Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Química

Camila Assis

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

Belo Horizonte
2018

UFMG/ICEx/DQ. 1303ª

T. 590ª

Camila Assis

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 apresentada ao Departamento de Química do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Doutor em Ciências – Química

Belo Horizonte

2018

Ficha Catalográfica

A848a
2018
T

Assis, Camila
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 [manuscrito] / Camila Assis. 2018.
[xiv], 134 f. : il.

Orientador: Marcelo Martins de Sena.

Tese (doutorado) – Universidade Federal de Minas Gerais – Departamento de Química.
Inclui bibliografia.

1. Química analítica - Teses 2. Análise espectral - Teses 3. Quimiometria - Teses 4. Controle de qualidade – Teses 5. Café - Teses 6. Espectroscopia de infravermelho – Teses 7. Espectrometria de massa - Teses 8. Fluorescência de raio X – Teses I. Sena, Marcelo Martins de, Orientador II. Título.

CDU 043

Elaborada pela Biblioteca do Departamento de Química - UFMG

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

50 anos química pós-graduação-ufmg

"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"

Camila Assis

Tese aprovada pela banca examinadora constituída pelos Professores:

Prof. Marcelo Martins de Sena - Orientador
UFMG

Prof. Marco Flôres Ferrão
UFRGS

Prof. Marcello Garcia Trevisan
UNIFAL

Profa. Letícia Malta Costa
UFMG

Prof. Bruno Gonçalves Botelho
UFMG

Belo Horizonte, 26 de setembro de 2018.

Agradecimentos

Muitas razões para agradecer e muitas pessoas para homenagear durante todo esse processo.

Agradeço à Jesus Cristo, meu fiel e amado amigo. Grata por Sua infinita misericórdia, bondade e fidelidade.

Aos meus pais, Mario e Meré, donos do meu amor e da minha admiração. Privilégio sem tamanho ser filha de vocês e poder aprender todos os dias.

Às minhas irmãs, Amanda e Rafaela, aos Rodrigo´s e aos meus sobrinhos amados: Elisa, André e Rodriguinho. Vocês me enchem de ânimo e alegria.

Agradeço ao meu orientador, Dr. Marcelo Sena, por toda amizade, auxilio, paciência e por sua sempre boa vontade em ensinar. Obrigada pela dedicação e por toda disponibilidade para acrescentar tanto nesse trabalho.

Agradeço ao professor, Dr. Leandro Soares de Oliveira, por todo ensinamento e por, principalmente, fornecer a estrutura básica para que este trabalho fosse realizado.

Agradeço aos professores Dra. Clésia Nascentes e Dr. Rodinei Augusti, por toda colaboração e parceria.

Agradeço a amigos queridos: Bethania, Carol, Paula, Naara, Tati, Diego, Rodger e Marcelo por todo incentivo e ajuda em momentos preciosos.

Agradeço a Luiza, amiga querida, presente de Deus. Obrigada por me ouvir sempre com paciência, alegria e bom humor. Você é uma pérola. Obrigada por toda ajuda, fundamental, em BH. Sem você seria muito mais difícil.

Agradeço ao querido Tio Luiz por todo incentivo, enviando sempre um conteúdo interessante sobre café.

Agradeço aos membros do GQQATE, por todo companheirismo, ajuda e boa convivência nesses 4 anos de estudo.

Agradeço a CAPES por todo apoio financeiro.

Agradeço a todos que contribuíram, direta ou indiretamente, para que este trabalho fosse realizado.

Resumo

O principal objetivo desta tese foi desenvolver modelos multivariados para quantificar e caracterizar misturas de cafés Robusta e Arabica. Para este fim, 120 misturas de café moído (0-33% m/m, café Robusta em *blends* com café Arabica), preparadas com amostras obtidas de dez produtores diferentes, foram formuladas em três diferentes graus de torra: leve, média e escura. Diferentes técnicas instrumentais foram utilizadas: espectroscopias no infravermelho médio com refletância total atenuada (ATR-FTIR ou MIR), no infravermelho próximo (NIR), espectrometria de massas com ionização por *paper spray* (PS-MS) e fluorescência de raios X por reflexão total (TXRF). Modelos utilizando regressão por mínimos quadrados parciais (PLS) foram construídos individualmente para os espectros de cada técnica. Paralelamente, modelos de fusão de dados (diferentes combinações entre as técnicas) também foram construídos nos níveis baixo e médio, de forma a aproveitar a sinergia entre os conjuntos de dados. Os modelos foram otimizados por métodos de seleção de variáveis, tais como algoritmo genético (GA) e seleção de preditores ordenados (OPS). Em geral, os menores erros de previsão foram fornecidos pelo modelo de fusão de dados de nível baixo. Em todos os casos, os métodos de seleção de variáveis reduziram consideravelmente os valores de erro quadrático médio de previsão (RMSEP) e o número de variáveis, aumentando os valores de coeficiente de correlação entre valores previstos e de referência. Os modelos construídos foram interpretados através de seus vetores informativos e, de forma geral, compostos específicos presentes no café foram fundamentais para diferenciar as espécies, tais como trigonelina, açúcares e ácidos clorogênicos. Para os dados obtidos de TXRF, destacaram-se os elementos Mn e Rb como possíveis marcadores das espécies. Os melhores modelos (MIR e MIR-PSMS) foram validados e figuras de mérito adequadas

foram estimadas, corroborando sua exatidão, linearidade, sensibilidade e ausência de viés.

Palavras chave: café, *blends*, quimiometria, seleção de variáveis, fusão de dados.

Abstract

The main objective of this thesis was to develop multivariate models to quantify and characterize mixtures of Robusta and Arabica coffees. For this purpose, 120 blends of ground coffees (0.0-33.0% m/m), prepared with coffee samples originated from ten different farmers, were formulated at three different degrees of roasting: light, medium and dark. Different instrumental techniques were used: attenuated total reflectance Fourier transform infrared (ATR-FTIR or MIR) spectroscopy, near infrared (NIR) spectroscopy, paper spray ionization mass spectrometry (PS-MS) and total reflection X-ray fluorescence (TXRF). Models using partial least squares regression (PLS) were built individually for the spectra from each technique. In the sequence, data fusion models (different combinations of techniques) were also built at low and medium levels, in order to take advantage of the synergy between the datasets. The models were optimized by variable selection methods, such as genetic algorithm (GA) and ordered predictors selection (OPS). In general, the smallest prediction errors were provided by the low-level data fusion models. In all the cases, the variable selection methods significantly reduced the mean square errors of prediction (RMSEP) and the number of variables, increasing the correlation coefficient values between predicted and reference values. PLS models were interpreted through informative vectors and specific coffee components were detected as marker species, such as trigonelline, sugars and chlorogenic acids. For the atomic data, the elements Mn and Rb were mostly detected as possible markers of the coffee species. The best models (MIR and MIR-PSMS) were validated and proper figures of merit were estimated, corroborating their accuracy, linearity, sensitivity and absence of bias.

Key words: coffee, blends, chemometrics, variable selection, data fusion.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AOAC	*Association of Official Analytical Chemists*
ATR	*Attenuated Total Reflectance (*Reflectância Total Atenuada)
CI	*Chemical Ionization* (Ionização Química)
DESI	*Desorption Electrospray Ionization* (Ionização de dessorção por *eletrospray*)
DPR	Desvio Padrão Relativo
DW	Durbin-Watson
BF	Brown-Forsythe
EDXRF	*Energy Dispersive X-ray Fluorescence (*Fluorescência de Raios X por Dispersão em Energia)
EI	*Electron Ionization* (Impacto Eletrônico)
ELDI	*Electrospray Assisted Laser Desorption/Ionization*
ER	Erro Relativo
ESI	*Electrospray Ionization* (Ionização por *Electrospray*)
FOM	*Figure of Merit* (Figura de Mérito)
FTIR	*Fourier-transform Infrared Spectroscopy* (Espectroscopia no Infravermelho com Transformada de Fourier)
GA	*Genetic algorithm* (Algoritmo genético)
HCA	*Hierarchical Cluster Analysis* (Análise de Agrupamento Hierárquico)
HPLC	*High Perfomance Liquid Chromatography* (Cromotografia Líquida de Alta Eficiência)
ICO	*International Coffee Organization* (Organização Internacional do Café)
iPLS	*Interval Partial Least Square* (Quadrados Mínimos Parciais por Intervalos)
IV	Infravermelho
MIR	*Mid-Infrared* (Infravermelho Médio)
MSC	*Multiplicative Scatter* Correction (Correção do Espalhamento Multiplicativo)
NIPALS	*Non linear Iterative Partial Least Squares*
NIR	*Near-Infrared* (Infravermelho próximo)
nVars	Número de Variáveis
OPS	*Ordered Predictors Selection* (Seleção dos Preditores Ordenados)
PCA	*Principal Component Analysis* (Análise de Componentes Principais)
PCR	*Principal Component Regression* (Regressão por Componentes Principais)
PLS	*Partial least squares* (Mínimos Quadrados Parciais)
PLS-DA	*Partial Least Squares Discriminant Analysis* (Análise Discriminante por Mínimos Quadrados Parciais)

PS-MS	*Paper Spray Mass Spectrometry (Espectrometria de Massas com Ionização Paper Spray*)
QSAR	*Quantitative Structure Activity Relationship* (Relações Quantitativas Estrutura-Atividade)
R	Coeficiente de Correlação
R^2	Coeficiente de Determinação
RJ	Ryan-Joiner
RMN	Ressonância Magnética Nuclear
RMSEC	*Root Mean Squares Error of Calibration (*Raiz Quadrada do Erro Médio Quadrático de Calibração)
RMSECV	Root Mean Squares Error of Cross Validation (Raiz Quadrada do Erro Médio Quadrático da Validação Cruzada)
RMSEP	*Root mean square error of prediction* (Raiz quadrada do erro médio quadrático de previsão)
RPD	*Residual Prediction Deviation* ou *Ratio of Performance to Deviation* (Razão de Desempenho de Desvio)
SIMCA	*Soft Independent Modeling of Class Analogy* (Modelagem suave independente por analogia de classe)
SNV	*Standard Normal Variate* (Variação Normal Padrão)
SPA	Successive Projection Algorithm (Algoritmo das Projeções Sucessivas)
SVPII	Seleção de Variáveis Preditivas com Base em seus Índices de importância
TXRF	*Total Reflection X-ray Fluorescence* (Fluorescência de Raios X por Reflexão Total)
VIP	*Variable Importance in the Projection* (Importância de Variáveis na Projeção)
VL	Variáveis Latentes

Sumário

1. Introdução

O café é uma das bebidas mais populares do mundo. Estimativas indicam que é a segunda bebida mais consumida, ficando atrás apenas da água [1]. Nesse cenário, o café possui um dos maiores impactos socioeconômicos, se comparado às diferentes atividades ligadas ao comércio agrícola, podendo ser classificado como uma grande *"commodity"* mundial [2]. Sua popularidade se deve, essencialmente, aos seus efeitos fisiológicos benéficos à saúde, além de agradáveis sabor e aroma [3].

O cultivo de café está presente em mais de 60 países, destacando-se Colômbia, Brasil, Indonésia e Vietnã [4]. O Brasil, particularmente, possui papel de destaque, sendo o maior produtor mundial, respondendo por aproximadamente 36% da produção e cerca de 30% do mercado mundial exportador. A cafeicultura se faz presente em quinze Estados da Federação, possibilitando uma diversificada disposição da produção em todo o território brasileiro. Neste contexto, destaca-se o estado de Minas Gerais, responsável por mais da metade da produção do café nacional. Além disso, o país também se destaca como o segundo maior mercado consumidor mundial (atrás somente dos Estados Unidos), com aproximadamente 21,5 milhões de sacas (o que corresponde a 1,07 milhões de toneladas) consumidas em 2017 e com uma forte tendência de crescimento do consumo [5].

O nome genérico *Coffea* possui, aproximadamente, 103 espécies catalogadas [6]. Do ponto de vista econômico, as duas espécies mais importantes cultivadas no mundo são a Arabica (*Coffea arabica*) e a Robusta (*Coffea canephora*). O café Arabica responde por cerca de 56% da produção mundial, com o restante, 44%, proveniente do Café Robusta [7]. Dados recentes, referente aos doze meses encerrados em abril de 2018, estimam que as exportações de Arabica totalizaram

75,93 milhões de sacas, em comparação com 74,08 milhões de sacas no ano anterior; enquanto as exportações de Robusta totalizaram 44,32 milhões de sacas, em comparação com 44,63 milhões de sacas do ano de 2017 [8].

Ambas as espécies diferem não apenas em relação às suas características botânicas e composição físico-química, mas também em termos de valor comercial, com cafés Arabica apresentando preços de mercado entre 20-25% mais elevados [4]. O café Arabica é originário da Etiópia e cultivado em regiões acima de 800 metros de altitude e de clima ameno. Além disso, esta cultura necessita de condições especiais de processamento, controle de pragas (essencialmente ferrugem), entre outros fatores; exigências que contribuem para um maior custo na produção [9]. O café Robusta, em contrapartida, é adaptado à condições de temperaturas médias anuais entre 22 a 26°C, tolera bem a seca e altitudes abaixo de 500 metros, podendo ser cultivado ao nível do mar. Devido a essas características, esta variedade permite a produção em locais onde o café Arabica não se adapta.

Do ponto de vista sensorial, o café Arabica é um produto reconhecido como de maior qualidade e de aroma superior, enquanto o Robusta possui um sabor mais amargo, com maior quantidade de sólidos solúveis, antioxidantes e cafeína [3]. Os cafés de melhor qualidade, geralmente, utilizam exclusivamente combinações de Arabica, recebendo frequentemente a denominação de café *gourmet*.

Neste sentido, a instabilidade nos preços que ocorre no agronegócio de café, ao mesmo tempo em que gera problemas relacionados à qualidade dos produtos, abre caminho para outra discussão: a formulação de *blends* do café Arabica com o café Robusta no setor de cafés torrados e moídos. Define-se *blends* como a arte de combinar grãos de cafés [10], com características complementares – acidez com doçura, torra clara com torra escura – entre outros fatores. Nesse sentindo, pode-se

combinar cafés de diferentes espécies, regiões e safras. Como a diferença entre os preços das duas espécies aumentou consideravelmente nos últimos anos, há uma forte tendência de substituição (ilegal) de café Arabica pelo Robusta [11]. Como o café Robusta tem um preço mais baixo se comparado ao café Arabica, a elaboração de *blends* é altamente viável devido à diminuição nos preços de custo da produção.

Conforme discutido, os cafés Arabica e Robusta apresentam diferenças no teor de diversos componentes. Assim, variações nos *blends* empregados impactam também na composição e qualidade da bebida. Dessa forma, o menor custo de grãos de café Robusta abre a possibilidade para fraudes comerciais, principalmente as relacionados aos *blends* 100% de misturas de café Arabica, tornando a detecção e quantificação de tais misturas em amostras comerciais uma demanda analítica necessária para a proteção do consumidor [12].

O método oficial descrito na "*Association of Official Analytical Chemists*" (AOAC) para análise de café, especialmente para determinação de micotoxinas, é baseado em Cromatografia Líquida de Alta Eficiência (CLAE) [13]. No entanto, a necessidade de reagentes de alto grau de pureza, equipamento e manutenção de alto custo, grande tempo de análise e grande geração de resíduos, dificultam o seu emprego em uma rotina laboratorial [14]. Portanto, existe a necessidade de se desenvolver novas metodologias analíticas, mais rápidas e eficientes, que possam ser implementadas em análises de rotina para identificação das diferentes espécies de café e de suas respectivas proporções na formulação dos *blends*.

Dependendo da característica de qualidade que está sendo monitorada e da natureza da matriz, diversas técnicas instrumentais podem ser utilizadas para o controle de qualidade do café. Para análise de café, as principais técnicas destacadas na literatura são: espectroscopia de absorção molecular no infravermelho (médio e

próximo) [15–18], cromatografias líquida e gasosa [19–21]; espectrometria de massas [22–24]; ressonância magnética nuclear [11,25], [26]; espectroscopia Raman [27,28], entre outras. Apesar de todas as técnicas mencionadas possuírem eficácia adequada, a espectroscopia no infravermelho é uma técnica rápida, confiável, simples de executar e exige um mínimo preparo de amostra, com o procedimento de amostragem e análise geralmente levando cerca de 5 min [18]. Dessa forma, a espectroscopia no infravermelho, quando usada em conjunto com métodos quimiométricos é uma ferramenta eficaz para análises quantitativas.

Apesar de métodos desenvolvidos para a análise direta de café baseados na técnica de espectroscopia no infravermelho já terem sido descritos na literatura em vários artigos, as informações sobre a composição química das amostras e quais são os compostos responsáveis pela diferenciação entre as variedades não estão disponíveis ou são superficialmente descritas nessas publicações [22]. Nesse sentido, a espectrometria de massas aliada a métodos de ionização ambiente tem surgido como opção de análise rápida, direta, com pouco (ou nenhum) preparo de amostra, sendo descrita como uma técnica altamente seletiva e sensível. *Paper Spray* é uma técnica de ionização recentemente desenvolvida que tem demonstrado alta eficácia em análises de matrizes complexas [29]. Dentre as vantagens dessa fonte de ionização podem-se citar: consumo mínimo de solvente e de amostra, rapidez e não utilização de gases [30–32].

Além das técnicas previamente mencionadas, as quais são todas moleculares, informações relacionadas à composição elementar também podem ser fundamentais para uma melhor caracterização dos *blends*. Nesse sentido, a Fluorescência de Raios-x por Reflexão Total (TXRF - *Total Reflection X-r*ay *Fluorescence*) é uma técnica analítica importante, devido à diversas vantagens: simplicidade, baixos limites de

quantificação e detecção, rapidez, facilidade de operação, possibilidade de detectar quase todos os elementos da tabela periódica, entre outras [33].

Uma abordagem quimiométrica que tem se destacado na literatura, essencialmente nos últimos anos, é a fusão de dados, a qual pode ser definida, em linhas gerais, como a junção de blocos de dados provenientes de diferentes origens ou técnicas analíticas em um único modelo [34]. Dessa forma, a união de diferentes ferramentas analíticas pode promover uma maior caracterização de *blends* de café, levando a modelos de previsão com menores erros. A junção de técnicas de espectroscopia molecular, espectrometria de massas e informações atômicas pode permitir a construção de modelos mais eficientes, que incorporem informações complementares, a fim de aproveitar as vantagens de sinergia entre as fontes de dados combinadas.

2. Objetivo

A proposta desta tese é o desenvolvimento metodologias analíticas rápidas, eficientes, baratas, com um gasto consideravelmente menor de reagentes, e como consequência, mais amigáveis ambientalmente, para caracterizar as diferentes espécies de café (Robusta e Arábica) e determinar a constituição de seus *blends.* Os métodos convencionais de análise (via úmida) são demorados e custosos. Dessa forma, este trabalho pretende contribuir diretamente para melhorar a qualidade e inocuidade de produtos de origem vegetal, garantindo e ampliando a introdução do café brasileiro nos mercados nacional e mundial.

2.1 Objetivos Específicos e Metas

Desenvolver e otimizar metodologias analíticas baseadas em espectroscopia NIR, MIR, PS-MS e TXRF de forma a caracterizar os diferentes *blends* (Robusta e Arábica);

Desenvolver modelos de calibração multivariada (regressão por quadrados mínimos parciais, PLS) de forma a quantificar (e caracterizar) a presença de café Robusta em *blends* de café Arabica;

Utilizar diferentes métodos de seleção de variáveis de forma a eliminar variáveis irrelevantes, simplificando e aumentando a capacidade preditiva dos modelos;

Utilizar a estratégia de fusão de dados (níveis baixo e médio) de forma a aproveitar a sinergia entre os conjuntos de dados de diferentes origens, atômica e molecular.

Realizar uma validação analítica multivariada completa dos modelos com os melhores desempenhos, de forma a avaliar a sua qualidade.

3. Revisão bibliográfica

3.1. Café

O café é uma das bebidas populares mais consumidas em todo o mundo e seu consumo mundial está aumentando consideravelmente. Segundo dados da Organização Internacional do Café (ICO), cerca de 1,4 bilhões de xícaras de café são consumidas por dia em todo o mundo [8]. Entre os produtos economicamente valorizados, o café representa a segunda maior *commodity* mundial, envolvendo a África, Ásia e América Latina como os principais parceiros exportadores, além de mais de 70 países produtores, enquanto a exportação envolve mais de 150 países [8]. Além disso, estima-se que entre 75 e 125 milhões de pessoas trabalham no setor. Para o ano cafeeiro de 2017/2018, a produção mundial de café está estimada em cerca de 160 milhões de sacas, crescimento correspondente a 1,2%, se comparado ao ano anterior. Nesse cenário, o Brasil é o maior produtor mundial de café, respondendo por aproximadamente 36 % da produção mundial e cerca de 30 % do mercado mundial exportador. De acordo com a ICO *(International Coffee Organization)*, para a safra de 2017/2018, no Brasil, 51 milhões de sacas foram produzidas, respondendo por volta de 31,9% da produção mundial [8].

Do ponto de vista biológico, a *Rubiaceae* é uma das maiores famílias botânicas de plantas de florescimento e pode ser subdividida em três subfamílias: *Rubioideae*, *Cinchonoideae* e *Ixoroideae*. Essa família é composta por 43 tribos, 611 gêneros e 13.143 espécies. Todas as plantas de café pertencem à tribo *Coffea* da subfamília *Cinchonoideae*, que é de longe a planta mais importante, economicamente, dentro da família e uma das *commodities* mais importante do mundo [35,36]. Atualmente, mais de 100 espécies diferentes de café foram catalogadas, porém *Coffea arabica*, *Coffea*

canephora (variedade robusta ou conilon), *Coffea liberica* e *Coffea excelsa* são as espécies mais populares. Do ponto de vista comercial, apenas a C. arabica e a C. canephora var. robusta (usualmente conhecidas como Arabica e Robusta, respectivamente) possuem importância, e representam as duas espécies mais relevantes e amplamente cultivadas, respondendo por 56% e 44%, respectivamente, da produção mundial [4]. Considerando a safra 2017/2018, a produção mundial de café Arabica foi estimada em 97,16 milhões de sacas (queda estimada em 6,6% em relação ao ano anterior). Já o café Robusta atingiu cerca de 61,40 milhões de sacas (crescimento em torno de 11,5% se comparado a 2016).

A principal razão para a disparidade de preços entre Arabica e Robusta (2–3 vezes maior) é causada por muitas diferenças entre as duas espécies, tais como composição química, clima ideal de crescimento, aspectos físicos, características da bebida final, dentre outros [3]. O café Arabica é nativo das terras altas da Etiópia e atualmente é cultivado no continente americano, na África e na Ásia, limitando-se a regiões acima de 800 metros de altitude e de clima ameno. O o café Robusta é originário da África equatorial, e atualmente é cultivado em alguns países da África Central e Ocidental, na Ásia e na América do Sul. A espécie Robusta é de clima quente, com predominância de plantios para regiões de altitudes abaixo de 500 m. Dessa forma, o café Robusta é mais resistente a doenças, e cresce em terras planas, conduzindo a uma produção mais fácil e mecanizada.

Do ponto de vista sensorial, o café Arabica fornece uma bebida de qualidade superior (sabor e aroma) em comparação ao Robusta [23]. Em linhas gerais, o café Arabica tem um pronunciado sabor doce/floral, enquanto o café Robusta resulta em uma bebida mais amarga. As principais diferenças na composição química de Arabica e Robusta são mostradas na Tabela 1.

Tabela 1. Composição química do grão de café cru (adaptada de Costa [10])

Componentes	Café Arábica[a]	Café Robusta[a]
Cafeína	1,2	2,2
Trigonelina	1	0,7
Cinzas (41% corresponde a K)	4,2	4,4
Ácidos:		
Clorogênico Total	6,5	10
Alifáticos	1,0	1,0
Quínico	0,0	0,4
Açúcares:		
Sacarose	8	4
Redutores	0,1	0,4
Polissacarídeos	44	48
Lignina	3	3
Pectina	2	2
Proteína	11	11
Aminoácidos Livres	0,5	0,5
Lipídeos	16	10

[a]Valores expressos em g $100g^{-1}$ em base seca

Em forma de grão cru, a diferenciação entre Robusta e Arabica pode ser observada visualmente, pois os grãos de Arabica são normalmente maiores do que grãos de Robusta (Figura 1). Além disso, diferenças na forma e cor do grão são normalmente visíveis. Porém, no caso de grãos de café torrados a discriminação com base no tamanho pode levar a resultados enganosos [11].

Figura 1. Grãos de café Arabica (esquerda) e Robusta (direita)

Nesse sentido, o processo de torrefação é uma etapa crucial no processamento de café, pois causa transformações químicas, físicas e estruturais que resultam no sabor e aroma únicos da bebida. Em linhas gerais, a torrefação é um processo complexo do ponto de vista químico, uma vez que durante esta etapa centenas de reações químicas acontecem simultaneamente [37]. Durante a torra, os grãos mostram uma perda de peso constante, devido a uma desidratação inicial e, em seguida, sofrem um conjunto de processos físicos e químicos complexos. Paralelamente, os grãos são submetidos a um aumento significativo do volume, causado essencialmente pela formação interna de gás CO_2 e de outros produtos de decomposição térmica [38]. Um grande número de transformações químicas ocorre simultaneamente durante a torra (reações de Maillard e Strecker, degradação de proteínas, polissacarídeos, ácidos clorogênicos e trigonelina), conduzindo à formação de milhares de moléculas, as quais fornecem o aroma peculiar de café. Dessa forma, a cor dos grãos torrados é um importante marcador do processo de torra, dependendo, essencialmente, das condições desse processo e das características dos grãos verdes. Na realidade, a alteração da cor dos grãos é o resultado das transformações complexas que ocorrem durante as reações de Maillard, em que aminoácidos e açúcares redutores iniciam uma complexa cascata de reações durante o aquecimento, resultando na formação final de substâncias marrons chamadas de melanoidinas [39]. O grau de torrefação determina também um padrão de compostos voláteis que estão diretamente ligados ao aroma. É evidente, então, que a cor do café está intimamente relacionada com as propriedades organolépticas finais da bebida [11].

Dessa forma, é necessário o desenvolvimento de métodos analíticos cada vez mais rápidos, eficientes e confiáveis, com o objetivo de verificar e garantir a autenticidade das misturas de café. A discriminação entre espécies de café ou a

quantificação de misturas Robusta-Arabica já foram realizadas por ressonância magnética nuclear (RMN) [11], espectrometria de massas (MS) [22] e métodos cromatográficos [40]. Além disso, novas abordagens envolvendo metodologias mais complexas têm sido propostas recentemente, como a quantificação de misturas de café por um método baseado em DNA [41], e a discriminação de espécies baseada na análise de compostos orgânicos voláteis (VOC) usando o tempo de reação de transferência de prótons MS (PTR-ToF-MS) [42]. Métodos mais simples, rápidos e relativamente mais baratos foram desenvolvidos com base em técnicas espectroscópicas vibracionais, como Raman [35], infravermelho próximo (NIR) [38] e médio (MIR) [43].

3.2. Espectroscopia no Infravermelho

A espectroscopia no infravermelho (IV) é, atualmente, uma das técnicas analíticas mais bem difundidas no meio acadêmico [44]. Isso acontece por se tratar de uma técnica relativamente simples, rápida e, quando acoplada a acessórios de reflectância, não destrutiva. Sua descoberta é atribuída a Sir William Herschel, um astrônomo inglês que, em 1800, estava investigando o efeito de aquecimento solar dentro e fora da região do visível. Ele utilizou um prisma de vidro, transparente à radiação, para decompor a luz branca do sol em seus comprimentos de onda visíveis e, com o auxílio de um termômetro, verificou a temperatura produzida por cada cor. A mais intensa variação de temperatura foi observada quando o termômetro foi colocado logo acima da região vermelha do espectro visível, permitindo a descoberta da radiação não visível, que foi, inicialmente, denominada “radiação de raios caloríficos”, sendo mais tarde denominada Infravermelho [45].

Como ferramenta analítica, uma das principais aplicações da espectroscopia no infravermelho foi durante o período da segunda guerra mundial, sendo utilizada no setor de controle de qualidade em indústrias químicas alemãs [46]. Nesta época, a espectroscopia no infravermelho restringia-se basicamente a aplicações qualitativas, focando, principalmente, na elucidação de estruturas químicas das espécies. Porém, o desenvolvimento da microeletrônica e a popularização dos computadores possibilitaram, de maneira fácil e rápida, a aquisição de um grande número de dados e, como consequência, o desenvolvimento do tratamento de dados utilizando técnicas multivariadas/quimiométricas. Nesse sentido, é importante ressaltar que o desenvolvimento da quimiometria pode ser considerado como uma das fortes razões que permitiram a utilização da espectroscopia no infravermelho como uma ferramenta de análise em aplicações qualitativas e quantitativas em química analítica [46].

Em geral, a espectroscopia pode ser entendida como uma técnica instrumental baseada nas interações da radiação eletromagnética com a matéria. Logo, através de análise do espectro pode-se obter informações relevantes sobre a constituição, estrutura molecular e modo de interação entre moléculas [47]. O espectro eletromagnético, em geral, se estende desde comprimentos de onda curtos (raios cósmicos), até as ondas de rádio de baixa frequência. De acordo com o valor de energia da radiação eletromagnética, as transições entre os estados podem ser de vários tipos, dos quais as principais são as transições eletrônicas, vibracionais e rotacionais [47].

A região espectral no infravermelho compreende a faixa de radiação com números de onda no intervalo de aproximadamente 12800 a 4000 cm^{-1}. A região do infravermelho é usualmente dividida em infravermelho próximo (NIR – *Near Infrared Spectroscopy*), infravermelho médio (MIR – *Mid Infrared*) e infravermelho distante (FIR

- *Far Infrared*). Os limites aproximados para cada região espectral são mostrados na Tabela 2.

Tabela 2. Regiões espectrais do infravermelho (adaptada de Skoog [48])

Região	**Comprimento de Onda / nm**	**Número de Onda /cm^{-1}**
Próximo (NIR)	780 – 2500	12800 - 4000
Médio (MIR)	2500 - 50000	4000 – 200
Distante (FIR)	50000 – 1 mm	200 – 10

A energia associada a uma radiação eletromagnética é dada por [48]:

$$E = h\nu \qquad (1)$$

onde h é a constante de Planck (6,626 x 10^{-34} J s) e ν é a frequência de transição. Como a radiação eletromagnética pode ser tratada como uma onda, a energia envolvida neste tipo de transição pode ser também expressa por:

$$E = h\nu = h\frac{c}{\lambda} \qquad (2)$$

onde c é a velocidade da luz no vácuo (2,998 x 10^{8} m s^{-1}).

Nesse sentido, a espectroscopia no infravermelho trata das variações de energia molecular devido à absorção ou emissão de um fóton, que são suficientes apenas para causar alterações em modos vibracionais e rotacionais das moléculas. Assim, a radiação com números de onda entre 10000 e 100 cm^{-1}, ao ser absorvida por um composto, é transformada em energia vibracional e aquela com número de onda menor que 100 cm^{-1} convertida em energia de rotação molecular [44].

3.2.1 Espectroscopia na Região do Infravermelho médio (MIR)

Na região de 4000 a 400 cm^{-1}, denominada infravermelho médio (MIR), técnica utilizada neste trabalho, ocorrem as transições fundamentais (nas quais a molécula passa do estado fundamental para um estado excitado). Por essa razão, a espectroscopia no infravermelho médio tem ampla aplicação na caracterização de compostos orgânicos, pois cada ligação característica de um grupo funcional apresenta uma banda de vibração em uma frequência específica. Dessa forma, na análise de matrizes complexas, cujos espectros apresentam grande sobreposição espectral, a quantificação nessa região só é possível com o auxílio de técnicas quimiométricas.

Com relação à parte instrumental, os modelos mais antigos de espectrômetros no infravermelho eram do tipo dispersivo, formados basicamente por três partes: fonte de radiação infravermelha contínua, monocromador e detector. Com o passar dos anos, esses instrumentos foram gradualmente sendo substituídos pelos espectrofotômetros no infravermelho com transformada de Fourier (FT-IR). Apesar dos dois equipamentos fornecerem o mesmo tipo de espectro, o FT-IR teve alta aceitação por não possuir nenhum elemento dispersivo e detectar, simultaneamente, todos os comprimentos de onda. Nesse sentido, pode-se citar como vantagens dos espectrofotômetros FT-IR [48]: maiores velocidade de aquisição dos espectros e sensibilidade, melhor aproveitamento da potência luminosa, calibração em comprimento de onda mais exata, eliminação virtual de problemas de radiação espúria e emissão IV. Atualmente, em virtude dessas vantagens, a grande maioria dos novos equipamentos são sistemas FT-IR.

Além dos equipamentos de FT-IR, outro avanço na espectroscopia no infravermelho foi o acessório de refletância total atenuada (ATR – *attenuated total*

reflectance). A reflexão total atenuada ocorre quando um feixe de radiação está se propagando em um material com alto índice de refração e é refletido na superfície desse meio. O uso da técnica ATR baseia-se no fato de que, embora ocorra reflexão interna total na interface entre dois meios, a radiação, na realidade, penetra determinada distância no meio menos denso (amostra). Essa radiação, denominada onda evanescente, pode ser parcialmente absorvida, colocando-se uma amostra em contato com o meio mais denso (cristal, mais comumente, seleneto de zinco, germânio ou diamante). Dessa forma, a onda evanescente pode ser absorvida nos comprimentos de onda em que o material absorve e, como consequência, a radiação refletida terá sua intensidade atenuada nesses comprimentos de onda. Logo, o espectro de infravermelho é o resultado do registro dos comprimentos de onda em que a radiação foi atenuada.

Para os equipamentos que utilizam ATR, pode-se citar como vantagem a ampla gama de amostras que podem ser analisadas em diferentes condições e diferentes estados físicos. Na realidade, esses acessórios permitem a análise de sólidos e líquidos e tem grande utilidade para examinar materiais densos ou com alta absorção. Dessa forma, não são necessários processos tediosos de preparo de amostras, e podem ser realizadas tanto análises qualitativas quanto quantitativas [44].

3.2.2 Espectroscopia na Região do Infravermelho Próximo (NIR)

A Espectroscopia na região do infravermelho próximo (NIR) é a espectroscopia vibracional que emprega energia na faixa de 2,65 x 10^{-19} - 7,96 x 10^{-20} J, compreendendo a região de 750 - 2500 nm (13300 - 4000 cm^{-1}) [49]. Diferente da espectroscopia nas regiões do infravermelho distante e médio, a interação da radiação

eletromagnética na faixa espectral do NIR é dominada, principalmente, por sobretons, combinações e ressonâncias dos modos vibracionais fundamentais [50]. Dessa forma, estes modos ativos NIR são principalmente associados a modos vibracionais altamente anarmônicos de grupos funcionais moleculares contendo um átomo relativamente pesado (C, N, O ou S) ligado a um átomo de hidrogênio [50]. Além disso, vibrações de fortes ligações químicas entre átomos mais pesados, como o C=O, também podem ser detectadas pelo NIRS [50]. Dessa forma, qualquer modo vibratório ativo no NIR requer uma mudança no momento de dipolo que acompanha o movimento periódico dos átomos ligados, criando assim um dipolo oscilante. A frequência da radiação eletromagnética deve coincidir com a dos harmônicos e combinações dos modos vibracionais fundamentais. Além disso, um modo fundamental inativo pode combinar com um modo ativo para produzir um modo NIR ativo [50].

Com relação a instrumentação, um espectrofotômetro NIR pode ser construído com componentes ópticos simples [49]. Tal fato diminui, consideravelmente, o preço do equipamento, se comparado ao espectrofotômetro de infravermelho médio (MIR) [49]. Vale lembrar também que os instrumentos NIR, em geral, são mais robustos, porque suas partes ópticas não são diretamente afetadas pela umidade [49]. De maneira geral, os detectores mais empregados em instrumentos NIR são baseados em materiais fotocondutores de silício e sulfeto de chumbo, juntamente com fontes de radiação de alta potência (tungstênio ou uma lâmpada de halogênio, por exemplo), podendo fornecer uma alta relação sinal-ruído, o que compensa as menores intensidades das bandas de absorção [49].

De forma geral, a espectroscopia NIR se destaca por ser uma técnica rápida, não destrutiva, permitir análises *in situ* ou com acoplamento de fibra óptica,

extremamente versátil no que diz respeito a estado, forma e espessura da amostra [51]. Por outro lado, a região NIR possui um espectro altamente complexo, contendo muitas correlações, com informações dispersas em todo a região e fortemente influenciado por propriedades físicas da amostra, tais como a distribuição do tamanho de partículas, densidade e temperatura. Logo, a espectroscopia NIR é diretamente dependente de técnicas multivariadas de análise, com o objetivo de extrair o máximo de informação relevante de natureza química. É por essa razão que a quimiometria pode ser considerada um dos pilares sustentadores da espectroscopia NIR [50].

Para análise de café, tanto a espectroscopia NIR como a MIR já têm sua aplicação consolidada na literatura, em diferentes áreas: autenticação de espécies [51,52], qualidade do grão [54], classificação de origem [55], previsão do grau de torra [56], entre outros objetivos.

3.3. Espectrometria de Massas (MS)

A espectrometria de massas é uma ferramenta analítica que vem se destacando consideravelmente nos últimos anos, devido a uma série de vantagens, tais como: alta seletividade, baixos limites de detecção, ampla versatilidade, entre outras. Historicamente, estima-se que a técnica teve início em 1897 com o trabalho de J. J. Thomson, devido à descoberta do elétron e, consequentemente, a determinação da relação massa/carga, resultando no Prêmio Nobel de 1906 [57]. Desde então, o progresso dos métodos experimentais e, principalmente, os refinamentos nos instrumentos levaram a melhorias consideráveis nesta técnica. Nesse sentido, novas técnicas de ionização foram desenvolvidas, os analisadores aperfeiçoados e o desenvolvimento de equipamentos acoplados (especialmente

combinações de analisadores) resultaram em um aumento na resolução, sensibilidade, faixa de massas de trabalho e precisão [57].

Esta técnica se baseia, essencialmente, na ionização dos analitos e na separação dos mesmos de acordo com sua relação massa/carga (*m/z*), onde *m* é a massa atômica (em Da) e *z* a carga formal. Logo após, esses íons são detectados e processados por um sistema apropriado. Em linhas gerais, um espectrômetro de massas contém as seguintes partes, conforme ilustrado na Figura 2: entrada da amostra, que pode ser um cromatógrafo ou uma sonda de inserção direta, fonte de ionização, um (ou vários) analisadores de massas, detector e o computador para o processamento dos sinais [57].

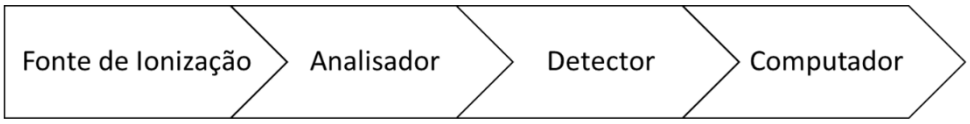

Figura 2. Componentes básicos de um espectrômetro de massas

Dessa forma, a técnica sempre executa os seguintes passos: 1) produção dos íons na fonte de ionização, 2) separação dos mesmos de acordo com sua relação m/z no analisador, 3) eventual fragmentação dos íons selecionados, 4) detecção dos íons, com a medida da abundância e conversão em sinais elétricos e 5) processamento dos sinais do detector que são transmitidos para o computador [57]. O resultado final é um espectro com a informação da abundância de um íon *versus* a relação massa carga (*m/z*).

Em linhas gerais, a ionização é uma etapa crucial para o bom desempenho da técnica. Atualmente, uma grande variedade de técnicas de ionização pode ser usada

em espectrometria de massas. Nessa etapa, os aspectos mais importantes são a energia interna transferida durante o processo de ionização e as propriedades físico-químicas do analito. Algumas técnicas de ionização são muito energéticas e causam extensa fragmentação. Outras técnicas são mais suaves e produzem apenas íons das espécies moleculares [57].

Até meados dos anos 80, a espectrometria de massas era utilizada tendo como base duas principais fontes de ionização: impacto eletrônico (EI) e ionização química (CI). A técnica EI, em linhas gerais, consiste em um filamento aquecido que emite elétrons. Estes últimos são acelerados em direção a um ânodo e colidem com as moléculas gasosas da amostra analisada, injetada na fonte [57]. Dessa forma, tal técnica funciona bem para muitas moléculas na fase gasosa, mas induz extensa fragmentação, de modo que os íons moleculares nem sempre são observados. Por outro lado, a CI consiste em produzir íons através de uma colisão da molécula de interesse com íons primários presentes na fonte. Logo, colisões de íon-molécula serão induzidas em uma parte definida da fonte. Como vantagem, a CI produz um espectro com menos fragmentação, no qual a espécie molecular é usualmente reconhecida. Dessa forma, a CI pode ser considerada complementar à EI [57].

A grande desvantagem da utilização dessas fontes é que a ionização é realizada em alto vácuo e, portanto, permitem apenas a análise de substâncias voláteis e termicamente estáveis [57]. Dessa forma, não é possível a análise de grandes moléculas orgânicas polares, dada a dificuldade de vaporização sem uma decomposição extensa [58]. Essas dificuldades foram contornadas, de forma geral, em 1989 em um trabalho publicado John Fenn propondo a fonte de ionização por *eletrospray* (ESI) [58]. Essa técnica inicialmente foi considerada como uma fonte de ionização mais direcionada à análise de proteínas. Mais tarde, seu uso foi estendido

não apenas a outros polímeros, mas também à análise de pequenas moléculas polares [57].

De uma forma geral, a ESI funciona pela aplicação de um forte campo elétrico, sob pressão atmosférica, a um líquido (solução contendo a amostra) que passa através de um tubo capilar [57]. Este campo induz uma acumulação de cargas na superfície do líquido localizada no final do capilar, que se partirá para formar gotículas altamente carregadas. Um gás injetado, a uma vazão baixa, permite que a dispersão do *spray* seja limitada no espaço. Essas gotículas passam então através de uma cortina de gás inerte aquecido, geralmente nitrogênio, ou através de um capilar aquecido para remover as últimas moléculas de solvente [57].

A partir do desenvolvimento da ESI, uma grande quantidade de novas fontes de ionização, mais simples e com menos etapas de preparo de amostras foram surgindo. Nos últimos 10 anos, muitos métodos distintos de ionização ambiente foram introduzidos, diferindo na maneira de processar as amostras e nos mecanismos de ionização [59]. Como exemplo, podemos citar: *desorption electrospray ionization mass spectrometry* (DESI-MS), *electrospray assisted laser desorption/ionization* (ELDI), *paper spray ionization* (PS), entre outros.

3.1.1 Fonte de Ionização por *Paper Spray* (PS-MS)

A fonte de ionização por *paper spray* (PS-MS) é uma técnica que combina características tanto da ESI quanto de métodos de ionização ambiente [60]. Essa técnica, relativamente recente, foi desenvolvida por Cooks *et al.*, no ano 2010 [60] e tem sido bastante difundida no meio científico, devido à sua utilidade para análises rápidas, qualitativas e quantitativas, de misturas altamente complexas, com um mínimo preparo de amostra.

A técnica se baseia na aplicação de uma alta tensão a um pedaço de papel cromatográfico na forma de um triângulo equilátero (material poroso) umedecido com um volume muito pequeno de solução contendo amostra [61] (Figura 3). O pedaço de papel triangular é conectado com a fonte de voltagem através de um clipe metálico. A amostra pode ser pré-carregada no papel, adicionada juntamente com o solvente eluente (10-50 µL, usualmente metanol) ou transferida da superfície usando o papel como absorvente. Em seguida, uma diferença de potencial é aplicada entre o papel e a entrada do espectrômetro, o que permitirá a formação de um fino *spray* contendo os analitos ionizados, devido ao grande acúmulo de cargas elétricas na ponta do papel. Neste caso, a formação das gotículas carregadas é realizada colocando a ponta do papel na frente da entrada do espectrômetro de massa, sem a utilização de qualquer gás para dessolvatação.

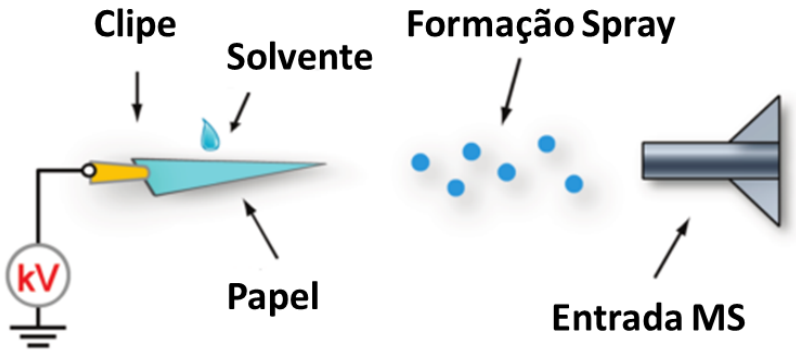

Figura 3. Esquema do Funcionamento da PS-MS (adaptada de Liu [61])

A técnica PS-MS possui uma série de vantagens, dentre as quais: ela demanda um consumo mínimo de solvente e de amostra, e não utiliza gases, podendo ser considerada uma metodologia de baixo custo. Além disso, por ser uma técnica de ionização ambiente, a PS-MS permite a análise de amostras ao ar livre, sem a

necessidade de métodos laboriosos de preparação de amostra ou pré-separação [62]. Entre suas desvantagens, podem-se citar a baixa reprodutibilidade, relacionada, principalmente, ao posicionamento correto do substrato de papel. Portanto, a otimização deste parâmetro é importante.

Conforme já mencionado, a PS-MS tem uma ampla aplicação na análise das mais variadas matrizes complexas. Diferentes trabalhos foram desenvolvidos para análise de alimentos [23, 30, 31, 63], matrizes biológicas [29,32,65–68], amostras de interesse forense [69–72], entre outros. Para a análise de café, o único trabalho encontrado na literatura foi desenvolvido por Garret et al. [23], com o objetivo de discriminar grãos verdes de café Arabica de três diferentes origens no Brasil, combinando PS-MS, Análise de Componentes Principais (PCA – *Principal Component Analysis*) e Análise Hierárquica de. Agrupamentos (HCA - *Hierarchical Cluster Analysis*). Para a quantificação de *blends* de café, destaca-se o trabalho de Garret *et al.* [22], utilizando ESI e calibração multivariada para a quantificação de *blends* de café. No entanto, este artigo é bastante limitado do ponto de vista analítico, desenvolvendo modelos a partir de poucas amostras, contendo pouca variabilidade e preparadas em poucos níveis de adulteração (em uma ampla faixa).

3.4. Fluorescência de Raios X por Reflexão Total (TXRF)

A fluorescência de raios X por reflexão total (TXRF) é uma técnica analítica multielementar, relativamente recente, bastante difundida por apresentar alta simplicidade e baixos limites de quantificação e detecção [33]. A principal vantagem está na sua capacidade de detectar quase todos os elementos da tabela periódica, nomeadamente do boro ao urânio [71]. Além disso, essa técnica é caracterizada por utilizar um volume mínimo de amostra, pela redução dos efeitos de ruído de fundo e

de matriz, pela possibilidade de operação em um ampla faixa de trabalho ($\mu g.g^{-1}$ até %), facilidade de operação, tempo de análise curto, sem consumo de gases ou água de resfriamento [72]. Vale ressaltar também a possibilidade de construção de instrumentos portáteis para analises *in situ*.

A TXRF foi utilizada pela primeira vez no final dos anos 80, mas sua popularidade atingiu o ápice no período de 1997–2005 [72]. De maneira geral, esta técnica se baseia na emissão da radiação incidente, a ângulos muito rasos, sobre uma superfície refletora, com o objetivo de se atingir reflexão total. Dessa forma, busca-se um ângulo de incidência no qual todo o feixe incidente seja refletido, minimizando a interação entre a radiação incidente e o material refletor. Assim, a fluorescência é observada com um detector posicionado muito próximo da amostra e em um ângulo reto em relação ao feixe primário. Com a eliminação dos efeitos de matriz e aumento da relação sinal-ruído, a TXRF leva a uma diminuição considerável dos limites de detecção [73]. Sendo assim, para que os raios X sejam totalmente refletidos em um plano (superfície de quartzo ou qualquer outro material), é necessário que o ângulo de incidência seja muito pequeno [74].

Na realidade, a TXRF é uma variação da espectroscopia de fluorescência de raios X por energia dispersiva (EDXRF) [73] com ângulos de incidência e de fluorescência na ordem de 0,1º e 90º, respectivamente. Por outro lado, na EDXRF esses ângulos são iguais a 45º. De maneira geral, técnicas de fluorescência de raios X (XRF) convencionais utilizam um tubo de raios X com um foco planar extenso para que o feixe primário seja cilíndrico. Dessa forma, ao atingir a amostra, o feixe de elétrons excita os átomos à fluorescência de raios X. Essa radiação de fluorescência é registrada, em geral, por um espectrômetro dispersivo de comprimento de onda. Por outro lado, a TXRF usa um tubo de raios X com foco em linha. Logo, o feixe primário

tem a forma de uma tira de papel. Dessa forma, ele atinge um suporte de vidro sob um ângulo muito pequeno e, portanto, é totalmente refletido no suporte plano [75] (Figura 5).

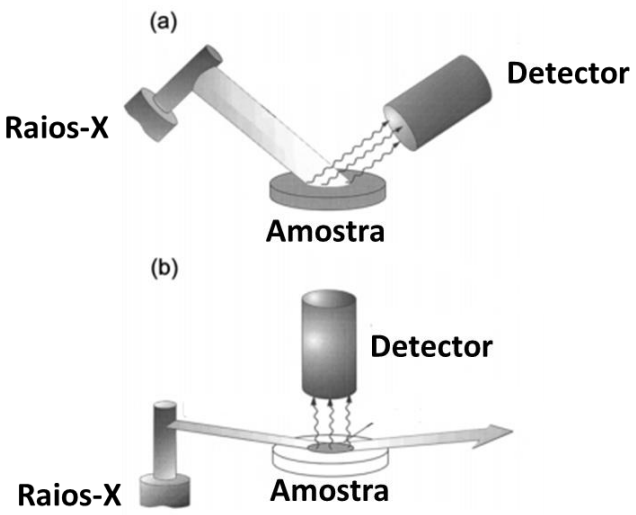

Figura 4. Comparação entre (a) XRF tradicional e (b) TXRF, (adaptada de Reinhold [75])

Com relação ao processo de preparo de amostras, a aplicação da TXRF necessita cumprir alguns requisitos básicos, tais como a formação de uma camada fina contendo amostra, geralmente definida por uma espessura crítica. Dessa forma, amostras com camadas mais finas que a espessura crítica não mostram efeitos de matriz significativos. Em geral, as etapas de preparo de amostras são necessárias para cumprir alguns objetivos, tais como: i) formação de película fina; ii) distribuição homogênea do analito e do padrão interno; iii) evitar a disseminação de solventes orgânicos; iv) enriquecimento do analito; e v) separação dos elementos interferentes [72].

Usualmente, usam-se como porta amostras discos de 30 mm de diâmetro e 3 mm de espessura, geralmente feitos de quartzo, safira ou acrílico. Nesse trabalho, foram utilizados vidros de quartzo que exigem a etapa prévia de hidrofobização (feita

por adição de 1–25 µL de solução de silicone em isopropanol e subsequente secagem por 30 min a 80° C). Essa etapa é necessária para uma correta fixação da amostra no substrato [72]. É de extrema importância que a limpeza dos porta amostras seja feita cuidadosamente, usando por exemplo, detergente neutro e HNO_3 30% v/v de e água, para evitar riscos de contaminação [72]. A Figura 6 sintetiza todos os processos de preparo dos discos de quartzo.

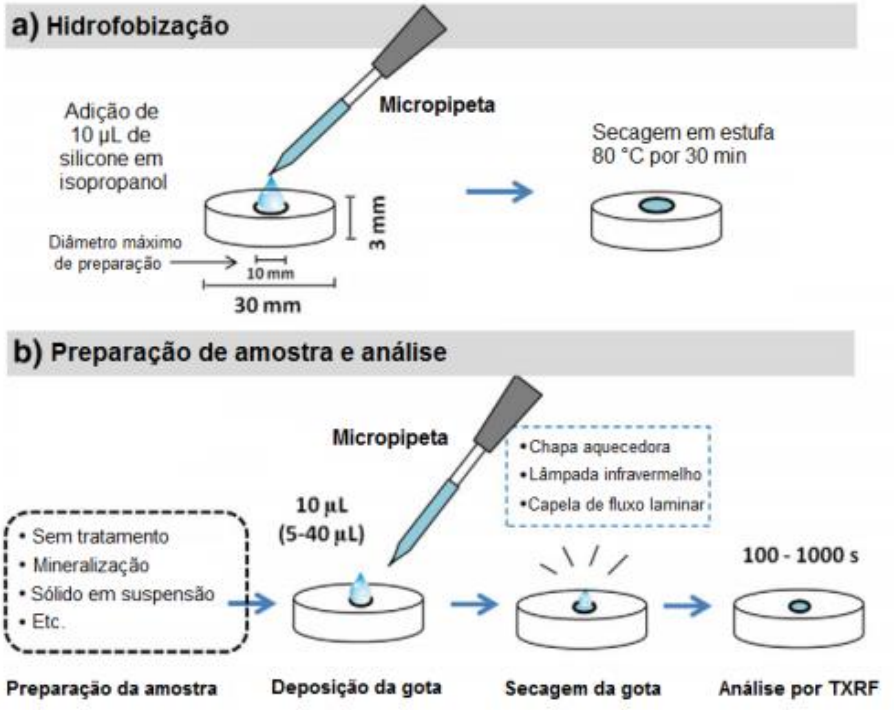

Figura 5. Etapas preparação discos de quartzo para análises por TXRF (adaptada de Rocha [76])

A TXRF é uma técnica bastante versátil, pois permite que suspensões e amostras sólidas digeridas possam ser analisadas diretamente [77]. Na literatura, diferentes trabalhos são encontrados analisando as mais diversas matrizes complexas: alimentos [78–80], amostras biológicas [81,82], amostras petroquímica [73], entre outras. Para análises de café, a grande maioria dos trabalhos envolvendo análises elementares foram feitos com objetivo de caracterização de espécies [83],

determinação de nutrientes e elementos tóxicos [84], classificação quanto à origem [85], entre outros. A utilização de TXRF para análise de café é mais rara e poucos trabalhos são encontrados [86]. Além disso, não foram encontrados artigos utilizando TXRF com o objetivo de quantificar e caracterizar *blends* de cafés Robusta e Arabica, que é o objetivo principal desta tese.

3.5. Quimiometria

A quimiometria pode ser definida como: “a disciplina que emprega métodos matemáticos e estatísticos para planejar ou selecionar experimentos de forma otimizada e para extrair o máximo de informação química dos dados” [87]. Em outras palavras, a quimiometria é vista como “a arte de extrair informações quimicamente relevantes a partir de dados produzidos em experimentos” [88], ou, em uma definição mais recente, a “ciência que relaciona medições feitas em um sistema ou processo químico com o estado do sistema através da aplicação de métodos matemáticos ou estatísticos” [89] . De forma prática, a quimiometria é uma interface entre a química e a estatística aplicada, que teve origem no final da década de 1960, quando a computação científica tornou-se acessível [90]. O desenvolvimento da quimiometria está intimamente relacionado com o desenvolvimento computacional e, como consequência, com o avanço dos instrumentos de medida analíticos que, cada vez mais, fornecem um volume maior de dados, os quais necessitam de metodologias multivariadas para sua interpretação.

O primeiro artigo científico a mencionar a palavra *Chemometrics* (quimiometria) no título foi publicado na década de 70 [91], no qual o autor afirma que a quimiometria tinha se desenvolvido a tal ponto que agora poderia ser classificada como uma área

de pesquisa funcional nas ciências químicas [92]. No Brasil, a primeira dissertação de mestrado envolvendo métodos quimiométricos foi defendida por Ieda S. Scarminio, em fevereiro de 1981, utilizando dados de espectrometria de emissão atômica para a discriminação de amostras de água mineral [93]. Atualmente, a quimiometria é uma disciplina consolidada, tanto no meio industrial como acadêmico. Os objetivos para os quais a quimiometria tem sido mais aplicada são [94]: (1) calibração multivariada, (2) relação estrutura e atividade (QSAR - *Quantitative structure–activity relationship*), (3) reconhecimento de padrões, classificação e análise discriminante e (4) modelagem multivariada e monitoramento de processos. A calibração multivariada foi rapidamente difundida a partir do final dos anos 1980, pois sua utilização possibilitou a realização de análises quantitativas sem a necessidade de resolução do sinal analítico [94], permitindo a modelagem na presença de interferentes. De forma paralela, os métodos de reconhecimento de padrões (análise exploratória de dados, classificação de amostras e resolução de curvas) e análise discriminante multivariada, tais como, PCA, SIMCA, PLS-DA, entre outros, se tornaram ferramentas bastante difundidas para o tratamento de dados químicos [94].

De maneira geral, o panorama da quimiometria para o futuro é bastante promissor. A experimentação tem, atualmente, exigido mais e mais recursos de tempo, pessoas qualificadas, instrumentação avançada, reagentes químicos, investimento financeiro, dentre outras demandas, necessitando a utilização do planejamento experimental estatístico para obter o máximo de informação possível dos dados [94]. Além disso, as técnicas instrumentais atuais têm fornecido um número cada vez maior de propriedades, variáveis, espectros, cromatogramas. Espectros que, antigamente, demandavam vários minutos para serem digitalizados com 20 variáveis, hoje são rapidamente digitalizados com milhares (ou mais) de variáveis. Por

conseguinte, no futuro provavelmente haverá ainda um maior número de variáveis sendo rapidamente gerado, em conjuntos de dados de ordem superior. Vale lembrar, também, que a quimiometria permite desenvolver métodos analíticos em consonância com os princípios da Química Verde, na medida em que estes minimizam o uso de reagentes e de energia, além da geração de resíduos.

3.5.1. Regressão por Quadrados Mínimos Parciais (PLS)

O método de regressão por quadrados mínimos parciais (PLS) foi desenvolvido por H. Wold, em 1966 [95], sendo sua aplicação, inicialmente, limitada ao campo da econometria. A partir dos anos 1980, o PLS tem ganhado importância em muitos ramos da química: química analítica, físico-quimica, química clínica e controle de processos industriais [96]. Basicamente, este método busca modelar a máxima correlação entre dois conjuntos de variáveis, contidas na matriz **X**, de variáveis independentes (espectros ou outros sinais analíticos), e no vetor ***y***, ou matriz **Y**, que contém a(s) variável(is) dependente(s) (concentrações de analitos ou propriedades a serem previstas). Nesse sentido, a única limitação sobre as dimensões dos conjuntos de dados (**X** e **y**) é que eles devem conter o mesmo número de amostras (linhas) [97]. Estudos têm demonstrando que o PLS é uma boa alternativa frente aos métodos clássicos de regressão linear múltipla e regressão por componentes principais (PCR), devido à sua maior robustez. Um método robusto significa que os parâmetros do modelo não mudam significativamente quando novas amostras provenientes da população total original são incluídas na calibração [96]. Além disso, a principal utilidade do método PLS está na sua capacidade em analisar dados com alto ruído, problemas de colinearidade e um número grande de variáveis.

A compreensão básica de qualquer modelo PLS está no fato de que o sistema estudado é modelado por um pequeno número de variáveis latentes (VL). Estas VL's são estimadas como médias ponderadas (combinações lineares) das variáveis originais [98]. Dessa forma, é necessário, primeiramente, realizar a compressão/decomposição dos dados, o que gerará as VL's, que descrevem as direções de maior variância dos dados no espaço multidimensional. O PLS faz a decomposição simultânea das variáveis independentes (matriz **X**) e das variáveis dependentes (**y** ou **Y**). Esta decomposição simultânea é a principal diferença do PLS em relação ao PCR, pois este último realiza a decomposição apenas da matriz **X** numa primeira etapa, para depois correlacionar os escores obtidos com as variáveis dependentes. Diferentes algoritmos estão disponíveis na literatura para o cálculo da regressão PLS, tais como [99] NIPALS (*Non linear Iterative Partial Least Squares)* e SIMPLS, embora na prática não haja diferença significativa entre eles em termos de desempenho.

Em linhas gerais, o PLS (quando aplicado para dados químicos) é um método de calibração inversa, ou seja, as variáveis a serem previstas (**y**, concentrações ou outras propriedades) são dependentes dos sinais analíticos (**X**, espectros, por exemplo). A decomposição das matrizes é realizada de acordo com as equações a seguir [96]:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E} \qquad (3)$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F} \qquad (4)$$

onde **T** e **U**, e **P** e **Q** são os escores e pesos (*loadings*) de **X** e **Y**, respectivamente, enquanto **E** e **F** são as matrizes correspondentes de resíduos.

O vetor de coeficientes de regressão linear (**b**) correlaciona os blocos X e Y de forma linear, para A variáveis latentes, de acordo com a equação a seguir.

$$\mathbf{u_a} = \mathbf{b_a t_a} \qquad (5)$$

onde $\mathbf{u_a}$ e $\mathbf{t_a}$ denotam a a-ésima coluna de **U** e **T**, respectivamente, enquanto que $\mathbf{b_a}$ é o vetor dos coeficientes de regressão.

O número máximo de variáveis latentes é, usualmente, igual ao pseudo-posto de **X**. Depois da escolha do número apropriado de variáveis latentes, o modelo de regressão pode ser construído e utilizado para a previsão do perfil de concentração de amostras desconhecidas, de acordo com a equação 6:

$$\hat{\mathbf{y}}_\mathbf{i} = \mathbf{x}_\mathbf{i}^{'}\mathbf{B} \qquad (6)$$

onde $\mathbf{x_i'}$ é o vetor linha (1 X J) contendo o espectro da nova amostra, $\hat{y_1}$ é o valor previsto previsto da amostra e **b** (J x 1) é o vetor de coficientes de regressão (ou **B** (JxK)) é a matriz de coeficientes de regressão, no caso da previsão de K variáveis dependentes simultaneamente).

A etapa de definição do número de variáveis latentes é de extrema importância, pois pode evitar problemas tais como subajuste ou sobreajuste do modelo. O subajuste (quando um número insuficiente de VL é utilizado) é caracterizado por excluir informação relevante do modelo. Já o sobreajuste (escolha de um número maior de VL que o adequado) é caracterizado por incluir informação irrelevante no modelo, geralmente ruído. De um modo geral, a maior preocupação na construção de modelos PLS é com a ocorrência de sobreajuste, um tipo de erro muito mais comum na literatura [100]. Dessa forma, para evitar sobre ou subajuste, uma das técnicas mais empregadas para a escolha do número de variáveis latentes é a validação cruzada. Esta técnica separa uma parte das amostras de calibração e constrói o modelo com a parte restante. Logo após, estima-se os erros de previsão para as amostras que foram separadas, utilizando diferentes números de VLs. Esse processo

é repetido para outras amostras, até que todas tenham ficado de fora. A técnica de validação cruzada tem como objetivo o cálculo do erro quadrático médio de validação cruzada (RMSECV), de acordo com a equação 7:

$$RMSECV = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \qquad (7)$$

onde $\widehat{y_i}$ é o valor previsto para a amostra e n o número de amostras utilizado na validação cruzada.

O número de variáveis latentes que corresponder ao modelo de menor RMSECV deve ser escolhido. Existem diferentes tipos de validação cruzadas, tais como: 1) *leave-one-out*, que remove uma amostra de cada vez (usada para um conjunto pequeno de amostras, normalmente até 20), 2) blocos contíguos, que separa grupos de amostras de maneira sequencial; 3) venezianas (*venetian blinds*), que separa amostras sistematicamente espaçadas; e 4) subconjuntos aleatórios, que separa, aleatoriamente, blocos de amostras diferentes [95].

3.5.2. Seleção de Variáveis

O avanço da instrumentação analítica, nas últimas décadas, tem contribuído para a geração de conjuntos de dados cada vez maiores, exigindo métodos multivariados adequados para sua análise [101,102]. Nesse contexto, os métodos de seleção de variáveis têm desempenhado um papel fundamental na análise de conjunto de dados, especialmente dados de ordem superior [103], nos quais o número de variáveis é consideravelmente maior que o de amostras [104]. Os métodos de seleção de variáveis são baseados no princípio de que a escolha de um pequeno

número de variáveis, selecionadas a partir do conjunto de dados original, permitirá uma interpretação mais simples e com maior capacidade preditiva [105]. Em linhas gerais, os métodos de seleção de variáveis são utilizados para melhorar o desempenho e a robustez do modelo, fornecendo previsões mais confiáveis. Como muitos conjuntos de dados possuem variáveis irrelevantes, ruidosas ou com pouca influência preditiva, a remoção destas pode melhorar as previsões, reduzindo simultaneamente a complexidade do modelo [106,107].

Conforme já discutido, o modelo PLS é utilizado, principalmente, para prever um vetor de resposta (ou matriz), **y**, ao mesmo tempo em que fornece a compreensão de como as amostras e variáveis se relacionam entre si por meio da estimativa de variáveis latentes, que são combinações lineares das variáveis independentes, em **X** [102]. Dessa forma, um grande número de métodos para a seleção de variáveis em modelos PLS tem sido proposto, diferenciando-se entre si na estratégia empregada para a escolha das melhores variáveis.

Os métodos de seleção de variáveis podem ser classificados como contínuos ou discretos. Em geral, a estratégia dos métodos de seleção de variáveis existentes pode ser também classificada em dois tipos, baseados: 1) na inspeção dos coeficientes de regressão, VIP *scores* ou outros vetores informativos, obtidos do modelo PLS completo, ou 2) em algoritmos de busca por sensores/variáveis nos quais o erro de predição é mínimo. Os primeiros são, usualmente, mais atraentes porque, em geral, são consideravelmente mais rápidos do que os últimos [108].

Para dados de origem espectroscópica, diversos métodos de seleção de variáveis têm sido propostos, tais como: algoritmo genético (GA) [109] (tipo 2), seleção dos preditores ordenados (OPS) [110] (híbrido entre os dois tipos), algoritmo das projeções sucessivas (SPA) [111] (tipo 1), quadrados mínimos parciais por intervalos

(iPLS) [112] (tipo 1) e seleção de variáveis preditivas com base em seus índices de importância (SVPII) [113] (tipo 1).

3.5.2.1 Seleção dos Preditores Ordenados (OPS)

O método de seleção dos preditores ordenados (OPS) foi proposto por Teófilo e coautores [110] em 2008 e tem como objetivo central automatizar a seleção de variáveis utilizando como ponto de partida vetores que trazem informações sobre os preditores/variáveis mais importantes na matriz original (**X**).

Em linhas gerais, o objetivo do método é selecionar as variáveis mais preditivas, a partir de uma investigação sistemática dos chamados vetores informativos, os quais contêm informações descritivas do modelo de calibração construído com os espectros originais inteiros (Figura 6A) [110] (primeira etapa). O método OPS pode ser descrito resumidamente em quatro etapas: (1) seleção de um vetor informativo, ou de uma combinação de mais de um tipo; (2) organização das variáveis em ordem decrescente de valor absoluto; (3) construção e avaliação dos modelos por validação cruzada e seleção de janelas de variáveis ordenadas às quais serão adicionados incrementos de novas variáveis; e (4) comparação dos modelos através de parâmetros de qualidade [114]. O algoritmo para o método é baseado em uma decomposição bidiagonal, e há sete opções de vetores informativos para inicialização: coeficientes de regressão, vetor das correlações entre **X** e **y**, vetor residual, vetor de covariância, vetor de VIP *scores*, sinal analítico líquido (NAS) e vetor da razão sinal ruído.

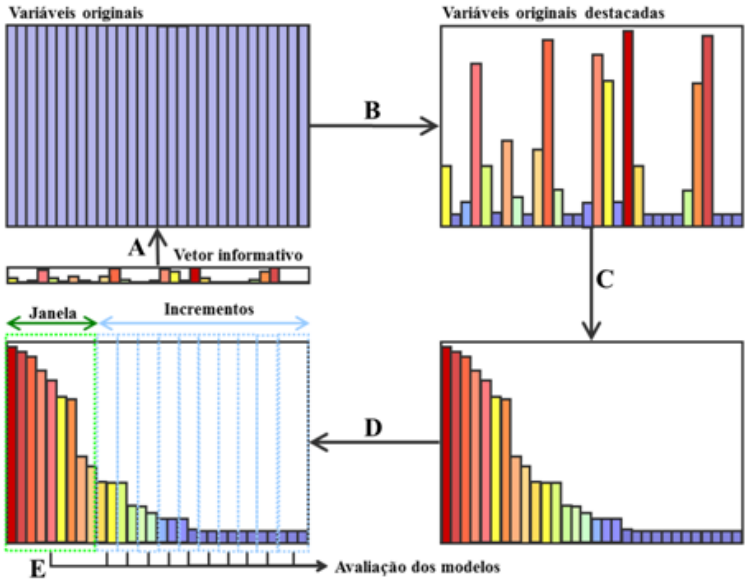

Figura 6. Representação esquemática do método de seleção de variáveis usando a seleção dos preditores ordenados (adaptada de Teófilo *et al*. [110])

Nas etapas seguintes (Figura 6B e C), as variáveis originais são organizadas (de maneira decrescente), de acordo com os correspondentes valores absolutos dos elementos do vetor informativo obtidos na etapa anterior. Quanto maior o valor absoluto, mais importante é a variável. Na próxima etapa (Figura 6D), são construídos diferentes modelos de calibração multivariada, que serão avaliados por validação cruzada do tipo *leave-N-out* (um grupo de N amostras deixadas de fora de cada vez). Um primeiro conjunto de variáveis (janela) é selecionado para construir o primeiro modelo otimizado. Em seguida, essa matriz é expandida pela adição de um número de variáveis (incremento) e um novo modelo é construído e avaliado. Dessa forma, os incrementos são adicionados até que todas, ou algum percentual das variáveis, sejam avaliadas. Na última etapa, o melhor modelo será selecionado com base em parâmetros de qualidade da validação cruzada, tais como o RMSECV e o coeficiente de correlação (R) do ajuste entre os valores previstos e de referência. O modelo com

a melhor capacidade preditiva terá o menor valor de RMSECV e o maior valor de r [110].

Durante a utilização deste método, dois números ótimos de variáveis latentes *h* são selecionados: h_{OPS}, para testar os modelos intermediários com diferentes janelas e incrementos, e h_{Mod}, o número selecionado para o modelo final (normalmente, um número próximo do usado no modelo com os espectros inteiros). Geralmente, testando-se um número de VL igual a h_{Mod}, não se geram vetores de regressão suficientemente informativos para a seleção das variáveis. Dessa forma, h_{OPS} pode ser definido como o número máximo de VL a serem testadas na construção das janelas de variáveis intermediárias a serem otimizadas pelo método OPS [110].

O OPS apresenta as vantagens de ser computacionalmente eficiente, quando comparado com outros algoritmos de seleção de variáveis; totalmente automatizado; poder ser adaptado para tratar conjuntos de dados *multi-way* ou usado em análise discriminante [110]; e combinar os dois princípios de seleção de variáveis, (1) baseado na inspeção das informações obtidas do modelo construído com os espectros inteiros, e também (2) na busca empírica por variáveis que minimizem os erros de previsão (janelas de busca) [108]. A principal vantagem do método OPS é que ele faz uma combinação entre os dois diferentes grupos de métodos de seleção de variáveis, utilizando os coeficientes de regressão ou outro vetor informativo de seleção de variáveis e complementando com a busca pelos sensores/variáveis mais preditivos.

3.5.2.2 Algoritmo Genético (GA)

O algoritmo genético (GA) é classificado como um método de inteligência artificial, tendo sido bastante empregado na análise de dados espectroscópicos por modelagem multivariada. O algoritmo foi proposto por John H. Holland [115], um pesquisador da Universidade de Michigan, nos 1960, com o objetivo de otimizar sistemas complexos [116]. Em linhas gerais, o GA tem como princípio básico simular um processo de evolução natural, baseado na Teoria da Evolução Darwiniana [117]. Na área de quimiometria, a primeira aplicação [116] foi publicada no início dos anos 1990, por Lucasius e Kateman [118], usando GA para seleção de comprimentos de onda em espectros no ultravioleta.

Em linhas gerais, o GA consiste de cinco etapas: 1) codificação das variáveis, 2) criação da população inicial, 3) avaliação da resposta, 4) cruzamento e 5) mutação [116]. Na etapa inicial, as variáveis são codificadas em número binários (assimilando-as a cromossomos biológicos), sendo representadas por uma sequência de zeros e um (0 implica que o intervalo codificado é descartado, enquanto 1 significa a inclusão das correspondentes informações no modelo PLS) [119]. Dessa forma, através de um gerador aleatório (garantia da ausência de influência tendenciosa), formula-se uma população inicial de cromossomos, formada pela combinação de todos os indivíduos testados.

A etapa seguinte (avaliação da resposta) é considerada a mais importante no procedimento do algoritmo genético [116]. A resposta é uma característica que indica a habilidade que um indivíduo possui para sobreviver (ser selecionado). Em outras palavras, seria sua habilidade de produzir a melhor resposta (maior capacidade de previsão). Esta resposta (aptidão) é obtida calculando um modelo de regressão para cada indivíduo e estimando o valor do erro para um conjunto de amostras externas ou

por validação cruzada (RMSECV). Indivíduos (variáveis) com valores altos de RMSECV são descartados, reduzindo consideravelmente o tamanho da população.

Na próxima etapa, os melhores cromossomos (aqueles que levam a menores erros) sobrevivem, podendo sofrer mutação e se recombinar para produzir descendência [119]. A partir da escolha de diferentes indivíduos, uma nova população é formada pelo cruzamento aleatório entre os cromossomos. Dessa forma, os descendentes guardam informações de seus progenitores, através do material genético proveniente do cruzamento [116]. Nesta altura, haverá uma forte tendência para o prevalecimento de características dominantes (variáveis mais preditivas), o que leva a convergência do algoritmo (situação ótima).

A etapa final (mutação) consiste, basicamente, em uma perturbação das informações contidas dentro dos genes. A etapa de mutação é necessária, pois ela pode superar alguns problemas ocorridos durante o processo de seleção, notadamente, buscando evitar a seleção de mínimos locais. Por exemplo, se uma variável (inicial) não for selecionada em um cromossomo original, ela nunca será selecionada nas gerações futuras, caso a mutação não exista [120]. De forma prática, a mutação ocorre a uma taxa baixa, sendo realizada aleatoriamente pela troca de um por zero e vice-versa em alguns indivíduos/variáveis [116]. Após uma série de *n* gerações que tenham sido avaliadas, o subconjunto que apresentar melhores parâmetros estatísticos é finalmente utilizado para construção do modelo [119].

O GA apresenta as seguintes vantagens [116]: não requer informações sobre o gradiente da superfície de resposta; descontinuidades da superfície de resposta não afetam o desempenho da otimização; a presença de mínimos locais não reduz a sua eficiência; realiza buscas simultâneas em várias regiões do espaço amostral; e pode ser aplicável a uma ampla gama de otimizações.

Para finalizar, a Figura 7 mostra um esquema do funcionamento geral do GA.

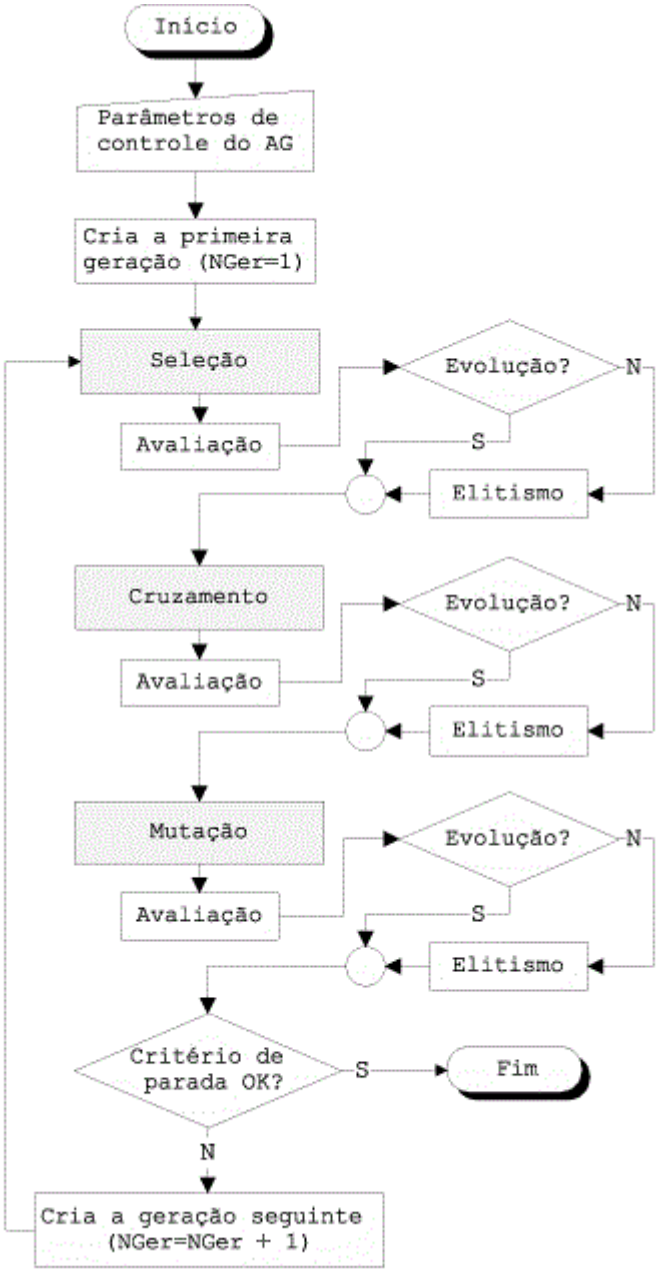

Figura 7. Esquema de funcionamento do Algoritmo Genético (adaptada de Ferreira [121])

3.5.2.3 Algoritmo das Projeções Sucessivas (SPA)

O Algoritmo das Projeções Sucessivas (SPA) foi proposto por Araújo e colaboradores [111], em 2001, com o objetivo de minimizar os problemas de colinearidade em modelos de regressão linear múltipla (MLR). Em linhas gerais, o SPA emprega operações simples de projeção em um espaço vetorial para obter subconjuntos de variáveis com colinearidade mínima (mínimo de informação redundante). O princípio básico é que cada nova variável selecionada é a única entre todas as demais variáveis, que tem o valor máximo de projeção sobre o subespaço ortogonal da variável selecionada anterior [120].

Este método foi, inicialmente, proposto para a construção de modelos de calibração multivariada sendo, subsequentemente, ampliado para resolver problemas de classificação [122].

O SPA é composto, basicamente, por três fases. A primeira fase consiste em operações de projeção envolvendo as colunas de matriz **X**cal (matriz de calibração, com as dimensões $N_{cal} \times K$), geralmente centrada na média das colunas, que geram K cadeias de M variáveis cada ($M=\min(N_{cal}-1, K)$) [123]. De uma forma geral, o algoritmo para resolver este método é descrito a seguir [124]:

Passo 1: (inicialização) faça:

De k=1 até k

$z^1 = x_{kj}$

$x^1_j = x_j$, j=1,..., K

$\mathbf{SEL}(1,k) = k$

i= 1

Passo 2: Calcular a matriz de projeção P_i no subespaço ortogonal a z_i:

$$P_i = I - \frac{z^i (z^i)'}{(z^i)' z^i}$$

onde $\boldsymbol{I}$ é a matriz identidade de dimensões n x n.

Passo 3: Calcular os vetores projetados x_{ji}+1 a partir de:

$$x_j^{i+j} = P^i x_j^i$$

para todos os j = 1,...,K.

Passo 4: Determinar o índice de j* do vetor de maior projeção e armazená-lo na posição (i+1, k) na matriz **SEL**:

$$j^* = arg(max\left\| x_j^{(i+j)} \right\|) \text{ e } SEL\left(i+1,k\right) = j^*$$

Passo 5: Fazer z_i+1 = x_{j^*i+1} (vetor que define a próxima operação de projeção)

Passo 6: Fazer i = i +1. Se i < M, retorne para o Passo 2.

A etapa seguinte consiste em avaliar subconjuntos de variáveis extraídos das cadeias e armazená-las na matriz **SEL** [123]. De forma geral, são utilizadas as variáveis com índices {**SEL** (1, k), **SE**L (2, k), ..., **SEL** (m, k)} para construir um modelo MLR. Como m varia de um a M e k varia de um a K, um total de M x K subconjuntos de variáveis são testados. Logo após, o modelo é aplicado para o conjunto de validação e o RMSE (m,k) é calculado. Na realidade, o cálculo do RMSE pode ser realizado de duas formas, utilizando validação cruzada ou um conjunto de teste independente.

A terceira fase do algoritmo consiste em um procedimento de redução de variáveis com a finalidade de eliminar variáveis não informativas. Para este efeito, as variáveis selecionadas na fase 2 são classificadas de acordo com um índice de relevância e um gráfico de RMSE *versus* número de variáveis incluídas no modelo é gerado. Considerando $RMSE_{min}$ o menor valor de RMSE observado, a solução ideal

é tomada como o menor número de variáveis tais que o RMSE não é significativamente maior do que o $RMSE_{min}$, de acordo com um teste t [123].

3.5.2.4 Seleção de Variáveis Preditivas com base em seus Índices de Importância (SVPII)

Este método foi proposto por Anzanello *et al.* [113,125] e, resumidamente, pode ser dividido em quatro passos: (1) a matriz de dados **X** (usualmente variáveis espectrais) e **Y** (matriz ou vetor de respostas dependentes) são divididas em dois conjuntos: treinamento e teste; (2) a regressão PLS é aplicada ao conjunto de treinamento, gerando quatro índices de importância das variáveis; (3) o modelo é aplicado ao conjunto de treinamento, sua exatidão é avaliada e variáveis irrelevantes são eliminadas iterativamente e (4) os melhores subconjuntos de variáveis são selecionados, utilizando duas abordagens diferentes, e o modelo é validado por um conjunto teste.

Na primeira etapa do algoritmo, o conjunto de dados é dividido, aleatoriamente, em conjuntos de treinamento/calibração e teste/validação. O conjunto de treinamento é utilizado para selecionar as variáveis mais importantes, enquanto o conjunto de teste é usado para avaliar a exatidão do modelo PLS. Em geral, 75% das amostras são selecionadas para o conjunto de treinamento e 25% para o conjunto de teste. Os dados são normalizados antes da construção dos modelos para evitar efeitos de escala na estimativa de parâmetros [125].

Na segunda etapa, a regressão PLS é aplicada ao conjunto treinamento, gerando os índices de importância de cada variável *k*. Nesse contexto, os parâmetros de interesse na avaliação incluem: os coeficientes de regressão ($\boldsymbol{b}_k$), os pesos das

variáveis nas componentes principais (*loadings*) ($\boldsymbol{p}_{ka}$), os pesos da regressão PLS (*weights*) ($\boldsymbol{w}_{ka}$) e a porcentagem de variância explicada em **Y** pela componente *a*, (R^2_{Ya}). Em geral, parâmetros com valores elevados em módulo estão associados com variáveis importantes para explicar a variabilidade em **Y**. Dessa forma, quatro índices de importância são obtidos para cada variável, com o objetivo de guiar a remoção de variáveis com pouca representatividade. O índice de importância de cada variável *k* é denotado por $I_{k,}$ onde k = 1, …, K, sendo K o número total de variáveis. Quanto mais alto o valor de I_k, mais importante é a variável para predição. Os índices de importância são por definição não-negativos, já que neles os parâmetros do modelo PLS são elevados ao quadrado para evitar que seus altos valores positivos e negativos se cancelem [125].

O primeiro índice testado, $I^{(I)}_k$, relaciona os valores dos pesos da regressão PLS, $\boldsymbol{w}_{ka}$, e a fração de variância explicada em **Y**, R^2_{Ya}, pela componente a = 1, …, A. Este índice é baseado nos valores de VIP scores. O segundo índice, $I^{(II)}_k$, avalia a importância de cada variável baseado na magnitude dos coeficientes de regressão, $\boldsymbol{b}_k$. Paralelamente, o terceiro índice, $I^{(III)}_k$, depende da magnitude do quadrado do valor $\boldsymbol{p}$ dos pesos das variáveis nas componentes principais usadas no modelo. O quarto, e último, índice testado, $I^{(iv)}_k$, integra os coeficientes de regressão PLS, os pesos da regressão PLS, e a fração de variância explicada em **Y**. O objetivo deste índice é destacar as variáveis que afetam diretamente a variável de resposta, ou seja, com alto $\boldsymbol{b}_k$, cujos pesos vêm de componentes que explicam uma quantidade substancial de variação em **Y**.

Na terceira etapa do método, um modelo PLS é construído utilizando K variáveis. O desempenho do modelo é avaliado com base nos valores de RMSE. Em seguida, a variável com menor valor de I_k é removida do conjunto treinamento,

gerando um novo modelo PLS, de K-1 variáveis, recalculando, em seguida, o RMSE. Esse processo iterativo é repetido até que reste apenas uma variável. Este procedimento é replicado para todos os índices de importância [125].

Para a última etapa, um gráfico contendo os valores de RMSE (Figura 8), calculados após cada remoção de variáveis, *versus* a porcentagem de remoção de variáveis é construído. De maneira geral, duas abordagens diferentes são propostas para selecionar o melhor subconjunto de variáveis: (1) RMSE Mínimo (MR) e Distância Euclidiana Mínima (MED). A primeira proposição seleciona o subconjunto que produz o RMSE mínimo. Em casos nos quais existam RMSEs mínimos idênticos, o modelo com o menor percentual de variáveis retidas é selecionado. Na segunda proposição, MED, calcula-se a Distância Euclidiana entre cada ponto descrito por coordenadas (RMSE, porcentagem de variáveis retidas) para um ponto ideal hipotético. As coordenadas do ponto ideal são definidas pelo usuário, recomendando-se estabelecer valores próximos de zero tanto para RMSE quanto para a porcentagem, uma vez que valores pequenos denotam alta exatidão e parcimônia em termos de retenção de variáveis, respectivamente. Logo, o ponto com a menor distância Euclidiana para o ponto ideal é selecionado. Para finalizar, um modelo PLS é construído com o subconjunto de variáveis com maior capacidade preditiva.

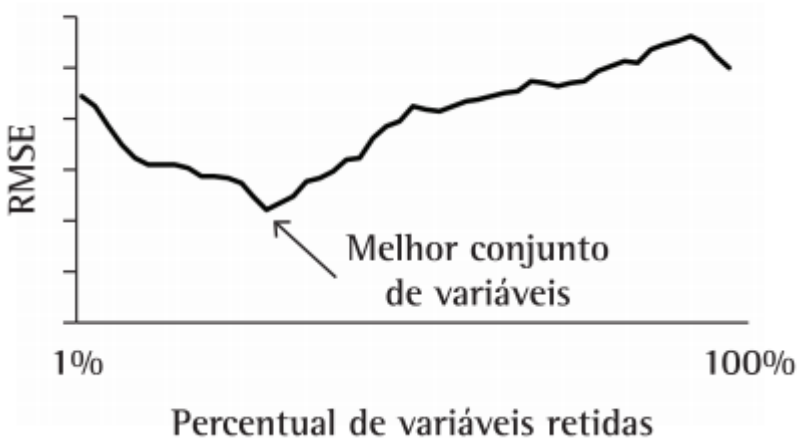

Figura 8. Perfil hipotético de RMSE à medida que as variáveis de processo são eliminadas do conjunto de treinamento/calibração. (adaptada de Zimmer e Anzanello [113])

3.5.2.5 Quadrados Mínimos Parciais por Intervalos (*i*PLS)

O método iPLS [112], o mais simples dentre os testados, baseia-se na divisão dos espectros originais em várias regiões (intervalos) contínuos, com o objetivo de encontrar uma região espectral que produza melhores previsões do que os espectros completos. Em geral, busca-se otimizar o poder preditivo dos modelos de regressão PLS e, concomitantemente, auxiliar na interpretação. Dessa forma, sua tarefa é fornecer uma visão geral da informação relevante em diferentes subdivisões espectrais, focando, assim, nas mais importantes regiões espectrais e removendo interferências de outras regiões.

De maneira prática, uma regressão PLS é executada em cada sub-intervalo (o número total é determinado pelo usuário) equidistante de todo o conjunto de variáveis na matriz de dados. Desta forma, é possível avaliar e identificar os subconjuntos de variáveis que apresentam informações mais relevantes, ou seja, as regiões espectrais

com contribuições menores e/ou responsáveis por informações ruidosas devem ser removidas. A comparação entre os modelos é baseada, principalmente, no parâmetro de validação RMSECV, mas outros parâmetros como R^2 (coeficiente de determinação) também são avaliados para garantir um modelo abrangente. Amostras anômalas (*outliers*) devem ser removidas antes da aplicação do método [112].

3.5.3. Fusão de Dados

Atualmente, com o amplo avanço das técnicas analíticas (que fornecem uma quantidade cada vez maior de informação em intervalos cada vez menores de tempo) e, como consequência, com o aumento da capacidade de processamento de dados (desenvolvimento acelerado dos microprocessadores), tornou-se tarefa comum a obtenção de um grande volume de dados [126], demandando novas ferramentas para melhor interpretação química desses resultados, e permitindo então o surgimento de modelos de fusão de dados.

A fusão de dados é, em linhas gerais, uma estratégia que tem como objetivo fundir (ou integrar) blocos de dados provenientes de diferentes técnicas analíticas em um único modelo [127], tais como: espectroscopias no infravermelho (IV) médio e próximo, Raman, no UV-Vis, espectrofluorimetria, espectrometria de massas, sensores de diversos tipos, cromatografia (líquida e gasosa), análise sensorial, entre outras. Além disso, variáveis discretas (variáveis físico-químicas provenientes de análises via úmida) também podem ser combinadas. Dessa forma, pode-se obter um conhecimento mais preciso (e completo) sobre uma amostra, construir modelos com melhores parâmetros estatísticos (classificações com uma menor taxa de erro e previsões com menores erros), se comparados com modelos que utilizam apenas uma

técnica instrumental. Além disso, como vantagens, essa estratégia permite identificar quais fontes de variações são comuns entre os blocos e quais são únicas para as matrizes individuais, permitindo uma interpretação química mais aprofundada [128].

É importante ressaltar que, para uma correta utilização dessa estratégia, é importante conhecer bem a matriz analisada, bem como o problema a ser solucionado, de forma que haja uma correta seleção das técnicas analíticas mais adequadas a serem fundidas [34]. Evidentemente, a fusão de dados é útil quando a informação complementar é modelada, uma vez que seu objetivo principal é aumentar a sinergia entre as técnicas fundidas, mesclando informações complementares [129].

Em linhas gerais, a fusão de dados pode ocorrer em três níveis (Figura 9): baixo, médio e alto. No nível baixo (ou nível de medição), os sinais (variáveis obtidas pelas diferentes fontes) são diretamente concatenados (após as etapas de pré-processamento) em uma única matriz, que tem tantas linhas quanto as amostras analisadas e tantas colunas quanto os sinais (variáveis) medidos pelos diferentes instrumentos [34]. Dessa forma, essa informação é então utilizada para calcular um modelo único que fornece a classificação ou previsão final. Apesar ser operacionalmente mais simples, neste nível a fusão ocorre diretamente nas matrizes de dados originais, ou seja, o arranjo final resultante normalmente conterá um número muito alto de variáveis, de modo que a principal desvantagem dessa estratégia é que o aumento na informação obtida pela adição de um ou mais blocos de dados para descrever a amostra pode não compensar a quantidade de variância irrelevante trazida pela adição dos mesmos blocos [128]. Nestes casos, a aplicação de métodos de seleção de variáveis pode contribuir para a construção de modelos mais simples, eliminando variáveis ruidosas ou irrelevantes.

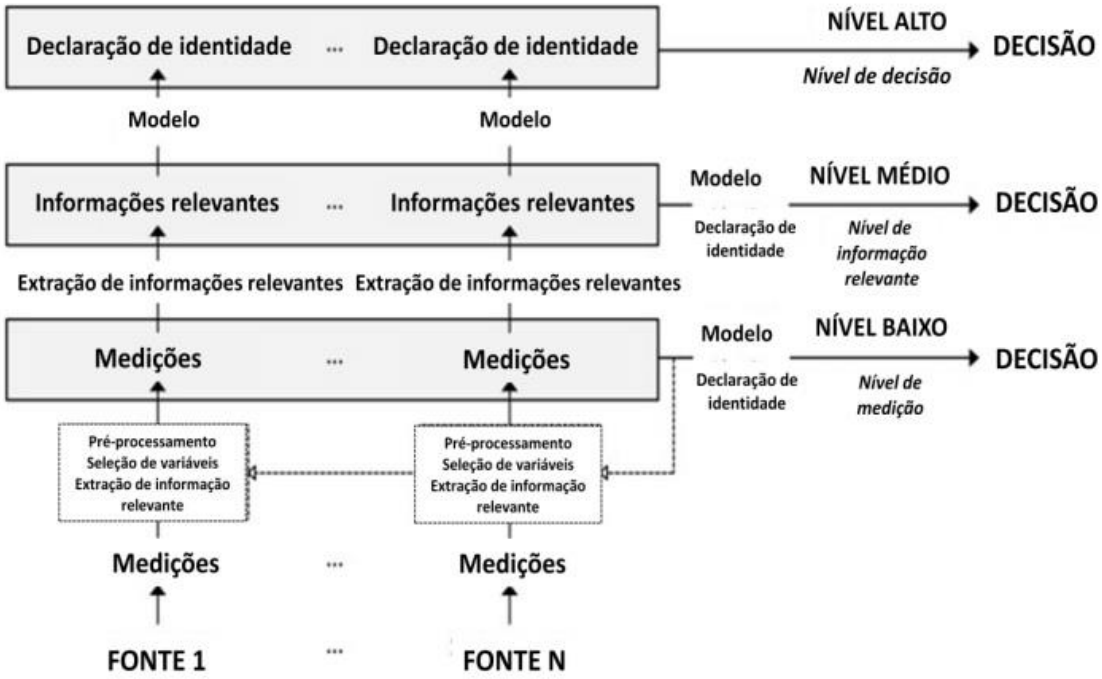

Figura 9. Representação dos três níveis fusão de dados (adaptada de Nunes [130])

Por outro lado, a fusão de dados nível médio (ou nível de informação relevante) primeiro extrai, separadamente, algumas informações relevantes de cada fonte de dados e, em seguida, concatena estas informações em um único arranjo (ou matriz), que pode ser utilizado para classificação ou calibração multivariada [34]. A abordagem mais comum na literatura é fundir o número de variáveis latentes (ou componentes principais) obtidos, isoladamente, dos sinais de cada instrumento. Dessa forma, os escores mais significativos de modelos PCA, PLS ou PLS-DA são concatenados para descrever a variação significativa nos diferentes blocos. O principal desafio, portanto, é encontrar a combinação ideal de informações extraídas que forneça o melhor modelo final [34]. Dessa forma, como esta técnica extrai apenas a variação relevante nas diferentes matrizes de dados, a estratégia de nível médio geralmente é mais eficaz do que a de baixo nível e não sofre das mesmas desvantagens [128]. Por outro lado,

a principal desvantagem é exigir um estágio preliminar de construção de modelos, no qual os recursos são extraídos dos diferentes blocos diferentes.

Já na fusão de dados de nível alto, também denominado nível de decisão, modelos de classificação (ou previsão) são construídos separadamente (para cada fonte de dados) e os resultados finais de cada modelo individual são integrados em uma única resposta final [128]. O principal desafio é construir modelos que funcionem adequadamente para cada bloco de modo que sua combinação tenha um desempenho consideravelmente melhor que os modelos individuais [34]. De uma maneira geral, a fusão de dados de alto nível, é mais comumente utilizada em problemas de classificação, nos quais a inferência Bayesiana (baseada na estimativa de probabilidade) é a mais utilizada [34].

De forma prática, a escolha do melhor nível de fusão é definida após a geração dos dados. Essa escolha depende de vários fatores, tais como: técnicas instrumentais selecionadas, volume de dados gerados, entre outros. Os resultados, após a aplicação da fusão, devem ser comparados com os dados isolados. Obviamente, se uma única técnica isoladamente fornecer resultados melhores, não há necessidade de aplicação de fusão de dados. Vale lembrar que o objetivo principal é aumentar a sinergia entre os dados fundidos, ou seja, a utilização de técnicas com informações complementares, para obtenção de modelos com melhores classificações ou previsões [34] .

Em relação às análises de café, alguns artigos recentes foram publicados na literatura combinando modelos de fusão de dados com métodos quimiométricos. Reis e colaboradores [17] utilizaram espectroscopia na região do infravermelho médio em dois modos de aquisição distintos, reflectância total atenuada e reflectância difusa, para detecção simultânea de múltiplos adulterantes em café torrado. Modelos de

classificação, utilizando PLS-DA, foram construídos e a metodologia de fusão de dados diminuiu, consideravelmente, o percentual de amostras erroneamente classificadas. Além disso, Dong e colaboradores [131] utilizaram sensores do tipo nariz e língua eletrônicos com o objetivo de caracterizar e classificar sete diferentes cultivares de café Robusta chinês em diferentes graus de torra. Os dados combinados tiveram desempenho muito melhor que a abordagem isolada na medição dos parâmetros de qualidade dos grãos de café, utilizando regressão PLS. Dankowska e colaboradores [132] combinaram espectrofluorimetria e espectroscopia na região do UV-Vis para quantificação de *blends* de café Arabica e Robusta, utilizando fusão nos níveis médio e baixo. Neste caso, a regressão em componentes principais foi usada para reduzir a multidimensionalidade dos dados. Se comparada aos modelos construídos com as duas técnicas individuais, a fusão de dados mostrou melhor desempenho analítico, indicando que há complementariedade entre as diferentes respostas instrumentais. No entanto, este artigo possui uma série de aspectos controversos, os quais necessitam ser questionados. Tal discussão será feita na seção Resultados & Discussão.

4. Materiais e Métodos

4.1. Preparo das Amostras

A matéria prima utilizada nesse trabalho, grãos verdes secos e descascados, foi obtida diretamente com proprietários rurais nos estados de Minas Gerais e Espírito Santo. As amostras de café Arabica (N=30) foram provenientes da região da Zona da Mata, Minas Gerais, nas seguintes cidades: Manhuaçu, Simonésia, São João do Manhuaçu e Matipó. As amostras foram adquiridas com 10 proprietários rurais diferentes, variando o ano da safra e a qualidade dos grãos (número de defeitos). O mesmo procedimento foi realizado para as amostras de café robusta (N=10), provenientes do Espírito Santo, obtidas nas seguintes cidades: Venda Nova do Imigrante, Domingos Martins e Cachoeiro do Itapemirim. Todas as amostras foram acondicionadas em ambiente refrigerado, a 2°C.

O processo de torrefação foi realizado com um torrador elétrico de bancada, marca Hottop, modelo KN8828p. As amostras (aproximadamente 100 g) foram torradas separadamente e três temperaturas diferentes foram utilizadas, 185, 195, 205ºC (temperatura inicial: 140ºC, taxa de aumento: 7ºC/min) caracterizando as torras, respectivamente, como leve, média e escura (Figura 10). Além disso, foi feito um acompanhamento de coloração e aroma, durante o processo de torra, de forma a deixar o processo mais homogêneo possível. Após a torra, as amostras foram moídas, à temperatura ambiente, em um moedor de café caseiro (Cadence modelo MDR301-127) e peneiradas (40 mesh). Na sequência, cafés torrados de diferentes origens de Arábica (N = 30) foram misturados, assim como os cafés Robusta (N = 10). Dessa forma, dois grandes conjuntos de amostras, Arábica e Robusta, foram obtidos e utilizados para preparar os *blends*.

Figura 10. Amostras de café Arabica em diferentes graus de torrefação: leve (esquerda), média (centro) e escura (direita)

4.1.1 Formulações dos blends

Os *blends* foram preparados (10g) em diferentes proporções de café Robusta em café Arabica (0,0-33,0%, variado de 1,0 em 1,0%). *Blends* de concentrações 3,0, 15,0, e 27,0% foram preparados em triplicata, de forma a avaliar a precisão do método. Os *blends* foram acondicionados em sacos plásticos (do tipo ziploc) e armazenados em geladeira (5-7ºC). Esse material foi utilizado para as leituras no MIR e NIR. Dessa forma, 40 *blends* foram formulados para cada nível de torra, totalizando 120 amostras.

4.2. Análises por Espectroscopia MIR/NIR

Os espectros MIR foram obtidos em um espectrômetro IRAffinity-1S (Shimadzu, Kyoto, Japão) com um acessório fornecido pelo próprio fabricante para análise por ATR (8200H/8200HA), equipado com um prisma côncavo de ZnSe. As

amostras em pó foram utilizadas para essa leitura. A faixa investigada foi de 4000 a 800 cm^{-1}, resolução de 1 cm^{-1}. Os espectros foram obtidos através do *software* IR Solution (Shimadzu, Kyoto, Japão), armazenando a informação como absorbância (A). Para cada amostra, um total de 64 varreduras foram realizadas e a média foi armazenada. As leituras das amostras foram realizadas de maneira aleatória em relação aos teores, de forma a minimizar a presença de variação sistemática no modelo.

Para as leituras no NIR, o equipamento portátil RED-Wave-NIRX-SR (StellarNet, Tampa, EUA), no modo reflectância, foi utilizado. A faixa estudada foi de 900-2300 nm, resolução 1 nm. Assim como na análise no MIR, as leituras foram realizadas de maneira aleatória em relação aos teores, com um total de 64 varreduras por amostra, armazenando a média.

4.3. Preparação dos extratos de café

A extração das amostras de café foi realizada segundo os seguintes passos: 1,00 g de cada *blend* formulado (item 4.1.1) foi pesado e colocado em um tubo do tipo Falcon com capacidade de 15 mL. Logo após, foram adicionados 10 mL de água deionizada a cada tubo, levando ao banho-maria durante 20 min na temperatura controlada de 90º C. Em seguida, a mistura foi centrifugada por 10 minutos a 3000 rpm, seguida de uma filtração com papel quantitativo. A proposta da extração foi a escolhida pela simplicidade e pela tentativa de reproduzir a forma em que, geralmente, o café é coado. A solução foi então armazenada em um tubo Falcon e mantida à temperatura de -20 ºC para análise futura por PS-MS e TXRF.

4.4. Análises por Espectrometria de Massas

Os espectros PS-MS foram obtidos em um espectrômetro de massas Thermo Fisher LCQ FLEET (San Jose, EUA), na faixa m / z entre 100 e 500, em duplicata e em ordem aleatória. De cada extrato (item 4.3), uma alíquota de 2,0 µL foi carregada em um pedaço de papel cromatográfico 1 CHR, fabricado pela Whatman (Little Chalfont, Buckinghamshire, Reino Unido), cortado na forma de um triângulo equilátero (1,5 cm) e mantido por um clipe de cobre conectado à fonte de tensão do equipamento. Na sequência, aplicaram-se 40,0 µL de metanol de qualidade para HPLC (J. T. Baker Chemicals, Center Valley, EUA) ao papel triangular e a fonte de tensão foi ligada. Os dados foram coletados com o *software* Thermo Fisher Scientific Xcalibur 2.2 SP1. As condições experimentais otimizadas foram as seguintes: tensão de pulverização de papel: 5,5 kV; tensão capilar: 40 V; distância da ponta do papel à entrada do espectrômetro: 0,8 cm. Todos os espectros foram obtidos no modo positivo. Testes preliminares no modo negativo forneceram espectros com baixa estabilidade de sinal e baixa reprodutibilidade.

4.5 Análises por TXRF

As medidas por TXRF foram feitas utilizando o equipamento S2 PICOFOX™ (Bruker Nano GmbH, Karlsruhe, Alemanha), equipado com um tubo de molibdênio (Figura 13). Para o preparo das amostras, os seguintes passos foram realizados: 250 µL do extrato (item 4.3) foram misturados com 200 µL de água e 50 µL do padrão interno (solução de gálio 10 mgL^{-1}). Os espectros foram obtidos através do *software* SPECTRA 7.5.3 (Bruker, Karlsruhe, Alemanha) e os seguintes elementos foram

monitorados: P, S, Cl, K, Ca, Ti, Mn, Fe, Ni, Cu, Zn, Br, Rb, Sr. As análises foram realizadas em duplicata e de maneira aleatória com relação aos teores.

4.6 Construção dos Modelos Calibração Multivariada

Os espectros foram exportados na extensão .xls e importados pelo *sofware* Matlab 7.13 (Math Works, Natick, EUA). Os modelos foram desenvolvidos no pacote PLS Toolbox, versão 6.7.1 (Eigenvector Technologies, Manson, WA, EUA). As variáveis independentes (X) foram compostas por três matrizes de dados (X_{leve}, $X_{médio}$ e X_{forte}). As linhas da matriz **X** correspondem às amostras (N=40) e as colunas correspondem às variáveis (cada técnica instrumental gerou um número de variáveis diferentes). O vetor contendo a variável independente (**y**) foi construído com a informação da porcentagem (%) de café Robusta nos *blends*. O vetor **y** possui um número de linhas igual ao número de amostras na matriz **X**.

Durante o processo de construção dos modelos de calibração, as variáveis provenientes das técnicas MIR/NIR e PS-MS foram centradas na média em todos os cálculos. Além disso, diferentes pré-processamentos foram testados nas matrizes oriundas das técnicas MIR e NIR com o objetivo de encontrar o modelo com melhor exatidão, tais como [133]: (1) primeira derivada; (2) correção do espalhamento multiplicativo (MSC); (3) variação normal padrão (SNV). Para os dados de TXRF, as matrizes foram autoescaladas.

Para todos os modelos, diferentes métodos de seleção de variáveis foram otimizados e aplicados, tais como: OPS, GA, SPA, iPLS e SVPII. Dessa forma, um modelo PLS individual foi construído com cada tipo de conjunto de dados espectrais, ATR-FTIR, NIR, PS-MS e TXRF. Conforme mencionado, modelos de fusão de dados

foram construídos em dois níveis: baixo e médio. Para o modelo de fusão de dados de baixo nível, as matrizes espectrais foram pré-processadas separadamente, concatenadas e autoescaladas. Então, esse modelo foi otimizado por seleção de variáveis. Para o nível médio, os modelos PLS foram construídos separadamente para cada técnica analítica e otimizados por seleção de variáveis. Em seguida, os escores correspondentes ao número de variáveis latentes (nVL) significativo de cada modelo foram extraídos, concatenados e autoescalados. Para métodos de seleção de variáveis, diferentes condições foram testadas e aquelas que forneceram o menor RMSECV foram selecionadas.

Para a divisão do conjunto de dados em calibração e validação, as amostras foram separadas em 2/3 para o conjunto calibração e 1/3 para o conjunto de validação. Dessa forma, duas técnicas foram aplicadas: o algoritmo de Kennard-Stone [134] e o planejamento experimental, no qual as amostras foram separadas pela ordem CVC (calibração, validação, calibração), garantindo a presença de amostras em toda a faixa de concentração em ambos os grupos. Para a escolha do melhor número de VL, a técnica de validação cruzada foi aplicada, utilizando subconjuntos aleatórios, com 10 subconjuntos e 20 iterações.

O método SPA está disponível em: www.ele.ita.br/~kawakami/spa/. A rotina OPS está disponível em: http: //www.deq.ufv.br/chemometrics. O GA, disponível no PLS toolbox, foi aplicado utilizando os seguintes parâmetros otimizados: população (64); gerações (100); taxa de mutação (0,005); largura da janela (1); convergência (50); termos iniciais (10) e cruzamentos (duplo). O OPS foi executado com as seguintes configurações otimizadas: vetor informativo: coeficientes de regressão; janela: 5 e incrementos: 10. Para o iPLS, vários números de intervalos foram testados, variando de 2 a 25. Para o método SVPII, o índice de importância foi calculado com

base no vetor de regressão, o número de iterações foi testado de 1-10 e a solução final foi definida com o menor valor de RMSE e o menor percentual de variáveis retidas.

4.7 Validação Analítica Multivariada

A otimização por seleção de variáveis e a validação analítica foram baseadas nas seguintes figuras de mérito (FOM): raiz quadrada dos erros médios de validação cruzada (RMSECV), calibração (RMSEC) e predição (RMSEP), assim como os respectivos coeficientes de correlação entre os valores de referência e previstos (Rcv, Rc e Rp) e relação de desempenho de desvio (RPD). A RPD é obtida como a razão entre os desvios padrão dos valores de referência (para os conjuntos de calibração e validação) e os valores de RMSECV e RMSEP, respectivamente [135].

Além disso, as seguintes FOM também foram determinadas: veracidade, precisão, linearidade, faixa de trabalho, seletividade, sensibilidade, sensibilidade analítica e viés (*bias*). A veracidade foi avaliada através dos erros relativos de predição para cada amostra. A precisão foi avaliada no nível de repetibilidade, obtida pelo mesmo analista, determinando no mesmo dia três amostras em diferentes níveis de concentração, em três repetições para cada uma. A linearidade foi estimada através do coeficiente de correlação (r) de um gráfico de valores de referência *versus* valores previstos, adotando a metodologia proposta por Souza e Junqueira [136]. Esta metodologia propõe validar os modelos aplicando os testes de Ryan-Joiner (RJ), Brown-Forsythe (BF) e Durbin-Watson (DW) aos resíduos do modelo, a fim de avaliar sua normalidade, homocedasticidade e independência, respectivamente.

A sensibilidade (SEN) foi calculada como o inverso da norma do vetor de regressão **b**, $\|b_i\|$. A sensibilidade analítica (γ), que expressa a menor diferença de concentração que pode ser distinguida pelo modelo, considerando o ruído experimental como única fonte significativa de erro [137], foi calculada como a razão entre a SEN e a estimativa do ruído. O ruído foi estimado pelo desvio padrão agrupado de uma matriz de dados construída com 10 espectros do branco para cada técnica. Para os dados de ATR-FTIR e NIR, os espectros foram registrados com a célula de leitura vazia. Para os dados de PS-MS, os espectros foram obtidos com o papel cromatográfico contendo apenas o solvente. Já para o TXRF, as leituras foram feitas no porta-amostras contendo apenas a solução de padrão interno.

Além da avaliação de parâmetros estatísticos para determinar a eficiência do modelo, métodos de detecção de amostras anômalas (*outliers*) foram utilizados com o objetivo de evitar que estas comprometam o desempenho do modelo. Dessa forma, os dados foram analisados com o auxílio de uma rotina, identificando amostras com altos resíduos espectrais, altos resíduos nos valores de referência e alto *leverage* (medida da influência das amostras no modelo) [136,142]. Os métodos de detecção de *outliers* foram empregados no subconjunto de amostras de calibração.

5. Resultados e Discussão

5.1. Espectros no Infravermelho Médio

Os espectros obtidos no infravermelho médio para as amostras de café submetidas a torras leve, média e forte (N = 40), na faixa de 40000-800 cm^{-1}, são apresentados na Figura 11. Além dos modelos individuais construídos exclusivamente para cada nível de torra, também foi desenvolvido um modelo robusto (N = 120), no qual todas as amostras (em diferentes graus de torra) foram modeladas simultaneamente. Espectros de todas as amostras usadas para construir este modelo robusto são mostrados na Figura 11d.

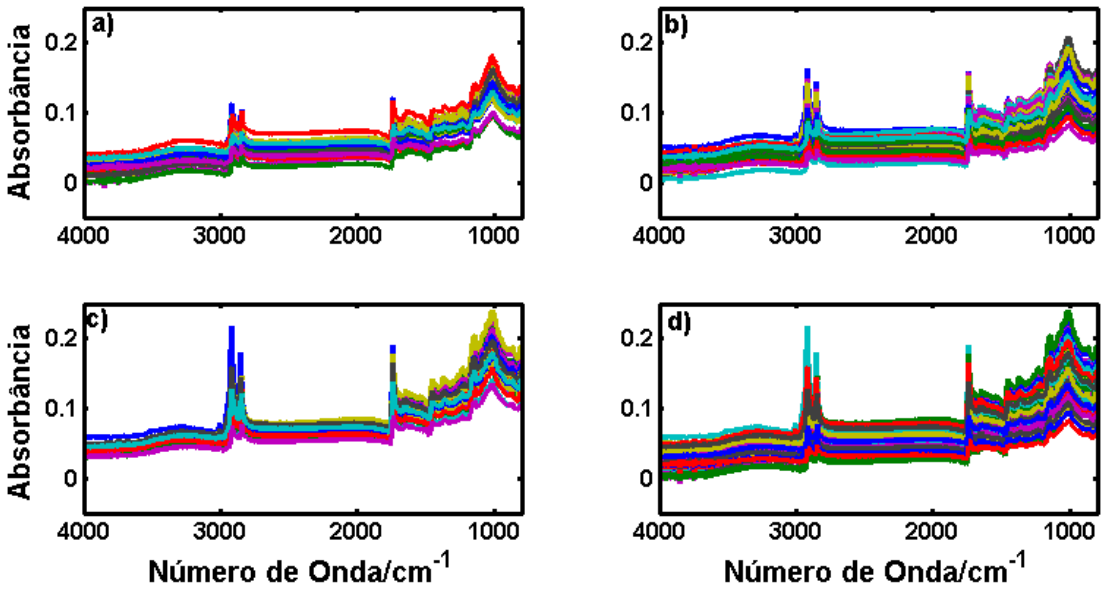

Figura 11. Espectros no Infravermelho Médio. (a) Amostras de torra leve, (b) torra média, (c) torra forte, e (d) junção de todos os espectros

Observando a Figura 11, pode-se notar que os espectros são bastante semelhantes visualmente. As diferenças são, essencialmente, nas intensidades de absorção. Dois picos presentes na região entre 3000-2800 cm^{-1} (a 2922-2920 e 2852-

2850 cm^{-1}) já foram previamente relatados para amostras de café torrado [140]. Este intervalo corresponde à absorção de ligações presentes em múltiplos constituintes do café. Essas vibrações incluem estiramentos de ligações C–H de hidrocarbonetos, e de ligações O–H de ácidos carboxílicos, estiramento assimétrico de ligações C–H de grupos metil ($–CH_3$) [141]. O pico a 1742 cm^{-1} é atribuído à ligação carbonila, geralmente relacionada a lipídios ou ésteres alifáticos presentes no café. O pico em 1543 cm^{-1} está relacionado à ligação C=C de anéis nitrogenados, muitas vezes associado a moléculas como a cafeína e a trigonelina, ambas presentes em quantidades significativas em cafés [18]. A banda intensa entre 1085-1050 cm^{-1} pode ser atribuída à vibração de deformação C–O axial, a banda entre 1420–1330 cm^{-1} é atribuída à deformação angular O–H, e a banda na região de 1300 a 1000 cm^{-1} é atribuída à vibração de ligação C–O–C de ésteres [17]. Além disso, a região de 1400 a 900 cm^{-1} é caracterizada por vibrações de vários tipos de ligações, como C-H, C-O, C-N e P-O [43]. A região entre 1800-800 cm^{-1} contém informações de impressões digitais importantes para a caracterização de cafés [43]. Em geral, essa região é caracterizada pela absorção de carboidratos que são responsáveis pela maioria das bandas no espectro do café torrado [140] .

Conforme mencionado, os espectros, nos três níveis de torra, são bastante semelhantes. No entanto, diferenças sutis podem ser notadas como, por exemplo, na região 1800-1680 cm^{-1}, relacionada à absorção de carbonila (especialmente na comparação entre torra leve e média). Lyman *et al*. [142] obtiveram espectros no infravermelho médio para amostras de extrato de café em diferentes níveis de torra. Esses autores acreditam que a diferença observada possa ser explicada devido ao aumento nos compostos relacionados a ésteres (1754-1744 cm^{-1}) e aldeídos (1741-1738 cm^{-1} e 1729-1723 cm^{-1}), e decréscimo em cetonas (1726 cm^{-1}), ácidos (1695 cm^{-}

^{1}) e ésteres de vinila e/ou lactona (1772 cm^{-1}). Estas alterações podem fornecer ao café de torra média um sabor mais encorpado, um maior equilíbrio entre sabor e aroma, e um sabor cítrico mais pronunciado, em comparação ao café de torra leve. Dessa forma, a diminuição relativa nos compostos ácidos é consistente com uma diminuição na qualidade do sabor [142]. Com relação às torras média e forte, mudanças importantes na concentração dos compostos carbonílicos são verificadas. Ocorrem aumentos nas quantidades de compostos relacionados a ésteres/lactonas insaturados (1772-1762 cm^{-1}), aldeídos/cetonas (1726 cm^{-1}), e ácidos (1706-1689 cm^{-1}). Por outro lado, há também diminuição na quantidade de compostos formados por ésteres (1755-1740 cm^{-1}) e aldeídos (em torno de 1739 cm^{-1}). Estas mudanças de concentrações são consistentes com os comentários dos provadores, de um sabor mais pesado e doce, com um sabor persistente de chocolate, característico da torra forte [142].

5.1.1. Desenvolvimento dos Modelos

Para todos os conjuntos de dados, o melhor pré-processamento foi a correção de espalhamento multiplicativo (MSC), seguido por centragem na média. Tal fato já era esperado, pois o MSC é um método que corrige os efeitos indesejados do espalhamento multiplicativo em, basicamente, duas etapas: (1) estimando os coeficientes de correção, linear e angular, a partir do espectro médio das amostras de calibração e, em seguida, (2) usando-os para corrigir o espectro original [133].

As regiões espectrais entre 4000-3600 cm^{-1} e 2820-1765 cm^{-1} foram eliminadas dos modelos, pois não apresentaram absorções significativas,

contribuindo apenas com ruído. Os parâmetros estatísticos obtidos na otimização dos modelos são mostrados na Tabela 3.

A detecção de amostras anômalas (*outliers*) foi realizada observando os resíduos espectrais (bloco X), altos valores de *leverage* (influência) ou grandes resíduos da previsão (bloco Y), com nível de confiança de 95%. Para cada modelo, no máximo duas amostras foram detectadas como *outliers* para o conjunto de calibração e o conjunto de validação.

A região espectral entre 3600 e 2970 cm^{-1} apresentou uma banda larga e de pouca intensidade associada às vibrações de estiramento de O-H e N-H. Modelos PLS com e sem essa região foram construídos e testados. Para os modelos nos três níveis de torra, as melhores previsões foram obtidas sem essa região, correspondendo a um total de 1151 variáveis. Por outro lado, para o modelo robusto, a inclusão dessa região proporcionou um modelo com melhor desempenho, utilizando um total de 1801 variáveis.

Tabela 3. Parâmetros estatísticos para modelos de torra leve, média, forte e robusto.

	Torra Leve						Torra Média					
	Total	**OPS**	***i*PLS**	**SPA**	**GA**	**SVPII**	**Total**	**OPS**	***i*PLS**	**SPA**	**GA**	**SVPII**
n_{VL}	5	5	5	4	5	5	5	5	5	5	5	5
n_{VL} ***OPS***	-	8	-	-	-	-	-	12	-	-	-	-
nVars	1151	250	575	346	76	241	1151	80	384	384	75	337
RMSECV(%)	3,0	2,0	2,3	2,9	1,3	3,8	3,1	2,0	3,0	3,7	1,9	3,9
RMSEC (%)	1,3	1	0,9	2,1	0,5	1,2	1,4	1,1	1,3	1,6	0,8	2,9
Rc	0,98	0,99	0,99	0,96	0,99	0,99	0,98	0,99	0,98	0,98	0,99	0,91
Rcv	0,93	0,97	0,96	0,92	0,99	0,86	0,92	0,97	0,92	0,88	0,97	0,85
RMSEP (%)	2,4	1,5	2,1	2,3	1,1	2,8	2,5	1,7	2,4	3,1	1,3	4,6
Rp	0,95	0,98	0,94	0,92	0,99	0,91	0,92	0,97	0,82	0,85	0,98	0,84
%ER	17,7	12,3	14,2	16,2	8,6	9,3	16,2	9,9	16,4	25,7	7,5	21,4

	Torra Forte						Robusto					
	Total	**OPS**	***i*PLS**	**SPA**	**GA**	**SVPII**	**Total**	**OPS**	**iPLS**	**SPA**	**GA**	**SVPII**
n_{VL}	5	4	5	5	4	4	8	8	8	8	8	7
n_{VL} ***OPS***	-	14	-	-	-	-	-	15	-	-	-	-
nVars	1151	140	575	385	127	972	1801	310	82	383	141	796
RMSECV(%)	3,1	3,1	3,5	4,5	2,4	4,6	3,9	2,4	3,5	3,5	2,3	3,8
RMSEC (%)	1,2	1,3	1,3	2,2	1,4	2,2	2,0	1,2	1,8	2,8	1,1	2,36
Rc	0,99	0,99	0,98	0,95	0,98	0,95	0,96	0,99	0,97	0,99	0,98	0,95
Rcv	0,91	0,92	0,88	0,81	0,95	0,80	0,87	0,95	0,90	0,99	0,95	0,87
RMSEP (%)	2,8	1,3	2,0	2,2	0,9	2,9	3,2	1,8	3,0	3,7	1,8	3,4
Rp	0,91	0,99	0,97	0,95	0,99	0,97	0,85	0,95	0,86	0,80	0,94	0,83
%ER	13,4	9,2	11,5	7,3	6,4	14,35	23,9	12,0	22,6	30,5	13,2	22,22

Observando os resultados da Tabela 3, pode-se notar que, em geral, a seleção de variáveis melhorou consideravelmente os modelos. Embora o *i*PLS, SPA-PLS e o SVPII não tenham melhorado significativamente os resultados, o OPS e o GA proporcionaram um grande aumento na capacidade preditiva dos modelos de calibração. Uma possível explicação para o pior desempenho do iPLS e do iSPA-PLS está na natureza complexa da matriz alimentar analisada. Como o café é uma mistura complexa de muitas substâncias que podem variar em conteúdo, dependendo de sua variedade, Arabica ou Robusta, espera-se a presença de variáveis preditivas distribuídas em todo o espectro MIR. Como o iPLS é um método que seleciona regiões espectrais contínuas localizadas entre intervalos específicos de números de onda, ele não foi capaz de produzir bons modelos preditivos. Para o método SVPII, uma possível explicação para o mau desempenho foi a escolha, durante a otimização, do tipo 2, ou seja, a solução final é uma combinação entre menor valor de RMSEC e menor percentual de variáveis retidas. Dessa forma, o algoritmo selecionou um conjunto muito pequeno de variáveis, eliminando informação importante para uma maior capacidade preditiva.

Por outro lado, OPS e GA podem realizar uma busca discreta de variáveis em todo o espectro. Assim, esses dois métodos de seleção de variáveis melhoraram consideravelmente o desempenho dos modelos, fato comprovado analisando todos os parâmetros estatísticos estimados.

Em geral, os resultados de OPS e GA não diferiram muito entre si. Para os três modelos construídos em cada nível de torra, o GA forneceu resultados um pouco melhores do que o OPS (Tabela 3). O principal parâmetro de comparação é o RMSEP, que foi reduzido em 54% (de 2,4% para 1,1%) para o modelo de torra leve, em 48% (de 2,5% para 1,3%) para o modelo de torra média e em 68% (de 2,8% para 0,9%)

para o modelo de torra escura. Os valores de RMSEC e RMSECV também foram significativamente menores para todos esses três modelos, e todos os coeficientes de correlação foram maiores. Erros relativos (ER) foram reduzidos, em média, de 17,7% para 8,6% (leve), 16,2% para 7,5% (médio) e 13,4% para 6,4% (escuro). Os modelos GA-PLS foram construídos com 4 ou 5 VL, representando mais de 99,05% da variância explicada no bloco X e mais de 98,89% no bloco Y. O número de variáveis utilizadas nos modelos foi reduzido para menos de 11% das variáveis originais (espectros inteiros): de 1151 para 76 (torra leve), 75 (torra média) e 127 (torra forte). Assim, esses resultados indicam que o menor número de variáveis selecionadas para cada modelo representa os conjuntos de número de onda mais ricos em informações específicas para predizer a composição de misturas de café Arabica-Robusta. A eficiência da seleção de variáveis foi demonstrada pela eliminação de muitas variáveis não informativas e ruidosas.

Além de modelos específicos construídos exclusivamente para misturas de café obtidas em graus fixos de torra, também foi construído um modelo robusto que incorporou todas as amostras. Este é o modelo mais importante, porque pode ser aplicado a qualquer amostra, independente do seu nível de torra. Para este conjunto de dados, o OPS forneceu resultados ligeiramente melhores do que o GA. Como era de se esperar, esse modelo mais complexo foi construído com um número maior de VL, 8, representando 98,61% da variância no bloco X e 98,53% no bloco y. Em comparação com o modelo de espectro completo (Tabela 3), o OPS-PLS reduziu o número de variáveis para cerca de 17% do espectro original (de 1801 a 310). Todos os parâmetros foram melhorados em relação ao modelo completo. Particularmente, o RMSEP foi reduzido em 44% (de 3,2% para 1,8%).

Além disso, dos fatores citados, também é importante observar os erros relativos individuais para as previsões das amostras. Os histogramas dos erros relativos para amostras nos conjuntos de validação para todos os quatro modelos são mostrados na Figura 12. Os maiores erros relativos, acima de ± 20%, foram observados para um pequeno número de amostras em baixas concentrações de café Robusta, o que comprova o bom desempenho dos modelos.

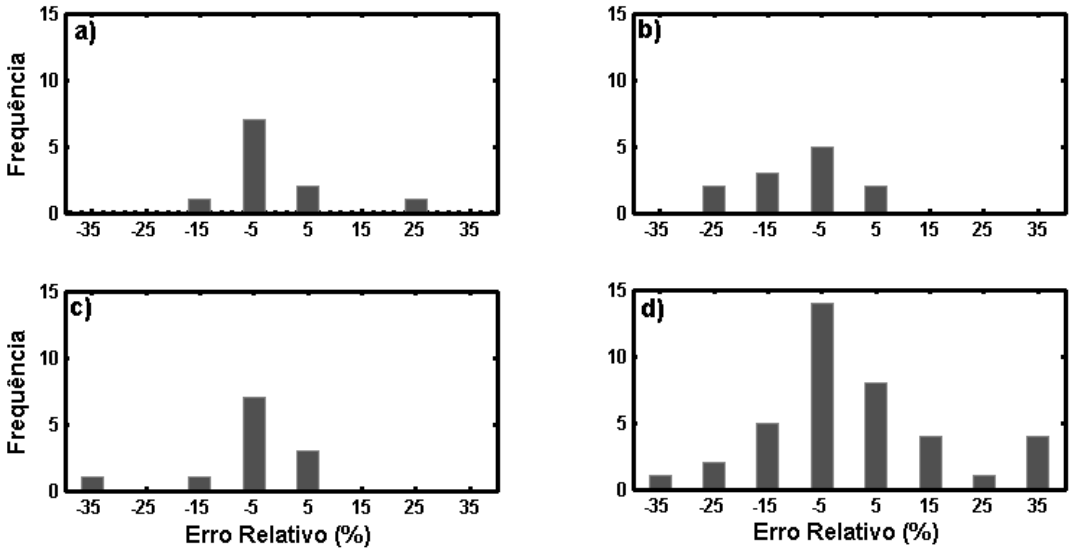

Figura 12. Erros relativos para os modelos: (a) Torra Leve, (a) Torra Média, (c) Torra Forte, (d) Robusto

A linearidade dos modelos pode ser avaliada analisando o ajuste dos valores de referência *versus* valores previstos. Estes gráficos, para todos os quatro modelos, são mostradas na Figura 13, na qual nenhuma tendência sistemática é observada para as distribuições residuais. Os coeficientes de correlação estimados para todos os modelos apresentaram valores acima de 0,94 (Tabela 3).

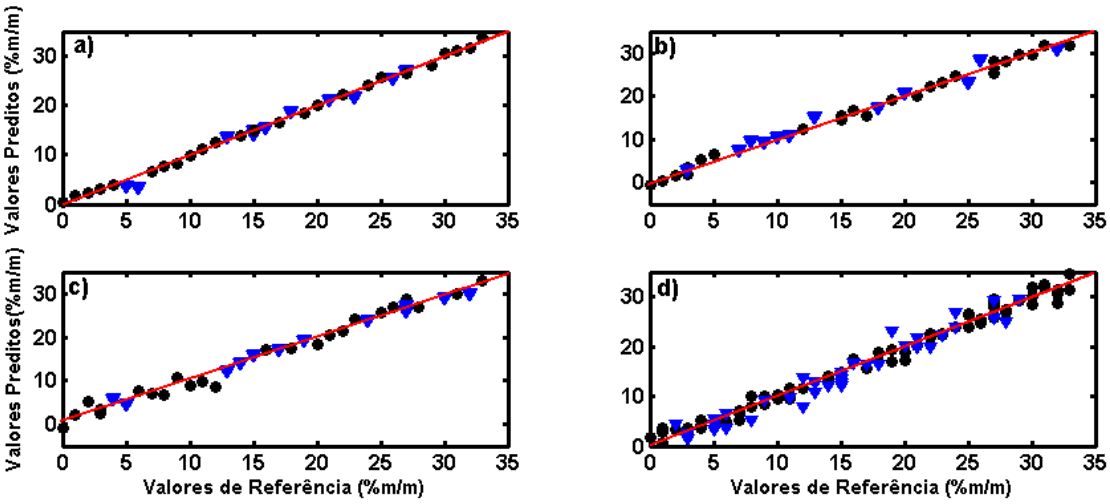

Figura 13. Valores medidos versus preditos para os modelos: (a) Torra Leve, (a) Torra Média, (c) Torra Forte, (d) Robusto

5.1.2 Interpretação das Variáveis Selecionadas

As variáveis selecionadas para os melhores modelos são mostradas na Figura 14. Para todos os modelos, o OPS selecionou um número maior de variáveis, se comparado ao GA (Tabela 3), variando de cerca de três vezes mais (torra leve) a quase o mesmo número de variáveis (torra média). No entanto, quase todas as variáveis escolhidas pelo GA também foram selecionadas pelo OPS. Ambos os métodos selecionaram variáveis distribuídas em todo o espectro MIR, não apresentando diferenças significativas em termos de interpretação do modelo. Esse resultado corrobora a natureza complexa das amostras e o fato de que a diferenciação entre os cafés Robusta e Arabica deve-se a diferentes compostos que absorvem em diferentes regiões, distribuídas por todo o espectro.

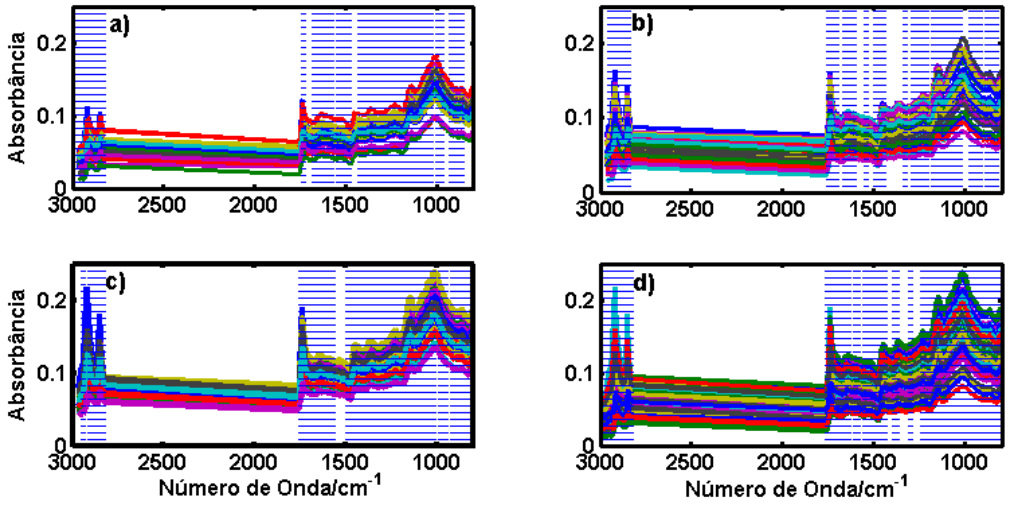

Figura 14. Variáveis selecionadas nos melhores modelos. (a) Torra Leve, (a) Torra Média, (c) Torra Forte, (d) Modelo Robusto

Observando a Figura 14, pode-se notar que a região entre 3000-2800 cm^{-1} foi selecionada para todos os melhores modelos. Esta região, atribuída a estiramento de ligações C-H de grupos metila, está relacionada a múltiplos constituintes do café, entre ele, cafeína [141]. Os cafés Robusta apresentam maior teor de cafeína que a espécie Arabica. Muitas variáveis também foram selecionadas na região da impressão digital, entre 1700-900 cm^{-1}. Os carboidratos exibem várias vibrações nessa região e, provavelmente, são responsáveis pela maioria das variáveis selecionadas [140]. O pico centrado em 1742 cm^{-1}, selecionado para todos os modelos, é atribuído à vibração de carbonila de lipídios ou ésteres alifáticos [18]. Os cafés Arabica contêm cerca de 60% mais lipídios do que os Robusta. Além disso, a região entre 1400-900 cm^{-1} é caracterizada por vibrações de vários tipos de ligações, como C-H, C-O e C-N. Desta forma, os ácidos clorogênicos, uma família de ésteres cujo teor é maior no Robusta que no Arabica, mostram fortes absorções na região de 1450-1000 cm^{-1} [16]. O pico em 1543 cm^{-1} está associado ao estiramento de ligações C=C de anéis de nitrogênio [18], presentes em moléculas como a cafeína e a trigonelina, cujos teores

diferem significativamente em função da espécie de café. Vibrações do modo esquelético de ligações glicosídicas de amido são observadas na região entre 950-800 cm^{-1} [16]. Bandas na região entre 1085-1050 cm^{-1} são geralmente atribuídas a deformações axiais de C-O, especialmente de ácido quínico em cafés; a região entre 1420–1330 cm^{-1} apresenta vibrações de deformações angulares de O–H; e as vibrações da ligação da ligação C–O–C de ésteres são observadas na região de 1300–1000 cm^{-1} [17].

Um aspecto importante a ser destacado é a inclusão da região correspondente à vibração de O-H (3600-2970 cm^{-1}), associada principalmente à presença de água, no modelo robusto. Como mencionado na seção anterior, um modelo completo de espectros incluindo essa banda apresentou melhores previsões apenas para o modelo robusto. Dessa forma, a modelagem de amostras em diferentes níveis de torra exige a inclusão dessa região para fornecer boas previsões.

5.2. Demais Técnicas Isoladas

O objetivo inicial do trabalho era a construção de modelos de calibração para cada nível de torra, separadamente. Ao construir os modelos no ATR-FTIR, conforme mencionado, percebeu-se que o modelo robusto, que engloba todos os graus de torra, apresentou um bom desempenho. Esse modelo apresenta vantagens sobre os modelos individuais, porque permite a modelagem de todos os níveis de torra simultaneamente. A vantagem de robustez do modelo deve ser realçada para amostras que estão em grau intermediário de torra, as quais poderiam não ser bem previstas pelos modelos individuais. Dessa forma, para todos os modelos que serão discutidos nesta tese, a estratégia de modelagem robusta será a única apresentada.

Além disso, pela experiência obtida nos dados de ATR-FTIR, concluiu-se que o OPS e GA apresentaram melhor desempenho na seleção de variáveis. Esta tendência foi observada também para outros tipos de espectros. Logo, para os demais modelos desenvolvidos, apenas os parâmetros estatísticos desses algoritmos serão mostrados.

5.2.1. Espectros NIR

Os espectros obtidos no infravermelho próximo para as amostras de café nas torras leve e média (N = 80), na faixa de 11000-5000 cm^{-1} (900-2000 nm), são apresentados na Figura 15. As regiões entre 800-900 e 2000-2300 nm foram deletadas por apresentarem alto ruído. Para as análises no NIR, não foi possível obter os espectros das amostras de torra forte, pois o equipamento apresentou problemas operacionais.

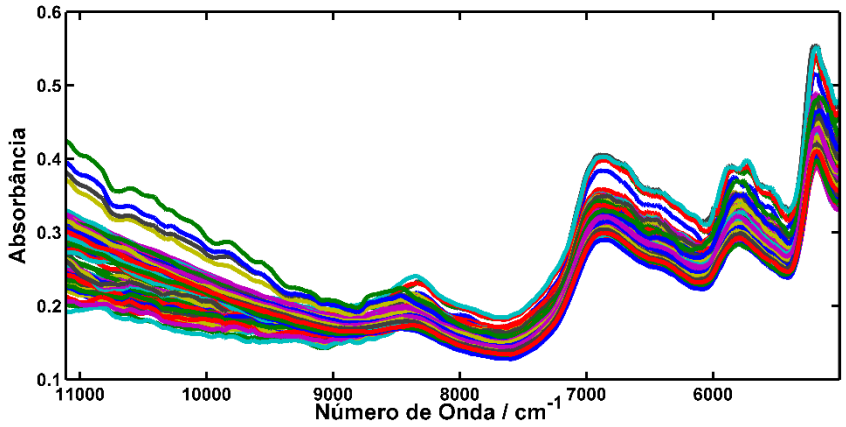

Figura 15. Espectros NIR para as amostras nas torras leve e média.

Observando a Figura 15, pode-se perceber que sinais intensos de bandas de combinação e sobretons estão presentes em regiões relacionadas à absorção da água (5200-5000 cm^{-1} e 7200-6800 cm^{-1}) [143]. Em geral, os números de onda entre 5300-5200 cm^{-1} estão relacionados às vibrações de ligações C=O e O-H. Números de

onda na região de 6100 a 4965 cm^{-1} podem ser atribuídos ao primeiro sobretom de vibrações de C-H. A região entre 9800-9000 cm^{-1} está relacionada ao segundo sobretom de vibrações de ligações C-H e N-H [143]. As vibrações em torno de 7212 cm^{-1} estão relacionadas ao segundo sobretom da deformação angular de ligações C-H. Além disso, outros picos específicos presentes, tais como em 5797 e 5666 cm^{-1}, estão relacionados ao conteúdo de lipídios. Em geral, o café Arabica é mais rico em ácidos graxos do que o Robusta.

As vibrações de grupos carbonila, CH e CH_2 são geralmente causadas por múltiplos constituintes do café, tais como proteínas, lipídeos, ácidos voláteis e não voláteis, ácidos clorogênicos, alcaloides, compostos aromáticos, entre outros [144]. Vale ressaltar também, conforme já relatado na literatura, que a intensidade dos sinais (relacionados à água) em torno de 6896 cm^{-1} (primeiro sobretom estiramento O-H) e 5154 cm^{-1} (banda de combinação de estiramento e deformação O-H) diminuíram gradualmente, de acordo com o nível da torra [56].

5.2.1.2 Desenvolvimento de Modelos de Calibração

Para a construção dos modelos de calibração no NIR, foi aplicado o pré-processamento MSC e, em seguida, os dados foram centrados na média. A detecção de *outliers* foi realizada e um total de 5 amostras foram deletadas dos conjuntos de calibração e validação.

A Tabela 4 contém os parâmetros estatísticos dos modelos otimizados.

Tabela 4. Parâmetros Estatísticos do modelo Robusto utilizando NIR

	NIR		
	Completo	**OPS**	***GA***
n_{VL}	*6*	4	6
n_{VL} ***OPS***		12	
nVars	*1100*	235	67
RMSECV(%)	*5,9*	5,1	4,1
RMSEC (%)	*3,6*	3,4	2,4
Rc	*0,86*	0,87	0,94
Rcv	*0,66*	0,74	0,83
RMSEP (%)	*4,4*	3,4	3,4
Rp	*0,8*	0,87	0,9

Em geral, os modelos de calibração utilizando NIR não apresentaram bom desempenho, mesmo após a otimização por seleção de variáveis. Os valores de RMSEC, RMSECV e RMSEP foram altos, se comparado ao melhor modelo utilizando MIR, por exemplo. Além disso, os parâmetros Rc, Rcv e Rp apresentaram, em sua maioria, valores inferiores a 0,94, indicando baixa linearidade entre os dados. Tal fato pode ter sido ocasionado por diversas razões sendo a principal uma questão instrumental. O equipamento utilizado é portátil e é fortemente influenciado pelas características externas, tais como temperatura, umidade, movimentação na sala, entre outros aspectos. Além disso, durante o processo de leitura no NIR, percebeu-se que o instrumento tinha baixa repetibilidade.

5.2.1 Espectros PS-MS

A Figura 16 mostra o espectro de massas de uma amostra representativa de um *blend* formado por 26,0% de Robusta, torra leve.

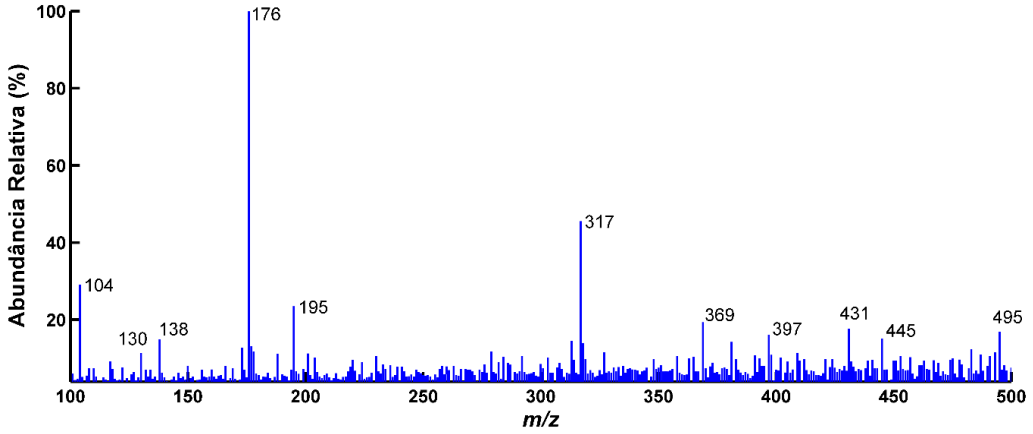

Figura 16. Exemplo de um espectro PS (+) - MS (concentração de 26,0% Robusta, torra média).

Em geral, para todas as amostras, em todos os níveis de torra, o sinal *m/z* mais intenso foi o de *m/z* 176. Este sinal base, relatado em outros estudos, é usualmente atribuído à trigonelina $[M + K]^+$ [22], um importante alcaloide presente no café. Além disso, o íon de *m/z* 138, observado em todos os espectros de massa, também pode ser relacionado à trigonelina $[M + H]^+$. Outros íons importantes, presentes em todos os espectros, podem ser atribuídos a importantes constituintes do café, tais como: *m/z* 104 (colina $[M + H]^+$) [23], *m/z* 195 (cafeína $[M + H]^+$) [22], *m/z* 317 (ácido linolênico $[M + K]^+$) [145], *m/z* 393 (ácido cafeoilquínico $[M + K]^+$) e *m/z* 381 (sacarose [M + K] +).

5.2.1.1 Desenvolvimentos dos modelos

Os dados provenientes da técnica PS-MS foram centrados na média, antes da etapa da construção dos modelos. Por ser um espectro altamente ruidoso, o modelo completo, que inclui todas as variáveis, teve um mau desempenho. Além disso, aplicando os métodos de seleção de variáveis, não foi possível a obtenção de bons

modelos. O melhor modelo, obtido com OPS, forneceu valores de RMSEC e RMSEP na ordem de 3,1% e 6,2%, respectivamente, com 6 VL.

5.2.2 Dados de TXRF

Para a construção dos modelos usando dados obtidos por TXRF, a matriz continha 120 linhas (correspondente às amostras) e 14 colunas, que estão relacionadas aos elementos investigados: P, S, Cl, K, Ca, Ti, Mn, Fe, Ni, Cu, Zn, Br, Rb e Sr. Os valores foram expressos em mgL^{-1}. Para esses dados, K tem a maior concentração, seguido de Ca e P, que têm valores próximos, e S e Cl. Os demais elementos possuem concentrações muito inferiores aos destacados. O anexo I contém, como ilustração, os resultados de concentração de cada elemento estimada por TXRF para as 40 amostras de torra leve. É importante frisar que, no caso desta técnica, ao contrário das outras, os espectros não foram tratados diretamente por métodos quimiométricos, mas as quantificações dos teores dos elementos obtidas dos espectros.

Os parâmetros estatísticos para os modelos completos e utilizando seleção de variáveis estão na Tabela 5. Para este conjunto, os dados foram autoescalados e 6 amostras foram identificadas como *outliers* (3 do conjunto calibração e 3 do conjunto validação).

Tabela 5. Parâmetros estatísticos para os dados obtidos por TXRF

	TXRF		
	Completo	**OPS**	***GA***
n_{VL}	*6*	4	4
n_{VL} ***OPS***		11	
nVars	*14*	8	6
RMSECV(%)	*2,5*	2,6	2,0
RMSEC (%)	*1,8*	2,2	1,5
Rc	*0,96*	0,95	0,97
Rcv	*0,94*	0,93	0,96
RMSEP (%)	*1,6*	2,2	1,2
Rp	*0,97*	0,95	0,98

Analisando a Tabela 5, pode-se perceber que o modelo completo, apesar de utilizar um número maior de VL, teve um melhor desempenho, se comparado ao método OPS. Por outro lado, o GA reduziu todos os valores de erros, aumentando os valores de R (tanto para calibração como validação).

O algoritmo GA selecionou 6 variáveis: Ca, Mn, Fe, Cu, Zn, Rb. Analisando os coeficientes de regressão, apenas Ca e Mn apresentaram valores negativos, o que indica que estão mais relacionados com o conteúdo de Arabica), enquanto Fe, Cu, Zn e Rb produziram coeficientes positivos.

Conforme mencionado, existem poucos trabalhos na literatura que relacionam conteúdo elementar com as espécies Arabica e Robusta. Artigos empregando calibração multivariada e técnicas atômicas com o objetivo de quantificar e caracterizar os *blends* de cafés são ainda mais raros. Santato *et. al* (2012) [85] construíram modelos para prever a origem de grãos de café verde utilizados em diferentes partes do mundo, combinando os perfis elementar (usando a espectrometria de massa com plasma indutivamente acoplado) e isotópico. Como resultado, eles observaram que os grãos Robusta apresentavam maiores concentrações de P e Cu e uma menor concentração de Mn. Martín *et. al* [83] mediram

o teor de onze metais, Zn, P, Mn, Fe, Mg, Ca, Na, K, Cu, Sr e Ba, escolhidos como descritores químicos, por espectrometria de emissão óptica com plasma indutivamente acoplado (ICP OES). Eles observaram que P, Mn e Cu são muito apropriados para caracterizar as variedades, Arabica e Robusta. Dessa forma, podemos concluir que tanto Cu como Mn podem ser classificados como marcadores químicos para diferenciação de Robusta e Arabica, respectivamente.

Os histogramas dos erros relativos para amostras nos conjuntos de calibração e validação, para o modelo TXRF-GA, são mostrados na Figura 17.

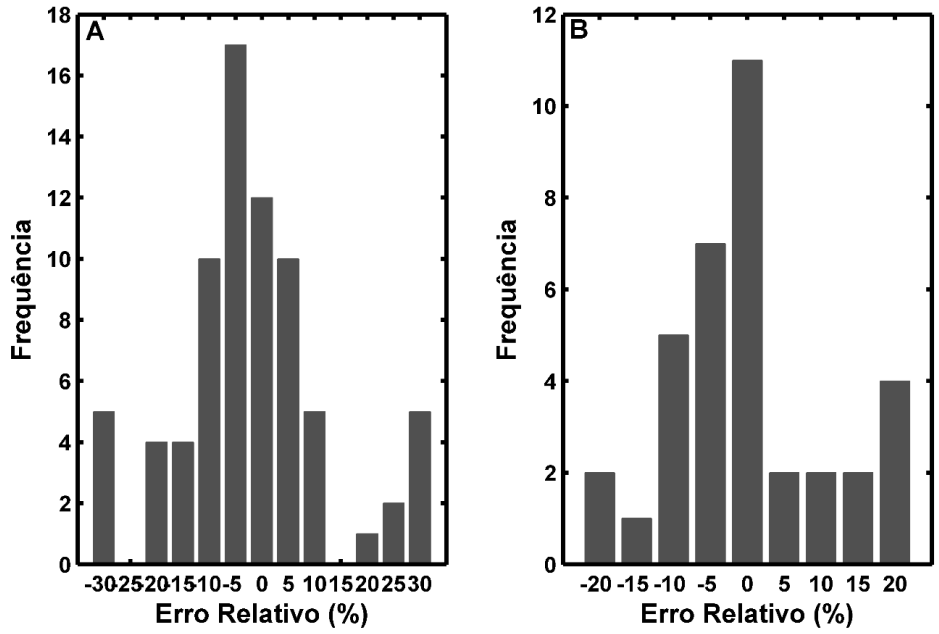

Figura 17. Histograma de erros relativos para os dados de TXRF para os conjuntos de: (A) calibração e (B) validação

Para avaliar a linearidade dos modelos, o gráfico contendo o ajuste (TXRF-GA) dos valores de referência *versus* valores previstos está na Figura 18. Os coeficientes de correlação obtidos foram todos acima de 0,96 (Tabela 5).

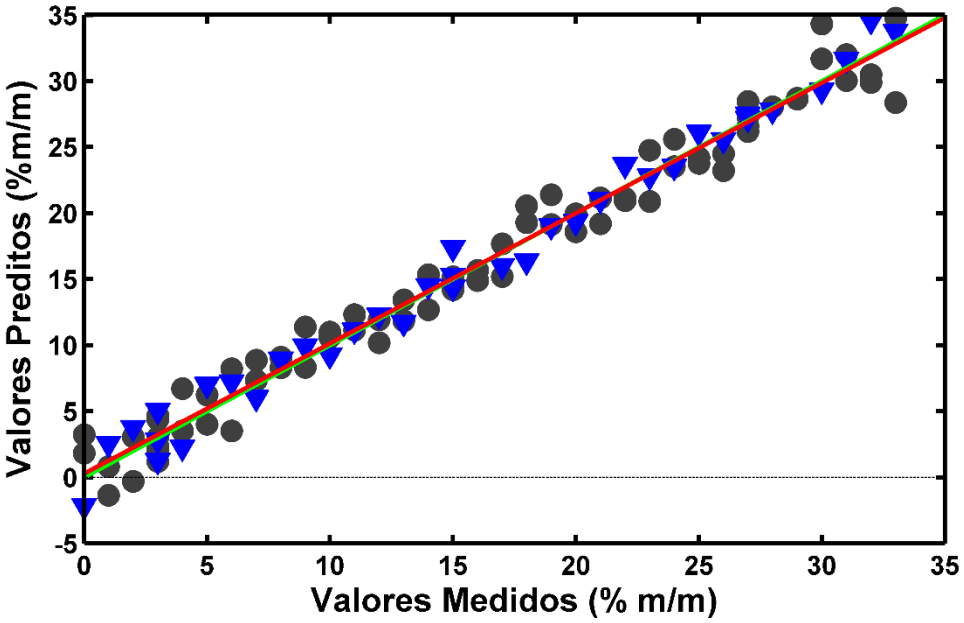

Figura 18. Valores medidos *versus* preditos para o modelo TXRF: Círculos correspondem ao conjunto calibração e triângulos ao conjunto de validação.

5.3 Fusão de Dados PS-MS e MIR

5.3.1 Desenvolvimento dos Modelos de Fusão de Dados

Conforme mencionado, para os dados de ATR-FTIR, duas regiões espectrais foram deletadas, 4000-3600 cm^{-1} e 2820-1765 cm^{-1}. Estas duas regiões não mostraram absorções significativas (Figura 11) e, portanto, contribuiriam apenas com ruído para os modelos. Como também mencionado anteriormente, dois métodos de seleção de variáveis, GA e OPS, foram aplicados para otimizar todos os modelos de calibração desenvolvidos. Esses modelos também foram submetidos a detecção de *outliers*. Amostras com grandes resíduos espectrais, altos valores de *leverage* ou grandes resíduos em Y [135], no nível de confiança de 95%, foram removidos dos

modelos. Em todos os modelos desenvolvidos, foram detectados no máximo três *outliers*. Para o melhor modelo, discutido a seguir, apenas dois *outliers* foram detectados no conjunto de calibração (2,4% das amostras de calibração).

Conforme discutido em outras seções, um modelo construído apenas com os espectros de ATR-FTIR apresentou resultados razoáveis. Por outro lado, os dados do PS-MS não forneceram bons modelos. Dessa forma, modelos de fusão de dados de nível baixo e médio, utilizando espectros completos e otimizados por GA e OPS, foram comparados com base nos valores de RMSEC, RMSEP e coeficientes de correlação para os conjuntos de calibração (Rc) e validação (Rp). Esses parâmetros são apresentados na Tabela 6.

Tabela 6. Parâmetros Estatísticos para dos modelos de fusão de dados

	Nível Baixo			**Nível Médio**		
	Completo	**OPS**	**GA**	**Completo**	**OPS**	**GA**
nVars	2202	230	193	2202	320	233
nVL	5	6	5	5	6	5
RMSEC(%)	2,7	1,0	2,1	1,89	1,5	1,7
R_c	0,96	0,99	0,98	0,97	0,99	0,98
RMSEP %)	3,2	1,7	2,5	4,3	2,3	1,9
R_p	0,94	0,98	0,97	0,87	0,97	0,98

Os resultados da Tabela 6 mostram que tanto o OPS quanto o GA melhoraram todos os modelos, descartando entre 85 e 91% das variáveis originais. Este fato demonstra o bom desempenho dos métodos de seleção de variáveis, eliminando informações ruidosas e irrelevantes que não estão correlacionadas com a propriedade de interesse (teor de café Robusta). Em relação ao nível de fusão de dados, os modelos de baixo nível foram, em geral, melhores que os de nível médio. Tal fato pode ser explicado pois a fusão nível médio não foi feita de forma totalmente otimizada. Na

realidade, a fusão nível médio é recomendada para grandes conjuntos de dados (dezenas de milhares de variáveis, por exemplo) devido a compressão dos dados. Como o conjunto de dados trabalhado não era tão grande, a fusão nível baixo se tornou mais vantajosa.

Para a fusão de dados de nível médio, o GA apresentou resultados ligeiramente melhores do que o OPS, enquanto o modelo OPS de baixo nível forneceu os menores erros de previsão. OPS tem a vantagem adicional de gastar muito menos tempo computacional do que o GA. Assim, a fusão de dados de baixo nível OPS-PLS foi escolhida como o melhor modelo.

O melhor modelo de fusão de dados forneceu resultados muito melhores do que o modelo OPS-PLS construído apenas com dados PS-MS, e ligeiramente melhor que o modelo OPS-PLS construído com espectros ATR-FTIR (neste caso, reduzindo o número de VL de 8 para 6, Tabela3). Para a fusão de dados, o OPS simplificou o modelo completo, reduzindo o número de variáveis de 2202 para 230, uma redução de 89,6%. Isso é muito importante em termos de interpretabilidade do modelo. Além disso, todos os parâmetros do modelo foram significativamente melhorados em relação ao modelo completo. O RMSEP foi reduzido em 47%, de 3,2% para 1,7%, enquanto o RMSEC diminuiu 63%, de 2,7% para 1,0%.

É interessante observar também os erros relativos individuais para cada amostra. Os histogramas dos erros relativos para amostras nos conjuntos de calibração e validação são mostrados na Figura 19.

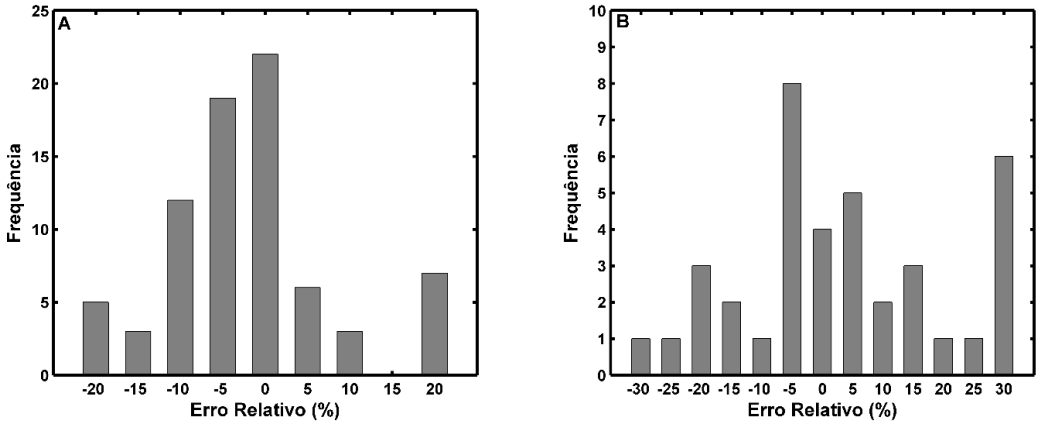

Figura 19. Erros Relativos: A) conjunto de calibração, B) conjunto de validação

Em geral, a linearidade de modelos de calibração multivariada é tipicamente analisada com base no ajuste dos valores de referência *versus* valores previstos (Figura 20). Coeficientes de correlação de 0,99 e 0,98 foram estimados para amostras de calibração e validação, indicando alta correlação entre os dados.

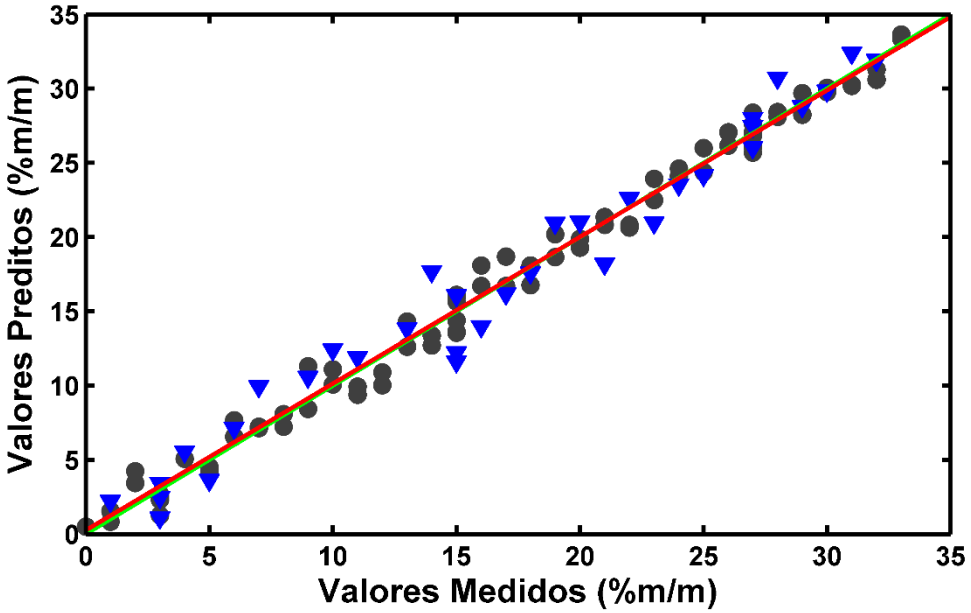

Figura 20. Valores medidos e preditos para o modelo PLS-OPS.

5.3.2 Interpretação das Variáveis Selecionadas

Das 2202 variáveis espectrais originais, a fusão de dados de nível baixo OPS-PLS selecionou 230 variáveis, sendo 111 dos espectros ATR-FTIR e 119 dos dados PS-MS. As variáveis selecionadas a partir dos espectros de ATR-FTIR são mostradas na Figura 21.

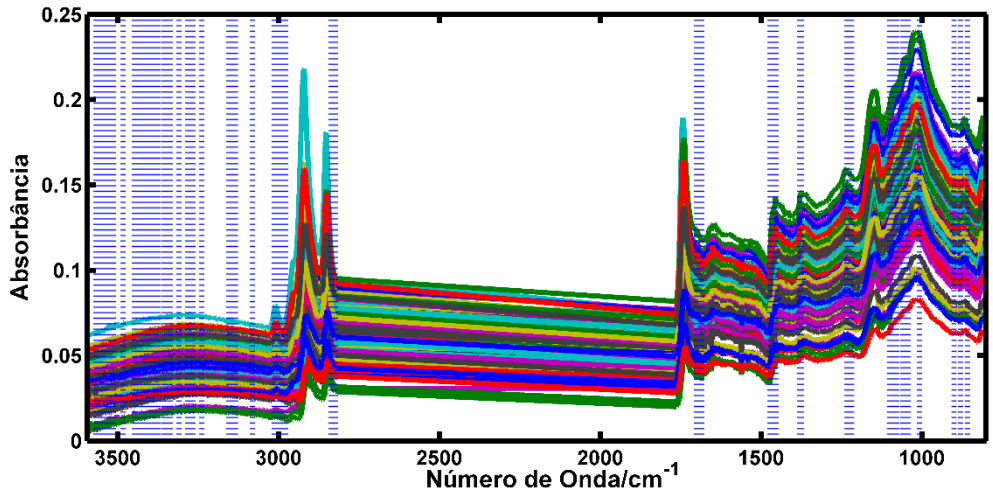

Figura 21. Variáveis selecionadas (ATR-FTIR) pelo modelo PLS-OPS de fusão de dados

A observação da Figura 21 permite concluir que, assim como para o modelo individual ATR-FTIR, as variáveis selecionadas estão distribuídas por toda a região espectral, indicando, mais uma vez, que a diferenciação entre os cafés Robusta e Arabica se deve, principalmente, a diversos compostos químicos absorvendo em diferentes regiões do espectro no infravermelho médio. Esta observação confirma a complexidade desta matriz alimentar e a necessidade de ferramentas quimiométricas para melhor prever e interpretar os dados.

A análise dos coeficientes de regressão é uma ferramenta adicional para a interpretação espectral do modelo. A Figura 22 mostra os coeficientes de regressão para os espectros ATR-FTIR (A) e PS-MS (B) no melhor modelo de fusão de dados.

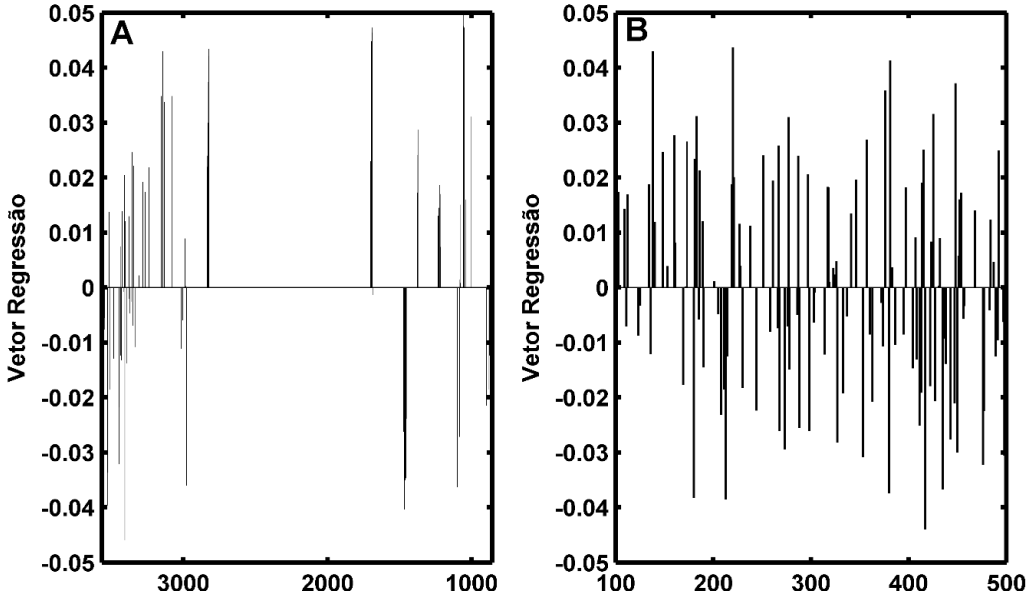

Figura 22. Vetor de regressão para: (A) ATR-FTIR e (B) PS-MS

De acordo com as Figuras 21 e 22(A), várias variáveis foram selecionadas na região espectral entre 3600-3000 cm^{-1}. Esta região é dominada por uma banda larga usualmente associada com o conteúdo de água. Além disso, algumas variáveis na região entre 3000-2800 cm^{-1} foram selecionadas. Esta região está relacionada a vibrações de ligações C–H dos grupos metila, e esses números de onda podem ser associados à cafeína e trigonelina [141], entre vários outros compostos. Assim, a presença de coeficientes de regressão negativos e positivos nesta região espectral pode ser explicada pelas diferenças químicas entre estas duas espécies de café, uma vez que os cafés Robusta possuem menores teores de trigonelina e maiores teores de cafeína que os cafés Arabica.

Na região de impressão digital (*fingerprint*), vários números de onda foram selecionados, alguns deles relacionados a moléculas de carboidratos [140]. O intervalo entre 1400-1000 cm^{-1} tem sido associado a vibrações de ácidos clorogênicos presentes em amostras de café [43]. Observou-se que a maioria dos coeficientes de regressão são positivos nesta região, o que é considerado consistente com a literatura. De fato, os ácidos clorogênicos são as principais espécies fenólicas presentes no café, as quais são formadas por reações de esterificação entre o ácido quínico e os ácidos hidroxicinâmicos. Esses compostos são particularmente importantes por estarem associados a efeitos benéficos para a saúde, e os cafés Robusta são mais ricos em ácidos clorogênicos do que os da espécie Arabica [146]. Isso se deve ao clima e a outros fatores ambientais que influenciam os teores de ácidos clorogênicos nos grãos de café, uma vez que os cafés Robusta são cultivados em climas mais severos do que os cafés Arabica.

Números de onda próximos a 1085 e 1050 cm^{-1}, que estão associados a picos atribuídos às deformações axiais de ligações C–O, também foram selecionados. Esses sinais analíticos podem estar relacionados ao ácido quínico, também mais concentrado em cafés Robusta (coeficientes de regressão positivos no modelo). Finalmente, algumas variáveis foram selecionadas na região entre 950-800 cm^{-1}, o que pode estar relacionado às vibrações do modo esquelético das ligações glicosídicas [16]. Os cafés Arabica são mais ricos em carboidratos do que os cafés Robusta e os coeficientes de regressão nessa região foram todos consistentemente negativos.

Como mencionado anteriormente, o OPS selecionou 119 variáveis dos dados de PS-MS. As variáveis selecionadas dos espectros de massa, em geral, corroboram e complementam a interpretação obtida a partir dos espectros de infravermelho. O íon

mais abundante, atribuído à trigonelina $[M+K]^+$, de *m/z* 176, não foi selecionado. É importante ressaltar que os íons mais abundantes não são necessariamente os mais discriminantes. No entanto, o íon de m/z 138, também associado à trigonelina ($[M+H]^+$), foi selecionado pelo OPS, corroborando a importância dessa molécula para a caracterização de espécies de café. Além disso, outros íons também foram selecionados, correspondendo a compostos já mencionados na literatura como discriminantes para espécies de café, tais como: ácido quínico (*m/z* 423, $[2M+K]^+$), glicose (*m/z* 219, $[M+K]^+$), sacarose (m/z 381, $[M+K]^+$), ácido ferúlico (m/z 395, $[M+K]^+$), ácido feruloilquínico (*m/z* 407, $[M+K]^+$) e ácido acetil-cafeoquínico (*m/z* 435, $[2M+K]^+$) [22,23]. Em geral, diferentes açúcares e ácidos clorogênicos foram importantes para discriminar essas espécies. Algumas variáveis selecionadas, atribuídas aos íons de *m/z* 395, 407 e 435, estão relacionadas a ácidos clorogênicos específicos, apresentando coeficientes de regressão positivos no modelo. Essa observação é consistente com os teores mais elevados desses compostos fenólicos no café Robusta. Assim, eles podem ser considerados marcadores químicos para a diferenciação entre espécies Robusta e Arabica [11,12]. Na verdade, esse é um padrão multivariado de marcadores químicos.

5.4 Fusão de Dados TXRF e NIR

Para desenvolvimento dos modelos de fusão de dados utilizando TXRF e NIR, as matrizes foram, inicialmente, pré-processadas individualmente (apenas as amostras de torra leve e média foram utilizadas). Os dados de TXRF foram autoescalados e, para os dados de NIR, foi aplicada MSC seguida por centragem na média. Para a fusão de dados de nível baixo, após essa etapa, as matrizes foram concatenadas e o autoescalamento foi aplicado. Para o nível médio, modelos foram construídos com as matrizes isoladas, os escores mais significativos foram extraídos de cada modelo individual, concatenados e usados para a construção dos modelos de fusão de dados.

Durante a etapa de otimização dos modelos, nenhuma amostra foi classificada como *outlier* e retirada do conjunto de dados. Em geral, os modelos utilizando fusão de nível baixo tiveram desempenho superior ao de nível médio. Dessa forma, a Tabela 7 contém os parâmetros estatísticos aplicando GA e OPS para os modelos de nível baixo.

Tabela 7. Parâmetros estatísticos estimados para os modelos de fusão de dados de nível baixo utilizando dados de TXRF e NIR

	NIR + TXRF		
	Completo	**OPS**	***GA***
n_{VL}	*5*	6	6
n_{VL} ***OPS***		15	
nVars	*1114*	55	32
RMSECV(%)	*4,9*	2,2	1,3
RMSEC (%)	*2,7*	1,5	0,9
Rc	*0,92*	0,98	0,99
Rcv	*0,75*	0,95	0,98
RMSEP (%)	*4,6*	1,8	1,5
Rp	*0,78*	0,98	0,98

Pela análise da Tabela 7, pode-se concluir que tanto a seleção de variáveis por GA como OPS melhoraram todos os parâmetros estatísticos, em comparação ao modelo que utilizou os dados completos. Na realidade, uma comparação entre os dois métodos indica que o GA forneceu todos os parâmetros estatísticos ligeiramente melhores que o OPS. É importante ressaltar que tanto GA como OPS diminuíram consideravelmente o número de variáveis (GA, de 1114 para 32). Tal fato é de extrema importância pois, além de tornar o modelo mais simples, permite que a interpretação do mesmo se torne mais fácil.

Com relação aos erros relativos, para o modelo GA-TXRF-NIR os valores médios para o conjunto calibração e validação foram, respectivamente, 9,1 e 7,7%. Tais valores indicam que o modelo tem alta exatidão, podendo ser aplicado. A Figura 23 representa os valores de erros relativos para todas as amostras.

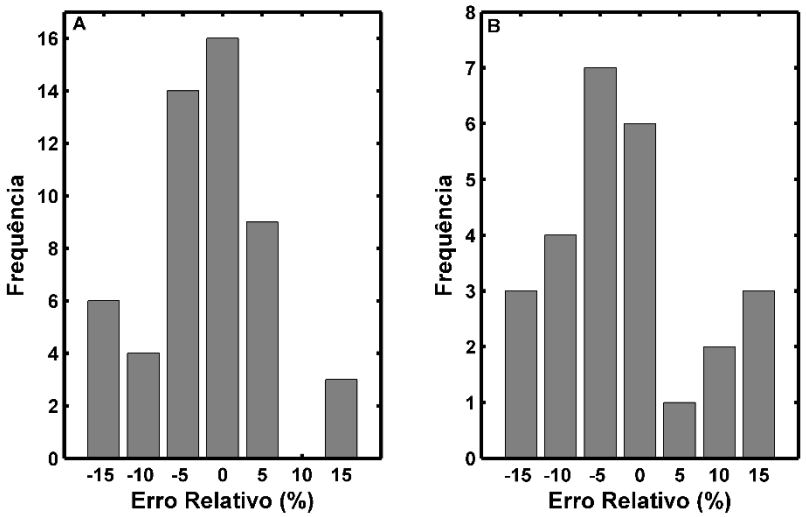

Figura 23. Erros relativos obtidos para as amostras individuais usando o modelo de fusão de dados de nível baixo construído com dados de TXRF e espectroscopia NIR: conjuntos de (A) calibração e (B) validação

O gráfico para avaliar a linearidade, valores medidos *versus* preditos, está esquematizado na Figura 24. Para o modelo GA-TXRF-NIR, coeficientes de correlação de 0,99 e 0,98 foram estimados para amostras de calibração e validação, respectivamente.

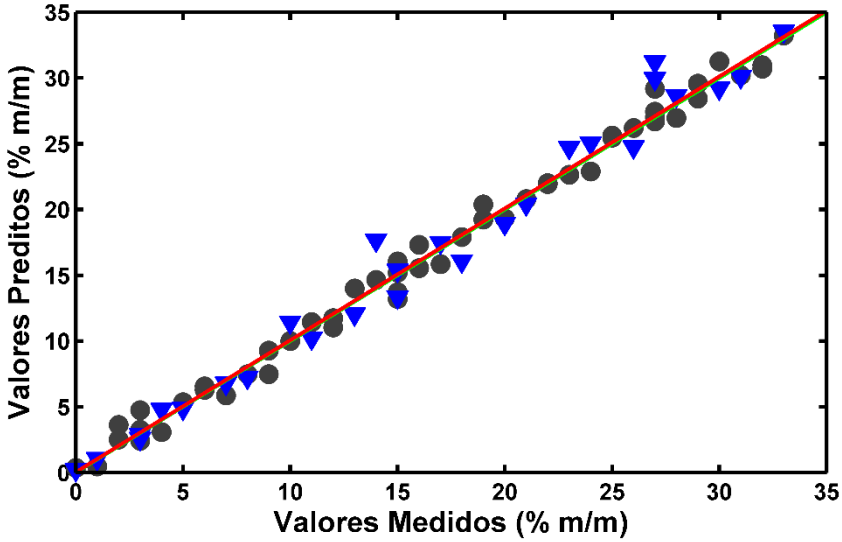

Figura 24. Valores medidos e preditos pelo modelo de fusão de dados de nível baixo GA-TXRF-NIR

5.4.1 Interpretação das Variáveis Selecionadas

Das 1114 variáveis originais, a fusão de dados de nível baixo GA-PLS selecionou 32 variáveis, 28 dos espectros NIR e 4 dos dados TXRF. As variáveis selecionadas a partir dos espectros de NIR são mostradas na Figura 25.

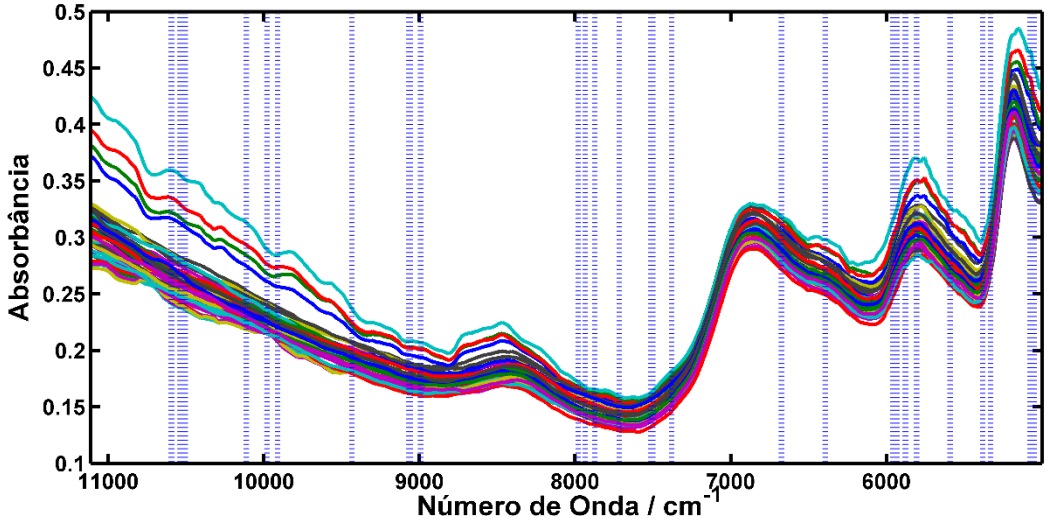

Figura 25. Variáveis selecionadas dos espectros NIR pelo modelo GA-NIR-TXRF

Observando a Figura 25, pode-se perceber as variáveis selecionadas se distribuem por toda a região espectral. Conforme esperado, regiões relacionadas à absorção da água (5200-5000 cm^{-1} e 7200-6800 cm^{-1}) foram selecionadas [143]. Tal região está mais relacionada com as variações espectrais devido aos diferentes graus do torra do que com o conteúdo de Robusta. Além disso, alguns números de onda na região entre 6100 e 4965 cm^{-1}, relacionados a vibrações de primeiro sobretom de ligações C-H, também foram selecionados. Vale ressaltar ainda que números de onda na região entre 9800-9000 cm^{-1}, relacionados aos sinais de terceiro sobretom de vibrações de ligações C-H e N-H, também foram selecionados. Em geral, a região compreendida entre 5797 e 5665 cm^{-1}, também selecionada pelo modelo, está relacionada com o conteúdo lipídico.

Diversas regiões espectrais selecionadas podem ser correlacionadas com compostos presentes no café. As regiões entre 5150–500 cm^{-1} e 6020–5925 cm^{-1}, por exemplo, podem ser associadas a vários compostos, tais como: ácidos clorogênicos, água, proteínas, carboidratos, cafeína, trigonelina, dentre outros. Além

disso, de forma mais específica, diversas variáveis situadas nas regiões entre 6666-5882 cm^{-1} e 5617-5405 cm^{-1} podem ser associadas a certas classes de substâncias submetidas a transformações químicas importantes durante o processo de torra, tais como: ácidos clorogênicos (5934 cm^{-1} - primeiro sobretom de vibrações de ligações C-H em estruturas aromáticas), carboidratos (5617 cm^{-1} - primeiro sobretom de estiramento de O-H, 5464 cm^{-1} – banda de combinação de estiramento de O-H e C-O) e aminoácidos (6622 cm^{-1} – primeiro sobretom de estiramento de N-H em proteína) [150,151].

Com relação ao TXRF, o modelo selecionou os seguintes elementos: K, Mn, Fe e Br. É importante ressaltar aqui o elemento Mn como importante marcador para diferenciar as duas espécies de café. Kivran *et. al* [149] utilizaram espectrometria de absorção atômica com forno de grafite, espectrometria de absorção atómica de chama e análise elementar para a determinação da origem de grãos de café Arabica. Entre os vários elementos investigados, constatou-se que o Mn é o mais adequado como indicador para esse propósito. Dessa forma, pode-se inferir que o conteúdo de Mn está mais relacionado ao café Arabica.

Neste contexto, é fundamental ressaltar a importância de um modelo de fusão de dados que encontre e utilize na modelagem, correlações atômico-moleculares. Na realidade**,** são poucos os artigos na literatura que exploraram essa possibilidade de interpretação aplicada a qualquer tipo de matriz.

5.5 Fusão de Dados MIR, PS-MS e TXRF

O processo para construção dos modelos utilizando dados MIR, PS-MS e TXRF foi similar aos demais modelos utilizando fusão de dados. O número total de variáveis foi 2216 (1801 do MIR, 401 PS-MS e 14 TXRF). As matrizes foram, individualmente, pré-processadas em uma etapa inicial (MIR: MSC e centragem na média, PS-MS: centragem na média; e TXRF: autoescalamento). Para a fusão de dados de nível baixo, as matrizes foram concatenadas e otimizadas por seleção de variáveis. Já na fusão de dados nível médio, modelos individuais foram construídos e os escores mais significativos de cada um foram extraídos.

Durante o processo de otimização, 4 amostras foram identificadas como *outliers* e retiradas dos conjuntos de calibração e validação. A Tabela 8 contém os parâmetros estatísticos obtidos para os modelos na fusão de dados de nível baixo. Conforme observado para os outros modelos, a técnica de fusão de dados de nível médio apresentou um desempenho um pouco pior.

Tabela 8. Parâmetros estatísticos para o modelo de fusão de dados de nível baixo usando MIR, PS-MS e TXRF

	MIR+PS-MS+TXRF		
	Completo	**OPS**	***GA***
n_{VL}	*5*	4	6
n_{VL} ***OPS***		20	
nVars	*2216*	125	165
RMSECV(%)	*4,0*	2,8	2,2
RMSEC (%)	*2,6*	1,9	1,1
Rc	*0,93*	0,96	0,99
Rcv	*0,84*	0,92	0,95
RMSEP (%)	*3,0*	2,5	1,3
Rp	*0,91*	0,94	0,98

Analisando a Tabela 8 é possível concluir que o GA mostrou o melhor desempenho, se comparado ao OPS e ao modelo completo. Apesar de o GA ter selecionado um número maior de VL, a análise da razão entre os valores de RMSEC e RMSEP indica que não houve sobreajuste no modelo. O GA selecionou um subconjunto de variáveis maior que o OPS, porém os valores de RMSEP, RMSECV E Rp justificam a escolha por esse método. Dessa forma, para o conjunto de dados utilizando MIR, PS-MS, TXRF e NIR, o modelo utilizando GA foi selecionado como o de melhor desempenho.

A média dos erros relativos, para o melhor modelo, foi de 12,4 e 18, 3% para os conjuntos de calibração e validação, respectivamente. Esses valores podem ser considerados altos, em uma primeira análise. Porém, observando o histograma de erros (Figura 26) percebe-se que a maioria das amostras está numa faixa de erro aceitável, e a distribuição se aproxima da normalidade. Amostras com baixa concentração de Robusta, apesar de terem valores de erro absoluto baixos, têm alto valor de erro relativo, colaborando para que a média seja alta.

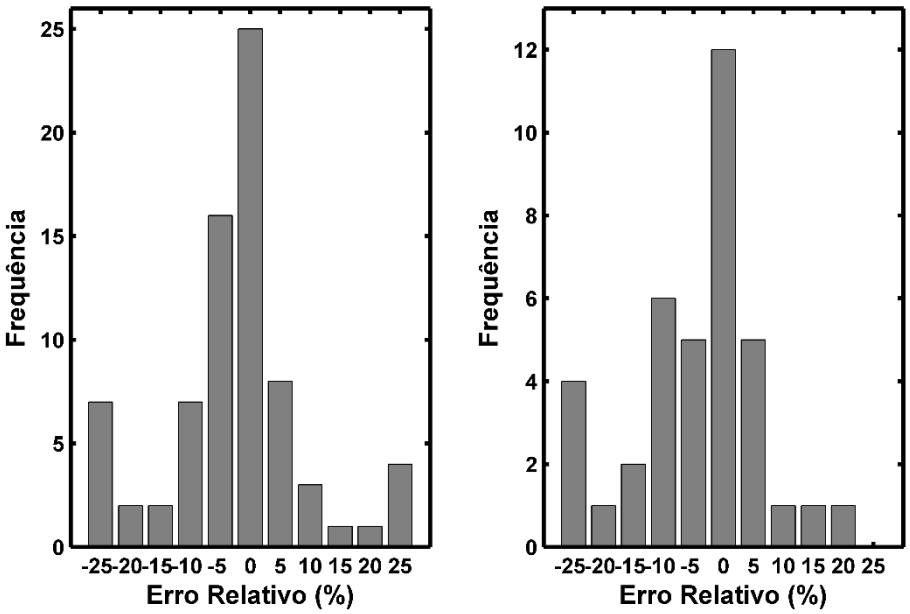

Figura 26. Histograma de erros relativos: conjuntos de (A) calibração e (B) validação

A linearidade deste modelo pode ser averiguada pelo gráfico dos valores medidos *versus* preditos (Figura 27). Para o melhor modelo, todos os valores de R foram superiores a 0,95.

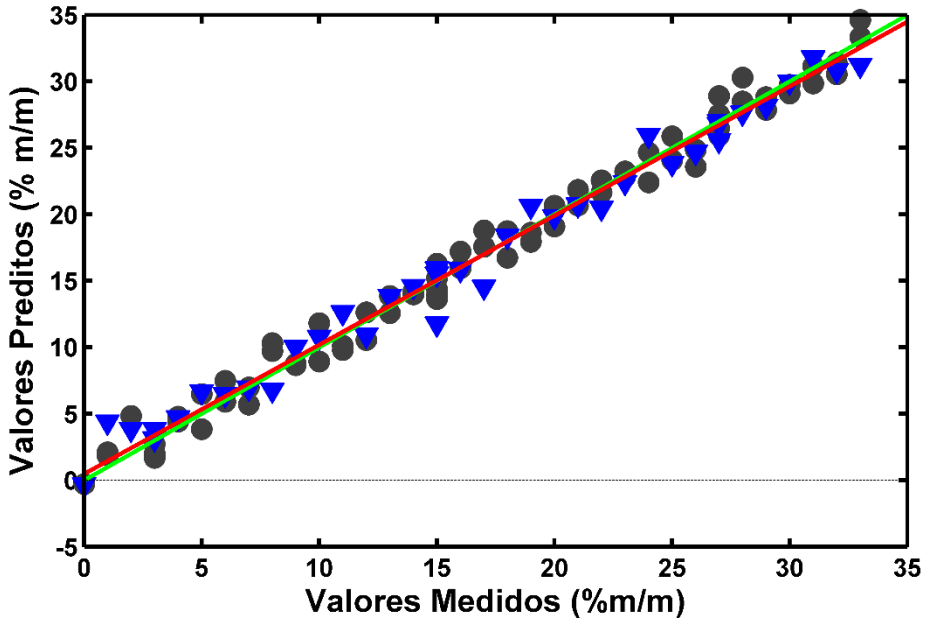

Figura 27. Valores medidos e preditos para os conjuntos: calibração (círculos) e validação (triângulos)

5.5.1 Interpretação das Variáveis Selecionadas

O algoritmo GA selecionou um total de 165 variáveis. Destas, 131 são provenientes dos espectros MIR, 31 dos espectros de PS-MS e 3 dos dados TXRF. A Figura 28 mostra as variáveis selecionadas (GA) dos dados provenientes dos espectros MIR.

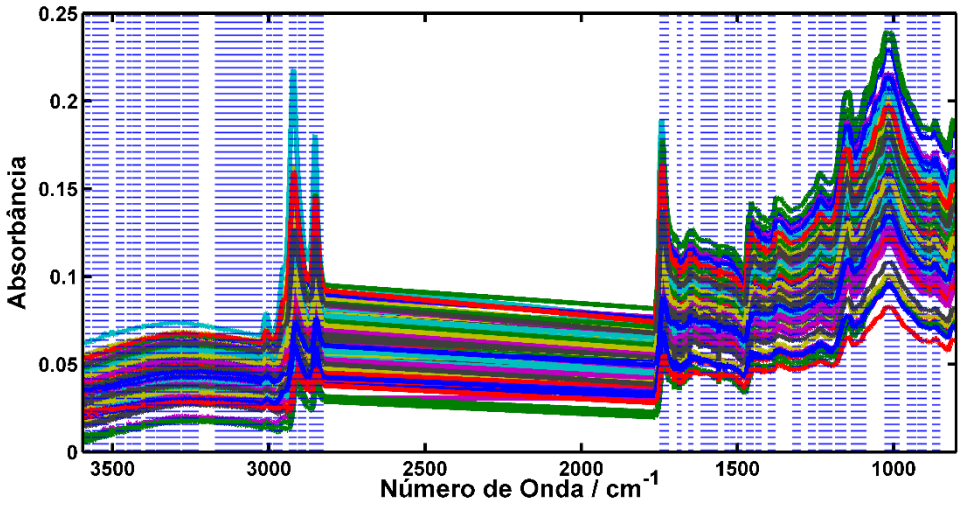

Figura 28. Variáveis selecionadas nos espectros MIR pelo modelo de fusão de dados de nível baixo usando MIR, PS-MS e TXRF

Uma análise da Figura 28 permite concluir que as variáveis selecionadas pelo modelo construído com dados das três técnicas instrumentais são bastante similares às variáveis selecionadas em outros modelos mais simples. As duas bandas estreitas presentes na faixa de 3000-2800 cm^{-1} têm sido relatadas para amostras de café torrados, sendo usualmente associadas com cafeína em bebidas [16]. Além disso, a banda centrada em 1743-1741 cm^{-1} foi atribuída à vibração da ligação carbonila em lipídios ou em ésteres alifáticos. Em 1543 cm^{-1}, acredita-se que a banda é associada ao estiramento C=C, presente em anéis nitrogenados, tais como os de cafeína e trigonelina, ambas presentes em quantidades significativas nos cafés. As bandas na região de 1150-900 cm^{-1} são associadas a carboidratos. Já a banda entre 1161 e 1153 cm^{-1} está relacionada aos ácidos clorogênicos, também presentes no café. A região entre 1800-800 cm^{-1}, que contém informações de *fingerprint* importantes para a caracterização de cafés, também foi selecionada [143,144].

Para os dados provenientes do PS-MS, a Figura 29 é uma representação do vetor regressão para as variáveis selecionadas.

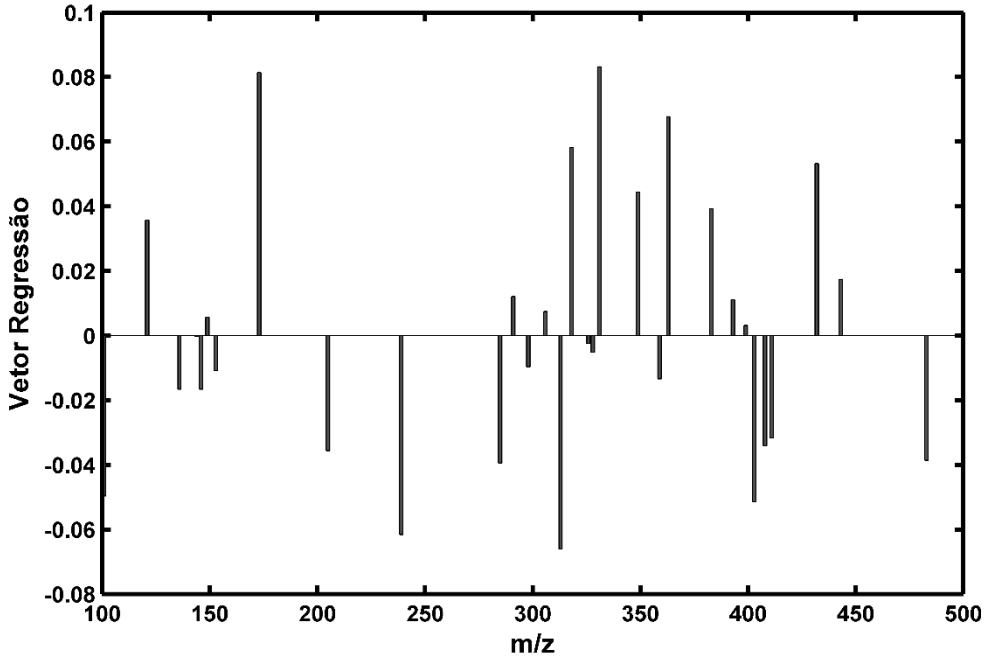

Figura 29. Vetor de regressão com as variáveis selecionadas nos espectros PS-MS pelo modelo de fusão de dados de baixo nível e usando MIR, PS-MS e TXRF

Os sinais em m/z 144 e 383, apesar de não identificados, já foram mencionados em outros artigos como associados a importantes compostos para separação de cafés Robusta e Arabica [145]. Ácidos carboxílicos formados no processo de torra, tais como ácidos octanóicos e decanóicos (m/z 145 e m/z 173) também foram selecionados [145]. Outros compostos importantes também foram selecionados: m/z 408 ($[M+K]^+$, ácido feruloilquínico [9]), *m/z* 317 (ácido linolênico $[M+K]^+$) e m/z 435 (ácido acetil-cafeoquiníco, $[2M + K]^+$). Em geral, ao se comparar as varáveis selecionadas por esse modelo com as selecionadas pela fusão de dados PS-MS/MIR, pode-se concluir que, para o segundo modelo, um número consideravelmente maior de variáveis foram selecionadas. Para o modelo PS-MS/MIR, compostos de diferentes classes foram selecionados. Já para o modelo MIR/PS-MS/TXRF, percebe-se que a

maioria das variáveis selecionadas podem ser classificadas como ácidos clorogênicos.

Para os dados provenientes de TXRF, os seguintes elementos foram selecionados: K, Br, Rb. Nesse contexto, é importante ressaltar o papel do Rb. Apesar de não encontrar referências na literatura, o Rb possui, entre os elementos do TXRF, o maior valor de vetor de regressão, em módulo. Tal fato pode indicar o metal Rb como um importante marcador para a diferenciação entre as espécies de café.

5.6 Modelo de fusão de dados MIR, PS-MS, NIR e TXRF

Além dos modelos já mencionados, modelos de fusão de dados contendo os dados oriundos de todas as técnicas utilizadas neste trabalho foram construídos. Para isso, os dados provenientes de cada técnica foram pré-processados separadamente. As estratégias de fusão de dados nos níveis baixo e médio foram aplicadas. Neste caso, como em todos os outros, a estratégia de nível baixo mostrou um melhor desempenho.

Para o melhor modelo, nenhuma amostra foi detectada nos conjuntos de calibração e validação como *outlier*. A Tabela 9 contém os parâmetros estatísticos para os modelos completo e utilizando GA e OPS.

Tabela 9. Parâmetros estatísticos para o modelo de fusão de dados de nível baixo usando todas as técnicas

	MIR+PS-MS+NIR+TXRF		
	Completo	**OPS**	***GA***
n_{VL}	*5*	4	6
n_{VL} ***OPS***		13	
nVars	*3316*	480	228
RMSECV(%)	*3,9*	2,5	1,9
RMSEC (%)	*1,8*	1,0	0,7
Rc	*0,96*	0,98	0,99
Rcv	*0,84*	0,94	0,97
RMSEP (%)	*3,2*	2,0	1,7
Rp	*0,90*	0,97	0,97

Analisando a Tabela 9, pode-se perceber que tanto o OPS como o GA melhoraram consideravelmente os parâmetros estatísticos dos modelos, diminuindo o número de variáveis a menos que 10% do total de variáveis originais. Uma análise mais criteriosa permite indicar o GA como método de melhor desempenho. Os valores de RMSEC, RMSECV e RMSEP foram reduzidos, em média, a cerca de 50% dos valores fornecidos pelo modelo que utilizou os espectros inteiros.

O histograma de erros relativos (%) obtidos para todas as amostras utilizando seleção de variáveis por GA é mostrado na Figura 30. A média dos erros relativos para o melhor modelo foi de 6,3 e 14,0% para os conjuntos de calibração e validação, respectivamente. Como para todos os modelos construídos previamente, erros relativos altos foram encontrados para amostras na faixa de 0-5% de café Robusta.

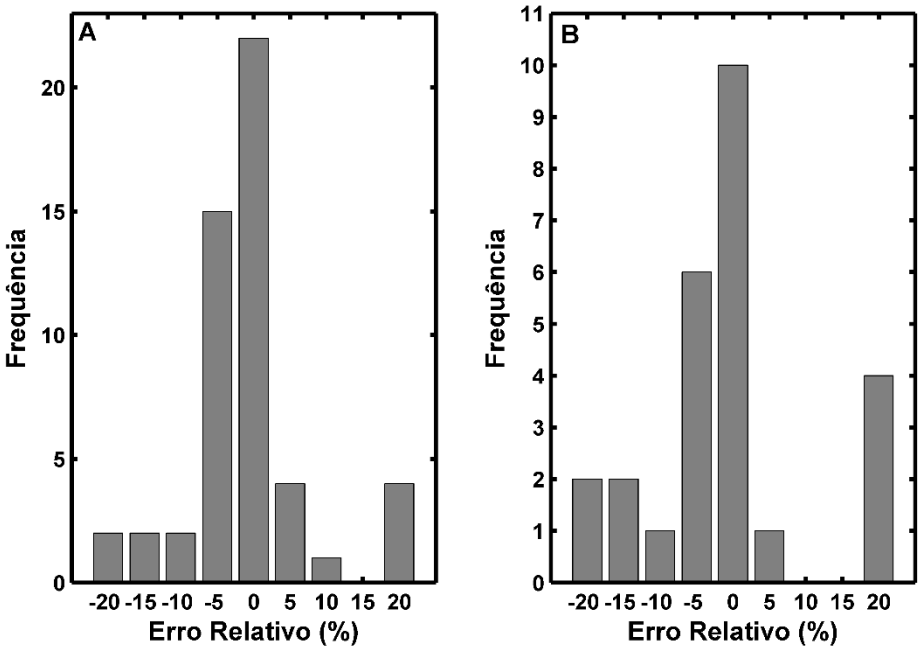

Figura 30. Histogramas de erros relativos para os conjuntos de (A) calibração e (B) validação

O gráfico para avaliar a linearidade deste modelo está esquematizado na Figura 31. Para o melhor modelo, coeficientes de correlação de 0,99 e 0,97 foram estimados para amostras de calibração e validação, respectivamente.

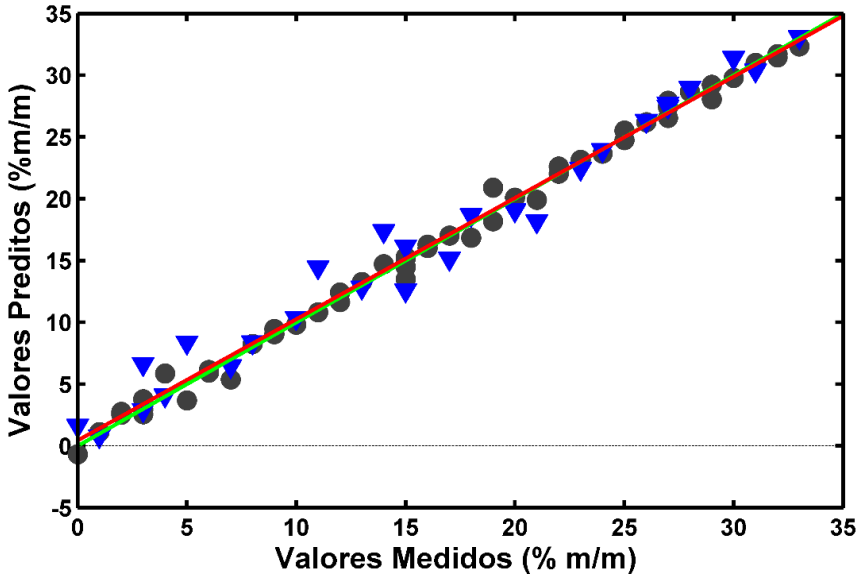

Figura 31. Valores medidos e valores preditos para o modelo de fusão de dados de nível baixo usando todas as quatro técnicas analíticas

5.6.1 Interpretação das Variáveis Selecionadas

O método GA selecionou 120 variáveis dos espectros MIR, 24 dos dados de PS-MS, 80 dos espectros NIR e 4 dos dados de TXRF, totalizando 228 variáveis. As variáveis selecionadas nos espectros de infravermelho médio estão destacadas na Figura 32. Vale ressaltar que as variáveis selecionadas aqui são bastante similares às variáveis já selecionadas em outros modelos apresentados anteriormente.

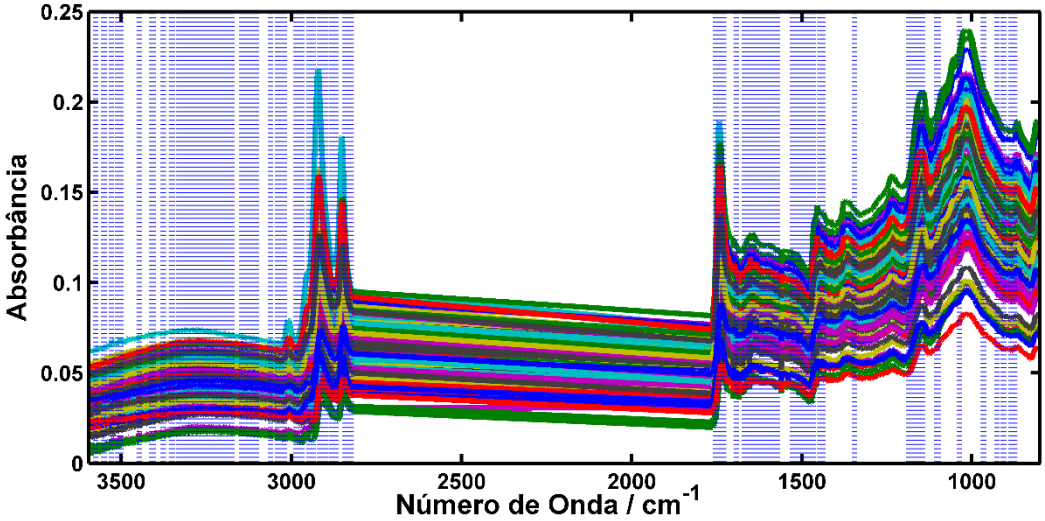

Figura 32. Variáveis selecionadas nos espectros MIR pelo modelo de fusão de dados de nível baixo otimizado por GA e usando todas as quatro técnicas

As variáveis selecionadas provenientes do MIR estão mostradas na Figura 35. A grande maioria destas também foi selecionada em outros modelos prévios, e suas atribuições já foram discutidas. Em geral, a região entre 3000-2800 cm^{-1} foi selecionada, sendo atribuída a vibrações de estiramento de ligações C-H de grupos metila, podendo estar relacionada ao conteúdo de cafeína na matriz [141]. Muitas variáveis também foram selecionadas na região da impressão digital, entre 1700-900 cm^{-1}, relacionadas ao conteúdo de carboidratos [140]. O pico centrado em 1742 cm^{-1}, selecionado para todos os modelos, é atribuído à vibração de carbonila de lipídios ou ésteres alifáticos [18]. A região entre 1400-900 cm^{-1} é caracterizada por vibrações de vários tipos de ligações, como C-H, C-O e C-N.

As variáveis selecionadas para o conjunto de dados provenientes de espectroscopia NIR (Figura 33) também são em sua maioria as mesmas selecionadas por outros modelos já discutidos. Regiões relacionadas a absorção de moléculas de água (5200-5000 cm^{1} e 7200-6800 cm^{1}) foram selecionadas [143]. A região entre

9800-9000 cm^{-1}, relacionada ao terceiro sobretom de vibrações de ligações C-H e N-H [143], também foi selecionada. Conforme já observado, diversas regiões espectrais selecionadas podem ser correlacionadas com compostos presentes no café, tais como: 5150–500 cm^{-1} e 6020–5925 cm^{-1}, associadas a ácidos clorogênicos, água, proteínas, carboidratos, cafeína, trigonelina, dentre outros.

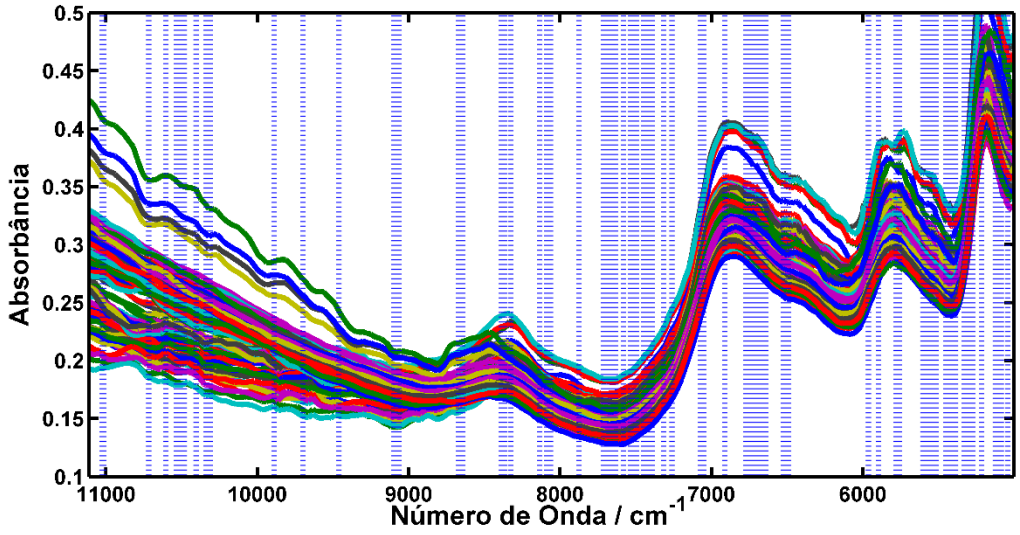

Figura 33. Varáveis selecionadas nos espectros NIR pelo modelo de fusão de dados de nível baixo otimizado por GA e usando todas as quatro técnicas

A Figura 34 é uma representação do vetor de regressão para as variáveis selecionadas a partir dos dados provenientes do PS-MS.

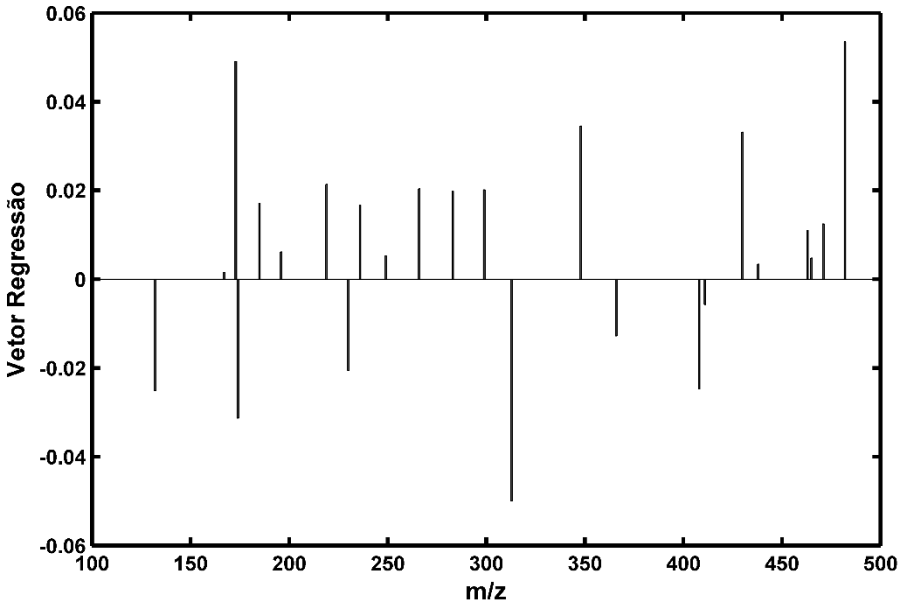

Figura 34. Vetor dos coeficientes de regressão (PS-MS) com as variáveis selecionadas por GA no modelo de fusão de dados construído com todas as quatro técnicas

A análise do vetor de regressão permite ampliar a possibilidade de interpretação dos modelos. O sinal de *m/z* 132, apesar de não identificado, já foi mencionado em outros artigos como importante composto para separação de cafés Robusta e Arabica [145]. O sinal de *m/z* 408 correspondente ao aduto $[M+K]^+$ do ácido feruloilquínico [9], também foi selecionado; os sinais de *m/z* 317 (ácido linolênico $[M+K]^+$), *m/z* 219 (glicose, $[M+K]^+$) e *m/z* 435 (ácido acetil-cafeoquínico, $[2M + K]^+$) também foram selecionados. Em geral, o GA selecionou um conjunto de variáveis bastante similar aos conjuntos selecionados nos modelos anteriores. Essas variáveis foram atribuídas a açúcares, ácidos clorogênicos, entre outros. Muitos sinais selecionados nos dados de massas não foram identificados/atribuídos.

Para os dados provenientes de TXRF, os seguintes elementos foram selecionados por este último modelo, Ti, Mn, Br e Rb. Vale lembrar que o Rb foi selecionado para, praticamente, todos os modelos apresentados anteriormente que

envolveram dados de TXRF. Isso indica que tal elemento pode ser fundamental para a diferenciação entre cafés Arabica e Robusta. O Mn também foi selecionado em outros modelos e já foi mencionado em outros trabalhos.

Com relação a trabalhos já presentes na literatura envolvendo o uso de fusão de dados na análise de misturas de espécies de cafés, é importante ressaltar, nesse momento, o trabalho de Dankowska e colaboradores [132]. Neste artigo, os autores combinaram espectrofluorimetria e espectroscopia de absorção molecular na região do UV-Vis para quantificação de *blends* de cafés Arabica e Robusta, utilizando fusão de nível médio e baixo. Em linhas gerais, modelos PCR foram construídos com 115 amostras. Porém, este artigo apresenta muitas deficiências, principalmente na parte de desenvolvimento dos modelos. De forma geral, não houve separação dos conjuntos calibração e validação, a validação cruzada foi realizada de duas formas (*leave-one-out* e reamostragem), embora não fique claro, no texto, qual das duas formas forneceu o menor valor de RMSECV. Além disso, o artigo não forneceu nenhum gráfico de linearidade ou maiores detalhes sobre os resultados para a previsão das amostras. Finalmente, o ponto mais crítico, o melhor modelo utilizando fusão de dados nível médio foi construído com 38 componentes principal (a maioria dos modelos testados tinha mais de 30 componentes). Para esse conjunto de dados, um modelo com um número tão grande de componentes principais (ou VL no caso de modelos PLS) é um verdadeiro absurdo, indicando alto grau de sobreajuste nos dados (e falta de conhecimento sobre métodos de calibração multivariada).

5.7 Comparação Entre os Modelos

Os valores de RMSEC, RMSECV e RMSEP dos modelos desenvolvidos são mostrados na Tabela 10.

Tabela 10. Comparação de parâmetros estatísticos entre os modelos desenvolvidos.

	MIR	TXRF	PS-MS/MIR	TXRF/NIR	MIR/PS-MS/TXRF	TODAS
RMSEC	1,2	1,5	1,0	0,9	1,1	0,7
RMSECV	2,4	2,0	2,2	1,3	2,2	1,9
RMSEP	1,8	1,2	1,7	1,5	1,3	1,7

Uma análise dos parâmetros estatísticos de todos os modelos permite concluir que o modelo utilizando apenas TXRF teve o melhor desempenho, se for considerado apenas o valor de RMSEP. Uma análise dos valores de RMSECV e RMSEC indica que o modelo utilizando todos as técnicas analíticas apresentou os menores erros. Porém, a escolha dos melhores modelos deve ser baseada em uma série de aspectos, dentre eles, a simplicidade na obtenção dos dados, o tempo de análise e o custo.

Avaliando os resultados da Tabela 10, pode-se afirmar que todos os modelos tiveram um bom desempenho, podendo ser aplicados na prática. O método utilizando MIR, além de ter incorporado a variância de todos os níveis de torra, permite aquisição rápida dos espectros, com um mínimo preparo de amostra. Tais características (além de bons parâmetros estatísticos) conferem a este método maior potencial de aplicação em uma rotina laboratorial (a utilização do TXRF demanda várias etapas de preparo de amostra, o que dificulta sua aplicação). Dessa forma, se forem considerados apenas os métodos desenvolvidos com apenas uma técnica, a espectroscopia no infravermelho próximo é a mais indicada.

Uma comparação entre os modelos de fusão de dados indica que os que foram construídos utilizando PS-MS/MIR e MIR/PS-MS/TXRF apresentaram os melhores

resultados. Os demais modelos, apesar de mostrarem menores valores de erros, não modelaram os três níveis de torra. Uma comparação entre os dois modelos mencionados mostra que não há grande diferença entre eles. Como o PS-MS/MIR utilizou apenas duas técnicas (o que elimina TXRF, técnica laboriosa) e possui um poder alto de interpretação do sistema (duas técnicas que fornecem informações diferentes e complementares), esse modelo foi considerado o melhor.

5.8 Validação Analítica Multivariada

Além de todos os parâmetros já apresentados para avaliar a qualidade dos modelos, uma validação analítica multivariada foi realizada para os modelos com os melhores desempenhos. Nesse caso específico, a espectroscopia no infravermelho médio forneceu o melhor modelo, se considerados apenas os conjuntos de dados provenientes das técnicas isoladas. Para a fusão de dados, a junção entre MIR e PS-MS foi selecionada como a de melhor desempenho.

5.8.1 Modelos com Dados de Infravermelho Médio

A Tabela 11 contém as figuras de mérito (FOM) calculadas para os modelos com melhores desempenhos, após a aplicação dos métodos de seleção de variáveis.

Tabela 11. Parâmetros para avaliar as principais FOM para os melhores modelos utilizando espectroscopia MIR.

FOM	Parâmetros	Leve	Médio	Forte	Robusto
		GA	**GA**	**GA**	**OPS**
Faixa de Trabalho	**-**	0-33 %	0-33 %	0-33 %	0-33 %
	Durbin-Watson	2,21	1,63	1,74	2,24
Linearidade	**Inclinação[a]**	0,99±0,09%	0,99±0,02%	0,98±0,03%	0,98±0,02%
	Intercepto[a]	0,03±0,17	0,12±0,33	0,27±0,50	0,32±0,45
Sensibilidade[b]	**-**	0,0066	0,0075	0,0068	0,14
Sensibilidade Analítica	**Γ**	2,0 ($\%^{-1}$)	2,2 ($\%^{-1}$)	1,9 ($\%^{-1}$)	41,2 ($\%^{-1}$)
	γ^{-1}	0,5%	0,5%	0,5%	0,02%
Ruído	**Desvio Padrão Agrupado (Branco)**	0,0033	0,0034	0,0035	0,0034
Exatidão	**Erro Médio Relativo**	8,6%	7,5%	6,4%	12,0%
Bias	***Bias* ± DPR**	0,5±1,0%	-0,3±1,3	0,1±0,9	-0,3±1,8
Precisão	**DPR repetibilidade (%)**	9,2/2,9/0,7	8,2/2,2/3,9	9,4/5,4/3,3	8,8/5,1/3,9
RPDcal/val	**-**	8,4/7,0	5,6/7,0	4,3/10,0	4,6/4,0

a Valores para a linha ajustada às amostras de calibração

b Valores expressos como a razão entre a absorbância e a porcentagem

c Valores nos níveis baixo (3,0%), médio (15,0%) e alto (27,0%)

Em geral, a exatidão de modelos de calibração multivariada é avaliada com base nos valores de RMSEC e no RMSEP (já discutidos na Tabela 3). Para o modelo robusto, esses valores foram, respectivamente, 1,1 e 1,8%. Além disso, também é importante observar os erros relativos individuais para as previsões das amostras. Os histogramas dos erros relativos para amostras nos conjuntos de validação para todos os quatro modelos j foram mostrados na Figura 15 e discutidos. Como mencionado, os erros acima de ± 20% foram observados para um pequeno número de amostras em baixas concentrações de café Robusta. Outro parâmetro importante é o RPD, usado para avaliar a veracidade dos modelos de calibração multivariada em termos absolutos. Modelos com RPD acima de 2,5 são considerados adequados para uso em controle de qualidade [150]. Em relação à precisão, o modelo robusto apresentou um DPR de 4,6 % para calibração e 4,0 % para validação, confirmando seu bom desempenho.

A linearidade dos modelos pode ser avaliada analisando o ajuste dos valores de referência *versus* valores previstos (Figura 16). Para estes dados, nenhuma tendência sistemática foi observada para as distribuições residuais. Os coeficientes de correlação para todas as curvas foram todos acima de 0,94 (Tabela 3). No entanto, coeficientes de correlação próximos de 1,0 não são suficientes para garantir a linearidade, sendo necessária a verificação dos comportamentos aleatórios dos resíduos. Esse comportamento foi verificado aplicando-se sequencialmente os testes RJ, BF e DW (Ryan-Joiner, Brown-Forsythe, Durbin-Watson, respectivamente), para confirmar num nível de confiança de 95% a normalidade, homocedasticidade e independência dos resíduos, respectivamente. Para o último teste, todos os valores de DW estimados (Tabela 13) estavam dentro do intervalo de aceitação, entre 1,50 e 2,50 [136]. A precisão no nível de repetibilidade também foi avaliada através do DPR

de triplicatas obtidas em três diferentes níveis de concentração, um baixo (3,0%), um médio (15,0%) e um alto (27,0%). Considerando as avaliações de veracidade, linearidade e precisão, as faixas de trabalho dos métodos foram estabelecidas entre 0,0 e 33,0%.

A sensibilidade foi estimada com base na norma Euclidiana dos vetores de regressão [139]. Esta é uma FOM dependente da técnica analítica empregada e da matriz analisada. Assim, a sensibilidade analítica (γ) foi estimada dividindo a sensibilidade por uma estimativa do ruído instrumental, que foi obtida calculando o desvio padrão agrupado de 10 espectros replicados da célula de ATR vazia. O inverso desta FOM, γ^{-1}, fornece a diferença que pode ser discriminada pelos métodos considerando o ruído experimental aleatório como a única fonte de erro. Além disso, esse valor indica a limitação do número de algarismos significativos usados para expressar os resultados. Apesar da estimativa de γ^{-1} para o modelo robusto (0,02%), o uso de apenas uma casa decimal foi considerado mais coerente para expressar os resultados para todos os modelos.

Para finalizar, o viés (*bias*) foi estimado com base apenas nas amostras do conjunto de validação. Com os valores de viés estimados e seus desvios-padrão (SDV), os testes em nível de confiança de 95% evidenciaram a ausência de erros sistemáticos em todos os modelos.

5.8.2 Modelo de Fusão de Dados MIR e PS-MS

A Tabela 12 contém as FOM calculadas para o modelo MIR/PS-MS, utilizando OPS e fusão de dados de nível baixo.

Tabela 12. Parâmetros para avaliar as principais FOM para o modelo de fusão de dados MIR/PS-MS de nível baixo utilizando OPS

FOM	Parâmetro	Robusto
		OPS
Faixa de Trabalho	**-**	0-33 %
Linearidade	**Durbin-Watson**	1,76
	Inclinação[a]	0,92±0,02%
	Intercepto[a]	1,3±0,7
Sensibilidade[b]	**-**	0,19
Sensibilidade Analítica	**Γ**	39,2 ($\%^{-1}$)
	γ^{-1}	0,02 %
Ruído	**Desvio Padrão Agrupado (Branco)**	0,0046
Exatidão	**Erro Médio Relativo**	10,21 %
Bias	***Bias* ± DPR**	0,2± 0,9
RPDc/RPDv		4,0/5,5

Conforme já mencionado, a exatidão pode ser avaliada por meio do RMSEC e RMSEP (Tabela 6), assim como através dos erros relativos individuais para cada previsão de amostra (Figura 26). Erros considerados altos (acima de 20% a 30%) foram estimados para um pequeno número de amostras, todas elas na menor faixa de concentração de café Robusta.

Coeficientes de correlação, 0,99 e 0,98, foram estimados para amostras de calibração e validação, respectivamente. Além disso, nenhuma tendência sistemática na distribuição dos resíduos foi observada. A fim de confirmar a aleatoriedade desses resíduos, sua normalidade, homocedasticidade e independência foram verificadas, respectivamente, aplicando sequencialmente os testes RJ, BF e DW. Todos esses três testes garantiram o comportamento aleatório dos resíduos de ajuste com nível de confiança de 95%. Particularmente, um resultado de DW de 1,76 está dentro do

intervalo de aceitação (1,50-2,50) [136], confirmando a ausência de autocorrelação dos resíduos.

A sensibilidade do método foi calculada com base na norma Euclidiana dos vetores de regressão. Como a sensibilidade depende da técnica analítica e da matriz estudada, foi estimada uma FOM mais robusta, sensibilidade analítica (**γ**). Uma estimativa conjunta do ruído instrumental (espectros do branco obtidos por ATR-FTIR e PS-MS e organizados de forma concatenada) foi usada neste cálculo. O inverso deste valor, $\boldsymbol{\gamma}^{-1}$, foi estimado em 0,02%, o que corresponde à concentração mínima distinguível pelo método. Além disso, esse valor define a limitação do número de dígitos significativos usados para expressar os resultados. No entanto, foi considerado mais realista expressar todas as previsões com apenas uma casa decimal.

O viés, definido como a média aritmética dos erros de predição, foi estimado apenas para as amostras de validação, juntamente com os desvios padrão dos erros de validação (SDV). Com esses dois parâmetros, um teste t com nível de confiança de 95% corroborou a ausência de erros sistemáticos no modelo. Finalmente, o RPD foi estimado para os conjuntos de calibração e validação. Este é um FOM para avaliar a exatidão de modelos de calibração multivariados em termos absolutos, independente do intervalo analítico. Valores de RPD de 4,0 (calibração) e 5,5 (validação) foram obtidos, confirmando o bom desempenho do modelo. Valores de RPD superiores a 2,4 são considerados desejáveis para bons modelos de calibração, enquanto RPD menores que 1,5 indicam modelos que não possuem qualidade suficiente para serem utilizados [135].

6. Conclusão

A análise dos parâmetros estatísticos para todos os modelos indicou que tanto o método OPS como o GA diminuíram significativamente os erros de validação cruzada e de previsão. Dessa forma, a aplicação de métodos de seleção de variáveis contribuiu para a construção de modelos mais robustos e simples de interpretar. Para a maioria dos modelos construídos, o método GA foi o que apresentou melhor desempenho.

Em geral, os modelos apresentaram boa capacidade de previsão quando aplicados para prever a composição de *blends*, podendo ser perfeitamente aplicados, substituindo análises morosas, dispendiosas e ambientalmente não amigáveis, por uma análise extremamente rápida, de baixo custo e que não exige gasto de reagentes em análises por via úmida.

Como técnica isolada, o modelo com melhor desempenho foi obtido com a utilização de espectroscopia MIR. Para modelos de fusão de dados, a junção de PS-MS e MIR forneceu modelos com erros menores.

Os modelos construídos foram interpretados e, de forma geral, compostos específicos presentes no café foram fundamentais para diferenciar as espécies, tais como trigonelina, açúcares e ácidos clorogênicos. Para os dados de constituição elementar, Mn e Rb destacaram-se os elementos como possíveis marcadores das espécies. Os melhores modelos (MIR e MIR-PSMS) foram validados e figuras de mérito adequadas foram estimadas, corroborando sua exatidão, linearidade, sensibilidade e ausência de viés.

A metodologia de fusão de dados que modela/explora correlações elementares–moleculares é uma contribuição importante deste trabalho, pouco presente na literatura. Esta abordagem pode ser aplicada a outras matrizes,

aumentando o potencial de construção e, principalmente, interpretação de modelos multivariados.

7. Referências Bibliográficas

[1] D. F. Barbin, A. L. de S. M. Felicio, D. W. Sun, S. L. Nixdorf, e E. Y. Hirooka, “Application of infrared spectral techniques on quality and compositional attributes of coffee: An overview”, *Food Res. Int.*, 61, p. 23–32, 2014.

[2] H. Grinshpun, “Deconstructing a global commodity: Coffee, culture, and consumption in Japan”, *J. Consum. Cult.*, vol. 14, p. 343–364, 2013.

[3] I. Ludwig, M. Clifford, H. Ashihara, e A. Crozier, “Coffee: biochemistry and potential impact on health”, *Food Funct.*, vol. 29, p. 1695–1717, 2014.

[4] CONAB, “Companhia Nacional de Abastecimento”. [Online]. Available at: http://www.conab.gov.br. [Acessado: 20-jun-2018].

[5] ABIC, "Associação Brasileira da Indústria de café". [Online]. Available at: http://www.abic.com.br/. [Acessado: 01-jul-2018].

[6] A. P. Davis, R. Govaerts, D. M. Bridson, e P. Stoffelen, “An annotatated taxonomic of the genus coffea (Rubiaceae)”, *Bot. J. Linn. Soc.*, vol. 152, p. 465–512, 2006.

[7] FAOSTAT, “Food and Agriculture Organization of the United Nations”. [Online]. Available at: http://faostat3.fao.org/home/E. [Acessado: 20-jun-2018].

[8] ICO, “International Coffee Organization”, [Online]. Available at: http://www.ico.org/. [Acessado: 01-jul-2018].

[9] H. van der Vossen, B. Bertrand, e A. Charrier, “Next generation variety development for sustainable production of arabica coffee (Coffea arabica L.): a review”, *Euphytica*, vol. 204, p. 243–256, 2015.

[10] M. M. Costa e L. C. Trugo, “Determinação de compostos bioativos em amostras comerciais de café torrado”, *Quim. Nova*, vol. 28, p. 637–641, 2005.

[11] L. R. Cagliani, G. Pellegrino, G. Giugno, e R. Consonni, “Quantification of

Coffea arabica and Coffea canephora var. robusta in roasted and ground coffee blends", *Talanta*, vol. 106, p. 169–173, 2013.

[12] E. Schievano, C. Finotello, E. De Angelis, S. Mammi, e L. Navarini, "Rapid Authentication of Coffee Blends and Quantification of 16-O-Methylcafestol in Roasted Co ff ee Beans by Nuclear Magnetic Resonance", *J. Agric. Food Chem.*, vol. 62, p. 12309–12314, 2014.

[13] *AOAC INTERNATIONAL - Official Methods of Analysis of the Association of Official Analytical Chemists International*. 1995.

[14] R. M. R. Ribeiro, S. Fujii, M. Silva, A. B. Nogari, M. B. S. Scholz, E. Y. S. Ono, O. Kawamura, E.Y. Hirooka, "Espectrofluorimetria comparada à cromatografia líquida de alta eficiência na detecção de ocratoxina A em café", *Semin. Agrar.*, 29, p. 331–338, 2008.

[15] C. Pizarro, I. Esteban-Díez, e J. M. González-Sáiz, "Mixture resolution according to the percentage of robusta variety in order to detect adulteration in roasted coffee by near infrared spectroscopy", *Anal. Chim. Acta*, vol. 585, p. 266–276, 2007.

[16] N. Reis, A. S. Franca, e L. S. Oliveira, "Discrimination between roasted coffee, roasted corn and coffee husks by Diffuse Reflectance Infrared Fourier Transform Spectroscopy", *LWT - Food Sci. Technol.*, vol. 50, p. 715–722, 2013.

[17] N. Reis, B. G. Botelho, A. S. Franca, e L. S. Oliveira, "Simultaneous Detection of Multiple Adulterants in Ground Roasted Coffee by ATR-FTIR Spectroscopy and Data Fusion", *Food Anal. Methods*, vol. 10, p. 2700–2709, 2017.

[18] A. P. Craig, A. S. Franca, e L. S. Oliveira, "Discrimination between Immature and Mature Green Coffees by Attenuated Total Reflectance and Diffuse

Reflectance Fourier Transform Infrared Spectroscopy", *J. Food Sci.*, vol. 76, p. 1162–1168, 2011.

[19] D. Pacetti, E. Boselli, M. Balzano, e N. G. Frega, "Authentication of Italian Espresso coffee blends through the GC peak ratio between kahweol and 16-O-methylcafestol", *Food Chem.*, vol. 135, p. 1569–1574, 2012.

[20] G. Caprioli, M. Cortese, G. Cristalli, F. Maggi, L. Odello, M. Ricciutelli, G. Sagratini, V. Sirocchi, G. Tomassoni, S. Vittori, "Optimization of espresso machine parameters through the analysis of coffee odorants by HS-SPME-GC/MS", *Food Chem.*, vol. 135, p. 1127–1133, 2012.

[21] C. E. Mills, M. J. Oruna-Concha, D. S. Mottram, G. R. Gibson, e J. P. E. Spencer, "The effect of processing on chlorogenic acid content of commercially available coffee", *Food Chem.*, vol. 141, p. 3335–3340, 2013.

[22] R. Garrett, B. G. Vaz, A. M. C. Hovell, M. N. Eberlin, e C. M. Rezende, "Arabica and Robusta coffees: Identification of major polar compounds and quantification of blends by direct-infusion electrospray ionization-mass spectrometry", *J. Agric. Food Chem.*, vol. 60, p. 4253–4258, 2012.

[23] R. Garrett, C. M. Rezende, e D. R. Ifa, "Coffee origin discrimination by paper spray mass spectrometry and direct coffee spray analysis", *Anal. Methods*, vol. 5, p. 5944–5948, 2013.

[24] R. Hertz-Schünemann, T. Streibel, S. Ehlert, e R. Zimmermann, "Looking into individual coffee beans during the roasting process: Direct micro-probe sampling on-line photo-ionisation mass spectrometric analysis of coffee roasting gases", *Anal. Bioanal. Chem.*, vol. 405, p. 7083–7096, 2013.

[25] Y. B. Monakhova, W. Ruge, T. Kuballa, M. Ilse, O. Winkelmann, B. Diehl, F. Thomas, D. W. Lachenmeier, "Rapid approach to identify the presence of

Arabica and Robusta species in coffee using 1H NMR spectroscopy", *Food Chem.*, vol. 182, p. 178–184, 2015.

[26] V. A. Arana, J. Medina, R. Alarcon, E. Moreno, L. Heintz, H. Schafer, J. Wist, "Coffee's country of origin determined by NMR: The Colombian case", *Food Chem.*, vol. 175, p. 500–506, 2015.

[27] T. Wermelinger, L. D'Ambrosio, B. Klopprogge, e C. Yeretzian, "Quantification of the robusta fraction in a coffee blend via raman spectroscopy: Proof of principle", *J. Agric. Food Chem.*, vol. 59, p. 9074–9079, 2011.

[28] R. M. El-Abassy, P. Donfack, e A. Materny, "Discrimination between Arabica and Robusta green coffee using visible micro Raman spectroscopy and chemometric analysis", *Food Chem.*, vol. 126, p. 1443–1448, 2011.

[29] H. Wang, N. E. Manicke, Q. Yang, L. Zheng, R. Shi, R. G. Cooks, e Z. Ouyang, "Direct analysis of biological tissue by paper spray mass spectrometry", *Anal. Chem.*, vol. 83, p. 1197–1201, 2011.

[30] J. Deng e Y. Yang, "Chemical fingerprint analysis for quality assessment and control of Bansha herbal tea using paper spray mass spectrometry", *Anal. Chim. Acta*, vol. 785, p. 82–90, 2013.

[31] Z. Zhang, R. G. Cooks, e Z. Ouyang, "Paper spray: a simple and efficient means of analysis of different contaminants in foodstuffs", *Analyst*, vol. 137, p. 2556–2558, 2012.

[32] H. Wang, Y. Ren, M. N. McLuckey, N. E. Manicke, J. Park, L. Zheng, R. Shi, R. G. Cooks, e Z. Ouyang, "Direct quantitative analysis of nicotine alkaloids from biofluid samples using paper spray mass spectrometry", *Anal. Chem.*, vol. 85, p. 11540–11544, 2013.

[33] P. Wobrauschek, "Total reflection x-ray fluorescence analysis—a review", *X-*

Ray Spectrom., vol. 36, p. 289–300, 2007.

[34] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, e O. Busto, “Data fusion methodologies for food and beverage authentication and quality assessment - A review”, *Anal. Chim. Acta*, vol. 891, p. 1–14, 2015.

[35] T. Wermelinger, L. D’Ambrosio, B. Klopprogge, e C. Yeretzian, “Quantification of the robusta fraction in a coffee blend via raman spectroscopy: Proof of principle”, *J. Agric. Food Chem.*, vol. 59, p. 9074–9079, 2011.

[36] A. P. Davis, R. Govaerts, D. M. Bridson, M. Ruhsam, J. Moat, e N. A. Brummitt, “A Global Assessment of Distribution, Diversity, Endemism, and Taxonomic Effort in the Rubiaceae [1]”, *Ann. Missouri Bot. Gard.*, vol. 96, 68–78, 2009.

[37] A. S. Franca, J. C. F. Mendonça, e S. D. Oliveira, “Composition of green and roasted coffees of different cup qualities”, *LWT - Food Sci. Technol.*, vol. 38, p. 709–715, 2005.

[38] E. Bertone, A. Venturello, A. Giraudo, G. Pellegrino, e F. Geobaldo, “Simultaneous determination by NIR spectroscopy of the roasting degree and Arabica/Robusta ratio in roasted and ground coffee”, *Food Control*, vol. 59, p. 683–689, 2016.

[39] J. Shibao e D. H. M. Bastos, “Produtos da reação de Maillard em alimentos: Implicações para a saúde”, *Rev. Nutr.*, vol. 24, p. 895–904, 2011.

[40] U. Jumhawan, S. P. Putri, Yusianto, T. Bamba, e E. Fukusaki, “Quantification of coffee blends for authentication of Asian palm civet coffee (Kopi Luwak) via metabolomics: A proof of concept”, *J. Biosci. Bioeng.*, vol. 122, p. 79–84, 2016.

[41] M. C. Combes, T. Joët, e P. Lashermes, “Development of a rapid and efficient DNA-based method to detect and quantify adulterations in coffee (Arabica versus Robusta)”, *Food Control*, vol. 88, p. 198–206, 2018.

[42] I. Colzi, C. Taiti, E. Marone, S. Magnelli, C. Gonnelli, e S. Mancuso, “Covering the different steps of the coffee processing: Can headspace VOC emissions be exploited to successfully distinguish between Arabica and Robusta?”, *Food Chem.*, vol. 237, p. 257–263, 2017.

[43] J. Wang, S. Jun, H. C. Bittenbender, L. Gautz, e Q. X. Li, “Fourier transform infrared spectroscopy for kona coffee authentication”, *J. Food Sci.*, vol. 74, p. 385–391, 2009.

[44] L. C. A. Barbosa, *Espectroscopia no Infravermelho na caracterização de compostos orgânicos*. UFV: Viçosa, 2007.

[45] K. M. G. Lima, I. M. Raimundo, A. M. S. Silva, e M. F. Pimentel, “Sensores ópticos com detecção no infravermelho próximo e médio”, *Quim. Nova*, vol. 32, p. 1635–1643, 2009.

[46] P. A. Costa Filho e R. J. Poppi, “Aplicação de algoritmos genéticos na seleção de variáveis em espectroscopia no infravermelho médio. Determinação simultânea de glicose, maltose e frutose”, *Quim. Nova*, vol. 25, p. 46–52, 2002.

[47] L. F. C. Oliveira, “Espectroscopia Molecular”, *Cad. Temáticos Química Nov. Esc.*, vol. 4, p. 24–30, 2001.

[48] D. A. Skoog, D. M. West, F. J. Holler, e S. R. Crouch, *Fundamentos de Química Analítica*. São Paulo: Thomson, 2006.

[49] C. Pasquini, “Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications”, *J. Braz. Chem. Soc.*, vol. 14, p. 198–219, 2003.

[50] C. Pasquini, “Near infrared spectroscopy: A mature analytical technique with new perspectives – A review”, *Anal. Chim. Acta*, vol. 1026, p. 8–36, 2018.

[51] Y. Ozaki, “Near-infrared spectroscopy--its versatility in analytical chemistry.”, *Anal. Sci.*, vol. 28, p. 545–63, 2012.

[52] I. Esteban-Díez, J. M. González-Sáiz, e C. Pizarro, “An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS”, *Anal. Chim. Acta*, vol. 514, p. 57–67, 2004.

[53] S. Buratti, N. Sinelli, E. Bertone, A. Venturello, E. Casiraghi, e F. Geobaldo, “Discrimination between washed Arabica, natural Arabica and Robusta coffees by using near infrared spectroscopy, electronic nose and electronic tongue analysis”, *J. Sci. Food Agric.*, vol. 95, p. 2192–2200, 2015.

[54] K. Tolessa, M. Rademaker, B. De Baets, e P. Boeckx, “Prediction of specialty coffee cup quality based on near infrared spectra of green coffee beans”, *Talanta*, vol. 150, p. 367–374, 2016.

[55] K. M. Santos, M. F. V Moura, F. G. Azevedo, K. M. G. Lima, I. M. Raimundo, e C. Pasquini, “Classification of Brazilian Coffee Using Near-Infrared Spectroscopy and Multivariate Calibration”, *Anal. Lett.*, vol. 45, p. 774–781, 2012.

[56] L. Alessandrini, S. Romani, G. Pinnavaia, e M. D. Rosa, “Near infrared spectroscopy: An analytical tool to predict coffee roasting degree”, *Anal. Chim. Acta*, vol. 625, p. 95–102, 2008.

[57] E. De Hoffmann e V. Stroobant, *Mass Spectrometry - Priniples and Applications.*, vol. 29, John Wiley & Sons: Chichester, 2007.

[58] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, e C. M. Whitehouse, “Electrospray Ionization for Mass Spectrometry of Large Biomolecules”, *Science (80-.).*, vol. 246, p. 64–71, 1989.

[59] A. R. Venter, K. A. Douglass, J. T. Shelley, G. Hasman, e E. Honarvar, “Mechanisms of real-time, proximal sample processing during ambient ionization mass spectrometry”, *Anal. Chem.*, vol. 86, p. 233–249, 2014.

[60] H. Wang, J. Liu, R. G. Cooks, e Z. Ouyang, “Paper spray for direct analysis of complex mixtures using mass spectrometry”, *Angew. Chemie - Int. Ed.*, vol. 49, p. 877–880, 2010.

[61] J. Liu, H. Wang, N. E. Manicke, J. Lin, e R. G. Cooks, “Development , Characterization , and Application of Paper Spray Ionization”, *Anal. Chem.*, vol. 82, p. 2463–2471, 2010.

[62] R. D. Espy, A. R. Muliadi, Z. Ouyang, e R. G. Cooks, “Spray mechanism in paper spray ionization”, *Int. J. Mass Spectrom.*, vol. 325–327, p. 167–171, 2012.

[63] A. M. Hamid, P. Wei, A. K. Jarmusch, V. Pirro, e R. G. Cooks, “Discrimination of Candida species by paper spray mass spectrometry”, *Int. J. Mass Spectrom.*, vol. 86, p. 7500–7507, 2014.

[64] A. Naccarato, S. Moretti, G. Sindona, e A. Tagarelli, “Identification and assay of underivatized urinary acylcarnitines by paper spray tandem mass spectrometry”, *Anal. Bioanal. Chem.*, vol. 405, p. 8267–8276, 2013.

[65] Q. Yang, H. Wang, J. D. Maas, W. J. Chappell, N. E. Manicke, R. G. Cooks, e Z. Outang, “Paper spray ionization devices for direct, biomedical analysis using mass spectrometry”, *Int. J. Mass Spectrom.*, vol. 312, p. 201–207, 2012.

[66] C. S. Jhang, H. Lee, Y. S. He, J. T. Liu, e C. H. Lin, “Rapid screening and determination of 4-chloroamphetamine in saliva by paper spray-mass spectrometry and capillary electrophoresis-mass spectrometry”, *Electrophoresis*, vol. 33, p. 3073–3078, 2012.

[67] E. Domingos, T. C. Carvalho, I. Pereira, G. A. Vasconcelos, C. J. Thompson, R. Augusti, R. R. T. Rodrigues, L. V. Tose, H. Santos, J. R. Araujo, B. G. Vaz, e W. Romão, “Paper spray ionization mass spectrometry applied to forensic

chemistry – drugs of abuse, inks and questioned documents", *Anal. Methods*, vol. 9, p. 4400–4409, 2017.

[68] V. S. Amador, H. V. Pereira, M. M. Sena, R. Augusti, e E. Piccin, "Paper Spray Mass Spectrometry for the Forensic Analysis of Black Ballpoint Pen Inks", *J. Am. Soc. Mass Spectrom.*, vol. 28, p. 1965–1976, 2017.

[69] P. S. Ferreira, D. F. A. Silva, R. Augusti, e E. Piccin, "Forensic analysis of ballpoint pen inks using paper spray mass spectrometry", *Analyst*, vol. 140, p. 811–819, 2015.

[70] J. A. R. Teodoro, H. V. Pereira, D. N. Correia, M. M. Sena, E. Piccin, e R. Augusti, "Forensic discrimination between authentic and counterfeit perfumes using paper spray mass spectrometry and multivariate supervised classification", *Anal. Methods*, vol. 9, p. 4979–4987, 2017.

[71] A. K. Cheburkin e W. Shotyk, "Double-Plate Sample Carrier for a Simple Total Reflection X-Ray Fluorescence Analyser", *X-Ray Spectrom.*, vol. 25, p. 175–178, 1996.

[72] I. de la Calle, N. Cabaleiro, V. Romero, I. Lavilla, e C. Bendicho, "Sample pretreatment strategies for total reflection X-ray fluorescence analysis: A tutorial review", *Spectrochim. Acta - Part B*, vol. 90, p. 23–54, 2013.

[73] A. Cinosi, N. Andriollo, G. Pepponi, e D. Monticelli, "A novel total reflection X-ray fluorescence procedure for the direct determination of trace elements in petrochemical products", *Anal. Bioanal. Chem.*, vol. 399, p. 927–933, 2011.

[74] A. von Bohlen, "Total reflection X-ray fluorescence and grazing incidence X-ray spectrometry - Tools for micro- and surface analysis. A review", *Spectrochim. Acta - Part B*, vol. 64, p. 821–832, 2009.

[75] K. Reinhold e A. von Bohlen, "Total-reflection X-ray fluorescence moving

towards nanoanalysis: a survey", *Spectrochim. Acta - Part B*, vol. 56, p. 2005–2018, 2001.

[76] Chistiane Romanelli Rocha, "Desenvolvimento de método para análise de leite por espectrometria de fluorescência de raios X por reflexão total", UFMG, Belo Horizonte, 2015. (Dissertação)

[77] L. V. Resende e C. C. Nascentes, "A simple method for the multi-elemental analysis of organic fertilizer by slurry sampling and total reflection X-ray fluorescence", *Talanta*, vol. 147, p. 485–492, 2016.

[78] R. Souza, M. Grac, e S. M. Simabuco, "Trace Elements Content of Colostrum Milk in Brazil", *J. Food Compos. Anal.*, vol. 15, p. 27–33, 2002.

[79] R. O. R. Ribeiro, E. T. Máralém, E. F. O. Jesus, C. S. Carneiro, C. A. C. Júnior, E. Almeida, e V. F. Nascimento Filho, "Determination of Trace Elements in Honey from Different Regions in Rio de Janeiro State (Brazil) by Total Reflection X-Ray Fluorescence", *J. Food Sci.*, vol. 79, p. 738–742, 2014.

[80] H. Stosnach, "Analytical determination of selenium in medical samples, staple food and dietary supplements by means of total re fl ection X-ray fl uorescence spectroscopy", *Spectrochim. Acta Part B*, vol. 65, p. 859–863, 2010.

[81] D. Guimarães, M. L. Carvalho, M. Becker, A. Von Bohlen, e V. Geraldes, "Lead concentration in feces and urine of exposed rats by x-ray fl uorescence and electrothermal atomic absorption spectrometry", *X-Ray Spectrom.*, vol. 41, p. 80–86, 2012.

[82] M. C. Valentinuzzi e M. S. Gren, "Influence of smoking on the elemental composition of oral fluids : a TXRF approach", *X-Ray Spectrom.*, vol. 39, p. 372–375, 2010.

[83] M. J. Martín, F. Pablos, e A. G. González, "Characterization of green coffee

varieties according to their metal content", *Anal. Chim. Acta*, vol. 358, p. 177–183, 1998.

[84] H. Santos e E. Oliveira, "Determination of Mineral Nutrients and Toxic Elements in Brazilian Soluble Coffee by ICP-AES", *J. Food Compos. Anal.*, vol. 14, p. 523–531, 2001.

[85] A. Santato, D. Bertoldi, M. Perini, F. Camin, e R. Larcher, "Using elemental profiles and stable isotopes to trace the origin of green coffee beans on the global market", *J. Mass Spectrom.*, vol. 47, p. 1132–1140, 2012.

[86] S. J. Haswell e A. D. Walmsley, "Multivariate data visualisation methods based on multi-elemental analysis of wines and coffees using total reflection X-ray fluorescence analysis", *J. Anal. At. Spectrom.*, vol. 13, p. 131–134, 1998.

[87] M. M. C. Ferreira, A. M. Antunes, M. S. Melgo, e P. L. O. Volpe, "Chemometrics I: multivariate calibration, a tutorial", *Química Nova*, vol. 22. p. 724–731, 1999.

[88] S. Wold, "Chemometrics; what do we mean with it, and what do we want from it?", *Chemom. Intell. Lab. Syst.*, vol. 30, p. 109–115, 1995.

[89] D. B. Hibbert, "Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016)", *Pure Appl. Chem.*, vol. 88, p. 407–443, 2016.

[90] R. G. Brereton, "A short history of chemometrics: A personal view", *J. Chemom.*, vol. 28, p. 749–760, 2014.

[91] B. R. Kowalski, "Chemometrics: Views and Propositions", *J. Chem. Inf. Comput. Sci.*, vol. 15, p. 201–203, 1975.

[92] P. K. Hopke, "The evolution of chemometrics", *Anal. Chim. Acta*, vol. 500, p. 365–377, 2003.

[93] B. B. Neto, I. S. Scarmínio, e R. E. Bruns, "25 anos de quimiometria no Brasil",

Quim. Nova, vol. 29, p. 1401–1406, 2006.

[94] S. Wold e M. Sjöström, “Chemometrics, present and future success”, *Chemom. Intell. Lab. Syst.*, vol. 44, p. 3–14, 1998.

[95] R. G. Brereton, *Applied Chemometrics for Scientists*. Chichester, Inglaterra, 2007.

[96] P. Geladi e B. R. Kowalski, “Partial least-squares regression: a tutorial”, *Anal. Chim. Acta*, vol. 185, p. 1–17, 1986.

[97] D. A. Cirovic, R. G. Brereton, P. T. Walsh, J. A. Ellwood, e E. Scobbie, “Application of Partial Least Squares Calibration to Measurements of Polycyclic Aromatic Hydrocarbons in Coal Tar Pitch Volatiles”, *Data Process.*, vol. 121, p. 575–580, 1996.

[98] S. Wold, J. Trygg, A. Berglund, e H. Antti, “Some recent developments in PLS modeling”, *Chemom. Intell. Lab. Syst.*, vol. 58, p. 131–150, 2001.

[99] J. P. A. Martins, R. F. Teófilo, e M. M. C. Ferreira, “Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets”, *J. Chemom.*, vol. 24, p. 320–332, 2010.

[100] N. M. Faber e R. Rajkó, “How to avoid over-fitting in multivariate calibration-The conventional validation approach and an alternative”, *Anal. Chim. Acta*, vol. 595, p. 98–106, 2007.

[101] A. Höskuldsson, “Variable and subset selection in PLS regression”, *Chemom. Intell. Lab. Syst.*, vol. 55, p. 23–38, 2001.

[102] T. Mehmood, K. H. Liland, L. Snipen, e S. Sæbø, “A review of variable selection methods in Partial Least Squares Regression”, *Chemom. Intell. Lab. Syst.*, vol. 118, p. 62–69, 2012.

[103] Y. H. Yun, W. T. Wang, B. C. Deng, G. B. Lai, X. Liu, D. B. Ren, Y. Z. Liang,

W. Fan, e Q. S. Xu, “Using variable combination population analysis for variable selection in multivariate calibration”, *Anal. Chim. Acta*, vol. 862, p. 14–23, 2015.

[104] H.-D. Li, Y.-Z. Liang, Q.-S. Xu, e D.-S. Cao, “Model population analysis for variable selection”, *J. Chemom.*, vol. 24, p. 418–423, 2010.

[105] R. M. Balabin e S. V. Smirnov, “Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data”, *Anal. Chim. Acta*, vol. 692, p. 63–72, 2011.

[106] C. M. Andersen e R. Bro, “Variable selection in regression-a tutorial”, *J. Chemom.*, vol. 24, p. 728–737, 2010.

[107] D. Wu, P. Nie, Y. He, e Y. Bao, “Determination of Calcium Content in Powdered Milk Using Near and Mid-Infrared Spectroscopy with Variable Selection and Chemometrics”, *Food Bioprocess Technol.*, vol. 5, p. 1402–1410, 2012.

[108] N. Sorol, E. Arancibia, S. A. Bortolato, e A. C. Olivieri, “Visible/near infrared-partial least-squares analysis of Brix in sugar cane juice. A test field for variable selection methods”, *Chemom. Intell. Lab. Syst.*, vol. 102, p. 100–109, 2010.

[109] D. Broadhurst, R. Goodacre, A. Jones, J. J. Rowland, e D. B. Kell, “Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry”, *Anal. Chim. Acta*, vol. 348, p. 71–86, 1997.

[110] R. F. Teófilo, J. P. A. Martins, e M. M. C. Ferreira, “Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression”, *J. Chemom.*, vol. 23, p. 32–48, 2009.

[111] M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C.

Chame, e V. Visani, “The successive projections algorithm for variable selection in spectroscopic multicomponent analysis”, *Chemom. Intell. Lab. Syst.*, vol. 57, p. 65–73, 2001.

[112] L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, e S. B. Engelsen, “Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy”, *Appl. Spectrosc.*, vol. 54, p. 413–419, 2000.

[113] J. Zimmer e M. J. Anzanello, “Um novo método para seleção de variáveis preditivas com base em índices de importância”, *Production*, vol. 24, p. 84–93, 2014.

[114] J. M. Silla, C. A. Nunes, R. A. Cormanich, M. C. Guerreiro, T. C. Ramalho, e M. P. Freitas, “MIA-QSPR and effect of variable selection on the modeling of kinetic parameters related to activities of modified peptides against dengue type 2”, *Chemom. Intell. Lab. Syst.*, vol. 108, p. 146–149, 2011.

[115] J. H. HOLLAND, *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[116] P. A. Costa Filho e R. J. Poppi, “Algoritmo genético em química”, *Quim. Nova*, vol. 22, p. 405–411, 1999.

[117] C. B. Lucasius e G. Kateman, “Understanding and using genetic algorithms Part 1. Concepts, properties and context”, *Chemom. Intell. Lab. Syst.*, vol. 19, p. 1–33, 1993.

[118] C. B. Lucasius e G. Kateman, “Genetic algorithms for large-scale optimization in chemometrics: An application”, *Trends Anal. Chem.*, vol. 10, p. 254–261, 1991.

[119] H. C. Goicoechea e A. C. Olivieri, “A new family of genetic algorithms for

wavelength interval selection in multivariate analytical spectroscopy", *J. Chemom.*, vol. 17, p. 338–345, 2003.

[120] Z. Xiaobo, Z. Jiewen, M. J. W. Povey, M. Holmes, e M. Hanpin, "Variables selection methods in near-infrared spectroscopy", *Anal. Chim. Acta*, vol. 667, p. 14–32, 2010.

[121] M. M. C. Ferreira, C. A. Montanari, e A. C. Gaudio, "Seleção de variáveis em QSAR", *Quim. Nova*, vol. 25, p. 439–448, 2002.

[122] C. Soares, A. A. Gomes, M. Cesar, e U. Araujo, "The successive projections algorithm", vol. 42, 2013.

[123] S. F. C. Soares, R. K. H. Galvão, M. C. U Araújo, E. C. Silva, C. F. Pereira, S. I. E. Andrade, e F. C. Leite, "A modification of the successive projections algorithm for spectral variable selection in the presence of unknown interferents", *Anal. Chim. Acta*, vol. 689, p. 22–28, 2011.

[124] H. M. Paiva, S. F. C. Soares, R. K. H. Galvão, e M. C. U. Araujo, "A graphical user interface for variable selection employing the Successive Projections Algorithm", *Chemom. Intell. Lab. Syst.*, vol. 118, p. 260–266, 2012.

[125] M. J. Anzanello, K. Fu, F. F. Fogliatto, e M. F. Ferrão, "HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point", *Chemom. Intell. Lab. Syst.*, vol. 145, p. 1–6, 2015.

[126] C. D. Bernardes e P. J. S. Barbeira, "Different Chemometric Methods for the Discrimination of Commercial Aged Cachaças", *Food Anal. Methods*, vol. 9, p. 1053–1059, 2016.

[127] M. Spiteri, E. Dubin, J. Cotton, M. Poirel, B. Corman, E. Jamin, M. Lees, e D. Rutledge, "Data fusion between high resolution 1H-NMR and mass spectrometry: a synergetic approach to honey botanical origin

characterization”, *Anal. Bioanal. Chem.*, vol. 408, p. 4389–4401, 2016.

[128] A. Biancolillo, R. Bucci, A. L. Magrì, A. D. Magrì, e F. Marini, “Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication”, *Anal. Chim. Acta*, vol. 820, p. 23–31, 2014.

[129] K. M. Nunes, M. V. O. Andrade, A. M. P. Santos Filho, M. C. Lasmar, e M. M. Sena, “Detection and characterisation of frauds in bovine meat in natura by non-meat ingredient additions using data fusion of chemical parameters and ATR-FTIR spectroscopy”, *Food Chem.*, vol. 205, p. 14–22, 2016.

[130] K. M. Nunes, “Utilização de espectroscopia no infravermelho médio, fusão de dados e métodos quimiométricos de classificação na análise de fraudes em carnes bovinas in natura”, UFMG, Belo Horizonte, 2015. (Dissertação)

[131] W. Dong, J. Zhao, R. Hu, Y. Dong, e L. Tan, “Differentiation of Chinese robusta coffees according to species, using a combined electronic nose and tongue, with the aid of chemometrics”, *Food Chem.*, vol. 229, p. 743–751, 2017.

[132] A. Dankowska, A. Domagała, e W. Kowalewski, “Quantification of Coffea arabica and Coffea canephora var. robusta concentration in blends by means of synchronous fluorescence and UV-Vis spectroscopies”, *Talanta*, vol. 172, p. 215–220, 2017.

[133] Å. Rinnan, F. van den Berg, e S. B. Engelsen, “Review of the most common pre-processing techniques for near-infrared spectra”, *TrAC - Trends Anal. Chem.*, vol. 28, p. 1201–1222, 2009.

[134] R. W. Kennard e L. A. Stone, “Computer Aided Design of Experiments”, *Technometrics*, vol. 11, p. 137–148, 1969.

[135] B. G. Botelho, B. A. P. Mendes, e M. M. Sena, “Development and Analytical Validation of Robust Near-Infrared Multivariate Calibration Models for the

Quality Inspection Control of Mozzarella Cheese", *Food Anal. Methods*, vol. 6, p. 881–891, 2013.

[136] S. V. C. Souza e R. G. Junqueira, "A procedure to assess linearity by ordinary least squares method", *Anal. Chim. Acta*, vol. 552, p. 23–35, 2005.

[137] P. Valderrama, J. W. B. Braga, e R. J. Poppi, "Estado da arte de figuras de mérito em calibração multivariada", *Quim. Nova*, vol. 32, p. 1278–1287, 2009.

[138] P. Valderrama, J. W. B. Braga, e R. J. Poppi, "Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy", *J. Agric. Food Chem.*, vol. 55, p. 8331–8338, 2007.

[139] M. H. Ferreira, J. W. B. Braga, e M. M. Sena, "Development and validation of a chemometric method for direct determination of hydrochlorothiazide in pharmaceutical samples by diffuse reflectance near infrared spectroscopy", *Microchem. J.*, vol. 109, p. 158–164, 2013.

[140] E. K. Kemsley, S. Ruault, e R. H. Wilson, "Discrimination Between Coffea-Arabica and Coffea-Canephora Variant Robusta Beans Using Infrared-Spectroscopy", *Food Chem.*, vol. 54, p. 321–326, 1995.

[141] M. M. Paradkar e J. Irudayaraj, "Rapid determination of caffeine content in soft drinks using FTIR-ATR spectroscopy", *Food Chem.*, vol. 78, p. 261–266, 2002.

[142] D. J. Lyman, R. Benck, S. Dell, S. Merle, e J. Murray-Wijelath, "FTIR-ATR Analysis of Brewed Coffee : Effect of Roasting Conditions", *Jounal Agric. Food Chem.*, vol. 51, p. 3268–3272, 2003.

[143] J. R. Santos, O. Viegas, R. N. M. J. Páscoa, I. M. P. L. V. O. Ferreira, A. O. S. S. Rangel, e J. A. Lopes, "In-line monitoring of the coffee roasting process with

near infrared spectroscopy: Measurement of sucrose and colour", *Food Chem.*, vol. 208, p. 103–110, 2016.

[144] C. W. Huck, W. Guggenbichler, e G. K. Bonn, "Analysis of caffeine, theobromine and theophylline in coffee by near infrared spectroscopy (NIRS) compared to high-perfomance liquid chromatography (HPLC) coupled to mass spectrometry", *Anal. Chim. Acta*, vol. 538, p. 195–203, 2005.

[145] A. C. L. Amorim, A. M. C. Hovelli, A. C. Pinto, M. N. Eberlin, N. P. Arruda, E. J. Pereira, H. R. Bizzo, R. R. Catharino, Z. B. M. Filho, e C. M. Rezende, "Green and roasted Arabica coffees differentiated by ripeness, process and cup quality via electrospray ionization mass spectrometry fingerprinting", *J. Braz. Chem. Soc.*, vol. 20, p. 313–321, 2009.

[146] A. Panusa, R. Petrucci, R. Lavecchia, e A. Zuorro, "UHPLC-PDA-ESI-TOF/MS metabolic profiling and antioxidant capacity of arabica and robusta coffee silverskin: Antioxidants vs phytotoxins", *Food Res. Int.*, vol. 99, p. 155–165, 2017.

[147] I. Esteban-Díez, J. M. González-Sáiz, e C. Pizarro, "Prediction of sensory properties of espresso from roasted coffee samples by near-infrared spectroscopy", *Anal. Chim. Acta*, vol. 525, p. 171–182, 2004.

[148] J. S. Ribeiro, F. Augusto, T. J. G. Salva, e M. M. C. Ferreira, "Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics", *Talanta*, vol. 101, p. 253–260, 2012.

[149] V. Krivan, P. Barth, e A. F. Morales, "Multielement analysis of green coffee and its possible use for the determination of origin", *Mikrochim. Acta*, vol. 110, p. 217–236, 1993.

[150] B. M. Nicolai, K. Beullens, E. Bobelyn, A. Peirs, W. Saeys, K. I. Theron, e J.

Lammertyn., “Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy: A review”, *Postharvest Biol. Technol.*, vol. 46, p. 99–118, 2007.

ANEXO - Dados TXRF Torra Leve (mg/L)

%Robusta	P	S	Cl	K	Ca	Ti	Mn	Fe	Ni	Cu	Zn	Br	Rb	Sr
0	32,78	12,40	15,18	637,09	32,23	0,16	0,86	0,29	0,01	0,02	0,23	0,48	0,53	0,06
1	32,27	11,43	12,32	593,61	26,71	0,10	0,77	0,24	0,00	0,01	0,18	0,23	0,51	0,04
2	33,28	11,92	11,97	616,89	27,23	0,10	0,80	0,27	0,01	0,01	0,16	0,17	0,56	0,05
3	33,51	12,01	11,97	613,37	27,14	0,11	0,78	0,26	0,01	0,01	0,20	0,17	0,57	0,05
3	33,46	11,75	12,60	620,92	26,77	0,11	0,79	0,43	0,01	0,03	0,19	0,18	0,58	0,04
3	33,09	11,64	12,30	615,35	26,91	0,14	0,78	0,27	0,01	0,02	0,20	0,17	0,57	0,04
4	29,61	11,46	13,12	571,59	28,52	0,09	0,77	0,51	0,01	0,02	0,19	0,27	0,56	0,06
5	28,37	10,50	13,32	547,02	26,07	0,12	0,72	0,23	0,00	0,01	0,18	0,27	0,54	0,05
6	30,54	12,15	13,60	586,76	30,16	0,11	0,78	0,47	0,01	0,01	0,21	0,28	0,61	0,06
7	29,86	11,58	13,43	564,83	28,88	0,24	0,76	0,30	0,01	0,02	0,20	0,28	0,60	0,06
8	29,20	11,03	12,71	542,85	27,99	0,11	0,72	0,21	0,01	0,01	0,18	0,28	0,58	0,06
9	28,66	10,39	12,99	542,73	25,81	0,14	0,70	0,88	0,01	0,01	0,21	0,27	0,61	0,06
10	29,90	11,36	13,84	566,83	27,49	0,12	0,73	0,23	0,01	0,02	0,22	0,29	0,63	0,07
11	30,92	11,08	13,26	570,75	25,61	0,11	0,70	0,41	0,01	0,01	0,18	0,24	0,66	0,05
12	30,00	11,10	13,77	566,41	27,54	0,15	0,73	0,56	0,01	0,01	0,18	0,29	0,67	0,07
13	30,12	11,13	13,99	574,26	28,03	0,43	0,76	0,46	0,01	0,02	0,20	0,28	0,70	0,06
14	30,63	11,77	12,95	592,42	28,44	0,12	0,70	0,39	0,01	0,02	0,18	0,21	0,76	0,06
15	28,61	10,58	14,32	556,02	26,71	0,21	0,68	0,50	0,01	0,02	0,17	0,29	0,72	0,06
15	27,70	10,30	13,59	543,09	26,15	0,14	0,68	0,27	0,01	0,02	0,17	0,28	0,70	0,06
15	28,44	10,72	14,41	556,09	26,20	0,10	0,69	0,68	0,01	0,02	0,17	0,29	0,71	0,06
16	23,62	8,62	10,06	438,71	20,39	0,09	0,53	0,18	0,01	0,01	0,15	0,19	0,57	0,04
17	32,35	12,14	13,22	613,57	30,03	0,10	0,73	0,32	0,01	0,02	0,16	0,23	0,82	0,08
18	25,66	9,54	10,80	484,76	22,14	0,08	0,57	0,24	0,01	0,01	0,15	0,24	0,67	0,05
19	29,17	10,98	11,68	559,49	24,95	0,12	0,65	0,30	0,00	0,01	0,16	0,21	0,83	0,06

20	30,81	11,83	14,36	584,55	28,48	0,14	0,70	0,34	0,01	0,02	0,18	0,29	0,84	0,07
21	28,09	10,62	14,86	539,50	26,36	0,08	0,67	0,22	0,01	0,01	0,18	0,39	0,83	0,07
22	30,10	11,20	13,39	571,53	26,10	0,10	0,66	0,24	0,01	0,02	0,16	0,25	0,88	0,06
23	29,80	10,90	13,73	559,11	26,20	0,18	0,65	0,28	0,01	0,02	0,17	0,28	0,86	0,07
24	30,37	11,09	11,79	570,46	24,51	0,14	0,67	0,32	0,01	0,03	0,18	0,18	0,92	0,06
25	29,61	11,01	12,98	561,33	25,51	0,12	0,65	0,29	0,01	0,03	0,18	0,25	0,92	0,06
26	29,30	10,94	12,64	555,73	25,30	0,13	0,66	0,41	0,01	0,05	0,19	0,23	0,90	0,06
27	31,60	11,98	15,35	615,72	27,50	0,16	0,68	0,38	0,01	0,04	0,18	0,25	1,06	0,06
27	31,50	12,09	13,70	603,40	27,48	0,36	0,69	0,43	0,01	0,03	0,17	0,23	1,03	0,06
27	31,08	11,79	13,71	599,68	26,92	0,22	0,68	0,36	0,01	0,03	0,17	0,23	1,02	0,06
28	25,13	9,15	11,27	448,79	21,09	0,17	0,54	0,31	0,01	0,02	0,36	0,20	0,77	0,05
29	26,93	9,81	12,72	488,99	24,34	0,26	0,60	0,34	0,01	0,03	0,34	0,24	0,85	0,06
30	30,40	11,53	13,24	572,47	24,16	0,14	0,63	0,41	0,01	0,03	0,51	0,19	1,02	0,06
31	29,04	10,92	12,04	540,50	24,92	0,28	0,60	0,56	0,01	0,02	0,29	0,17	1,01	0,06
32	35,39	10,87	13,64	464,62	30,32	0,59	0,70	0,93	0,02	0,05	0,34	0,18	0,83	0,06
33	30,30	10,08	11,58	484,27	22,70	0,20	0,60	0,28	0,01	0,02	0,29	0,20	0,86	0,05