. 1–18, 2009.

TAHIR, E. H. *et al.* The use of analytical techniques coupled with chemometrics for tracing the geographical origin of oils: A systematic review (2013–2020), **Food Chemistry**, v. 366, p. 130633, 2022.

TAN, J. *et al*. Chemometric classification of Chinese lager beers according to manufacturer based on data fusion of fluorescence, UV and visible spectroscopies, **Food Chemistry**, v. 184, p. 30–36, 2015.

TARHAN, İ. A comparative study of ATR-FTIR, UV–visible and fluorescence spectroscopy combined with chemometrics for quantification of squalene in extra virgin olive oils, **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 241, p. 118714, 2020.

TEÓFILO, F. R., FERREIRA, C. M. M. Quimiometria II: planilhas eletrônicas para cálculos de planejamentos experimentais, um tutorial, **Química Nova**, v. 29, n. 2, p, 338-350, 2006.

TEÓFILO, F. R. *et al.* Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, **Journal of Chemometrics**, v. 23, n. 1, p. 32–48, 2009.

TEÓFILO, R F. **Métodos Quimiométricos: Uma Visão Geral - Conceitos básicos de quimiometria.** v. 1, n. 1, UFV, Viçosa, 2013.

TOCI, A. *et al*. Efeito do processo de descafeinação com diclorometano sobre a composição química dos cafés arábica e robusta antes e após a torração, **Química Nova**, v. 29, n.5, p. 965-971, 2006.

URÍČKOVÁ, V., SÁDECKÁ, J. Determination of geographical origin of alcoholic beverages using ultraviolet, visible, and infrared spectroscopy: A review, **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 148, p. 131–137, 2015.

VANDEN BRANDEN, K.; HUBERT, M. Robust classification in high dimensions based on the SIMCA Method, **Chemometrics and Intelligent Laboratory Systems**, v. 79, n. 1–2, p. 10–21, 2005.

VARÃO SILVA, T. *et al.* Tracing commercial coffee quality by infrared spectroscopy in tandem with pattern recognition approaches, **Vibrational Spectroscopy**, v. 116, p. 103295, 2021.

VEIGA, D. A. *et al*. Arabica coffee cultivars in different water regimes in the central Cerrado region, **Coffee Science**, v. 14, n. 3, p. 349–358, 2019.

VITALI ČEPO, D. *et al*. Application of benchtop total-reflection X-ray fluorescence spectrometry and chemometrics in classification of origin and type of Croatian wines, **Food Chemistry**, v. 13, p. 100209, 2022.

VON BOHLEN, A. Total reflection X-ray fluorescence and grazing incidence X-ray spectrometry - Tools for micro- and surface analysis. A review, **Spectrochimica Acta - Part B Atomic Spectroscopy**, v. 64, p. 821-832, 2009.

WANG, Y. Y. *et al.* Molecules Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy (ATR-FTIR) Combined with Chemometrics Methods for the Classification of Lingzhi Species, **Molecules**, v.24, p. 1-13, 2019.

WISE, M. B., *et al*. **PLS_Toolbox 4.0 for use with MATLAB™**. Disponível em: <www.eigenvector.com>.

WOLD, S., SJÖSTRÖM, M. **SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy, Chemometrics: Theory and Applications**, capítulo 12, 1977.

XU, L., *et al*. One-class partial least squares (OCPLS) classifier, **Chemometrics and Intelligent Laboratory Systems**, v. 126, p. 1–5, 2013.

XU, L. *et al.* A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers, **Chemometrics and Intelligent Laboratory Systems**, v. 139, p. 58–63, 2014a.

XU, L. *et al*. A MATLAB toolbox for class modeling using one-class partial least squares (OCPLS) classifiers, **Chemometrics and Intelligent Laboratory Systems**, v. 139, p. 58–63, 2014b.

ZONTOV, V. Y. *et al.* DD-SIMCA – A MATLAB GUI tool for data driven SIMCA approach, **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 23–28, 2017.

ANEXO

Tabela 1A – Resultados obtidos pela ANOVA para o planejamento de misturas para as condições de escolha do solvente.

	Soma Quadrática	Graus de liberdade	Média Quadrática	Valor de F	Valor p
Mistura Linear	13290,81	2	6645,41	650,37	<0,0001
AB	25,91	1	25,91	2,540	0,1553
AC	0,18	1	0,18	0,0177	0,8979
BC	23,00	1	23,00	2,250	0,1772
ABC	69,16	7	69,16	6,770	0,0353
Erro puro	71,52	7	10,22		
Total	13448,19	13			

Tabela 2A – Resultados obtidos pela ANOVA para o planejamento composto central para as condições de extração com o modo de estático.

	Soma Quadrática	Graus de liberdade	Média Quadrática	Valor de F	Valor p
A – Temperatura	262,82	1	262,82	4,55	0,0655
A^2	1392,59	1	1392,59	24,09	0,0012
Falta de ajuste	447,67	6	74,61	10,13	0,0926
Erro puro	14,73	2	7,36		
Total	2117,81	10			

Tabela 3A – Resultados obtidos pela ANOVA para o planejamento fatorial 2^3 para as condições de extração o modo de contato por ultrassom.

	Soma Quadrática	Graus de liberdade	Média Quadrática	Valor de F	Valor p
A – Temperatura	491,96	1	491,96	19,45	0,0031
B – Tempo	56,55	1	56,55	2,24	0,1785
B^2	391,74	1	391,74	15,49	0,0056
Falta de ajuste	165,37	5	33,07	5,65	0,1572
Erro puro	11,71	2	5,85		
Total	1117,33	10			

Tabela 4A – Resultados obtidos para os modelos DD-SIMCA e OCPLS para fusão PSMS e FTIR com os dados originais concatenados (a) e após seleção de variáveis.

a) Dados originais

		Treinamento	*Teste*		
Métodos	**PCs/ VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
DD-SIMCA	8	1,00	1,00	0,30	0,55
OCPLS	8	0,96	1,00	0,00	-
b) Seleção de variáveis					
DD-SIMCA	6	0,99	1,00	0,37	0,61
OCPLS	7	0,97	1,00	0,20	0,45

Tabela 5A – Resultados obtidos para os modelos DD-SIMCA e OCPLS para fusão PSMS e TXRF com os dados originais concatenados (a) e após seleção de variáveis.

a) Dados originais

		Treinamento	*Teste*		
Métodos	**PCs/ VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
DD-SIMCA	6	0,97	1,00	0,17	0,41
OCPLS	6	0,93	0,97	0,03	0,18
b) Seleção de variáveis					
DD-SIMCA	6	0,99	1,00	0,40	0,63
OCPLS	2	0,90	1,00	0,03	0,18

Tabela 6A – Resultados obtidos para os modelos DD-SIMCA e OCPLS para fusão FTIR, PSMS e TXRF com os dados originais concatenados (a) e após seleção de variáveis.

a) Dados originais

		Treinamento	*Teste*		
Métodos	**PCs/ VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
DD-SIMCA	7	1,00	1,00	0,27	0,52
OCPLS	6	0,93	0,97	0,03	0,18
b) Seleção de variáveis					
DD-SIMCA	6	1,00	1,00	0,30	0,55
OCPLS	4	0,91	0,93	0,07	0,25

Tabela 7A – Resultados obtidos para a fusão FTIR e TXRF com os dados originais concatenados (a) e com 80 variáveis selecionadas (b).

a) Dados originais

		Treinamento	*Teste*		
Métodos	**PCs/ VLs**	**Sensibilidade**	**Sensibilidade**	**Especificidade**	**Eficiência**
SIMCA	4	0,97	0,97	0,50	0,70
DD-SIMCA	4	0,97	1,00	0,13	0,37
OCPLS	5	0,93	0,97	0,20	0,43
b) Seleção de variáveis					
SIMCA	6	0,97	1,00	0,30	0,55
DD-SIMCA	6	0,99	1,00	0,13	0,37
OCPLS	5	0,93	1,00	0,20	0,45