UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

DEPARTAMENTO ACADÊMICO DE ALIMENTOS

CURSO DE ENGENHARIA DE ALIMENTOS

PATRICIA CASARIN DE LIMA

DISCRIMINAÇÃO DE ERVA-MATE PARA CHIMARRÃO QUANTO À ORIGEM GEOGRÁFICA E PRESENÇA DE AÇÚCAR UTILIZANDO FTIR E QUIMIOMETRIA

CAMPO MOURÃO

2019

PATRICIA CASARIN DE LIMA

DISCRIMINAÇÃO DE ERVA-MATE PARA CHIMARRÃO QUANTO À ORIGEM GEOGRÁFICA E PRESENÇA DE AÇÚCAR UTILIZANDO FTIR E QUIMIOMETRIA

Trabalho de conclusão de curso de graduação, apresentado ao Curso Superior de Engenharia de Alimentos (Departamento Acadêmico de Alimentos) da Universidade Tecnológica Federal do Paraná – UTFPR, Campus Campo Mourão, como requisito parcial para a obtenção do título de Bacharel em Engenharia de Alimentos.

Orientador: Prof. Dr. Evandro Bona

CAMPO MOURÃO

2019

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Campo Mourão
Departamento Acadêmico de Alimentos

TERMO DE APROVAÇÃO

DISCRIMINAÇÃO DE ERVA-MATE PARA CHIMARRÃO QUANTO À ORIGEM GEOGRÁFICA E PRESENÇA DE AÇÚCAR UTILIZANDO FTIR E QUIMIOMETRIA

por

PATRICIA CASARIN DE LIMA

Este Trabalho de Conclusão de Curso (TCC) foi apresentado no dia 26 de novembro de 2019 como requisito parcial para obtenção do título de Bacharel em Engenharia de Alimentos. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Evandro Bona

Prof. Dr. Augusto Tanamati

Prof. Dr. Odinei Hess Gonçalves

Nota: O documento original e assinado pela banca examinadora encontra-se na Coordenação do Curso de Engenharia de Alimentos da UTFPR campus Campo Mourão

AGRADECIMENTOS

Agradeço a Deus por me socorrer e fortalecer em sua infinita graça em cada momento em que o desespero e a sensação de fracasso tomavam conta da minha mente, por me dar perseverança e me proporcionar cada vitória nesses cinco anos. Com Ele nada é impossível.

Aos meus pais, Vilma e Moacir, e ao meu irmão, Alexandre, por todo amor, carinho, paciência e dedicação para comigo. Obrigada por cada tentativa de me fazer sorrir nas minhas piores horas, por cada palavra de incentivo. Ter vocês como família é um presente imensurável.

Ao meu orientador, Prof. Dr. Evandro Bona, por ser essa pessoa e profissional incríveis que eu tanto admiro. Trabalhar com você foi sempre um grande anseio, por isso sou extremamente grata por ter acreditado em mim e me oportunizar isso durante esses anos. Agradeço imensamente cada ensinamento e por transmiti-los de forma tão leve e paciente.

Agradeço ainda aos meus amigos de graduação, em especial Flávia Trivelato da Costa, Sidnei Macedo e Giovana Caroline Tonon. Muito obrigada por ouvirem meus desabafos e estarem sempre prontos para me consolarem. Obrigada por acreditarem tanto em mim quando eu mesma não o fazia, por me não me negarem ajuda e, principalmente, por tornar os momentos estressantes mais leves.

Agradeço também as minhas amigas do PPGTA, Franciele Leila Viell, Talita Butzske, Luana Dalagrana e Mariana Braga. Obrigada por todo o carinho de vocês e ajuda quando precisei. Vocês são pessoas maravilhosas.

RESUMO

LIMA, Patricia Casarin. Discriminação de erva-mate para chimarrão quanto à origem geográfica e presença de açúcar utilizando FTIR e quimiometria. 2019. 39 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2019.

Ilex paraguariensis ou erva-mate é uma planta comercialmente importante no setor alimentício, destacando-se o chimarrão, bebida típica dos estados do Sul brasileiro – Paraná (PR), Santa Catarina (SC) e Rio Grande do Sul (RS) – os maiores produtores nacionais. Assim como vinhos e outros produtos, a erva-mate muda química e sensorialmente dada a região de cultivo, o que estimula a prática de adulterações, como adição de açúcar para mascarar aspectos sensoriais indesejados ou atribuição errônea de procedência para agregar falso valor. Esse trabalho teve por objetivo utilizar a espectroscopia FTIR, alternativamente às metodologias padrão, aliada à quimiometria na classificação de erva-mate quanto origem e presença de açúcar, para possível aplicação no controle de adulteração. Adotou-se a técnica de pastilhamento com KBr como suporte das amostras, realizando-se leituras em duplicata no espectrômetro Shimadzu IRAffinity-1 para as 69 amostras provenientes dos três estados da região Sul. No software Matlab R2008b, os espectros passaram por pré-tratamentos, incluindo seleção de variáveis, para se trabalhar com as informações de real importância. Das repetições, extraíram-se os espectros médios, para os quais se empregou o método de classificação PLS-DA. Conseguiu-se um bom modelo discriminante quanto à origem da erva-mate empregando-se 15 variáveis latentes, o qual apresentou sensibilidade superior a 62% e seletividade superior a 89% para as três classes. Após validação cruzada, observou-se na previsão AUC igual a 0,9133. Já para a presença de açúcar, elegeu-se o modelo com 10 variáveis latentes, cuja exatidão foi 65%, sensibilidade 62% e 75% para as classes sem e com açúcar, respectivamente, e AUC 0,6923. Portanto, mesmo diante da quantidade elevada de variáveis latentes necessárias, o uso do PLS-DA em espectros FTIR foi satisfatório, visto a complexidade da matriz estudada e as próprias limitações de um método linear.

Palavras-chave: PLS-DA; FTIR; Controle de adulteração.

ABSTRACT

LIMA, Patricia Casarin. Discrimination of yerba mate for *chimarrão* about geographical origin and presence of sugar applying FTIR and chemometrics. 2019. 39 f. Trabalho de Conclusão de Curso – Departamento Acadêmico de Alimentos, Universidade Tecnológica Federal do Paraná. Campo Mourão, 2019.

Ilex paraguariensis or yerba mate is a commercially important plant in the food sector, especially *chimarrão*, a typical drink from South Brazilian states – Paraná (PR), Santa Catarina (SC) and Rio Grande do Sul (RS) – the largest national producers. Like wines and other products, yerba mate changes chemically and sensorially according to the region of cultivation, which encourages the practice of adulteration, such as adding sugar to mask unwanted sensory aspects or misleading provenance to add false value. The objective of this work was to use FTIR spectroscopy, alternatively to standard methods, combined with chemometrics in the classification of yerba mate as origin and presence of sugar, for possible application in the control of adulteration. The KBr pellet technique was adopted to support the samples. Duplication assays were performed on the Shimadzu IRAffinity-1 spectrometer for the 69 samples from the three southern states. In the Matlab R2008b, the spectra were pretreated, including variable selection, working with the information of real importance. From the duplication, the average spectra were extracted, for which the PLS-DA classification method was used. A good discriminating model was obtained about the origin of the yerba mate using 15 latent variables. The sensitivity was higher than 62% and selectivity greater than 89% for the three classes. After cross-validation, the AUC was 0.9133. For the presence of sugar, the model with 10 latent variables was chosen, whose accuracy was 65%, sensitivity 62% and 75% for the classes without sugar and with sugar, respectively, and AUC 0.6923. Therefore, even considering the high amount of latent variables required, the use of PLS-DA in FTIR spectra was satisfactory, considering the complexity of the studied matrix and the limitations of a linear method.

Keywords: PLS-DA; FTIR; Adulteration control.

SUMÁRIO

1. INTRODUÇÃO

É da espécie *Ilex paraguariesis* de origem latino-americana que se tem o popular produto conhecido como erva-mate, cuja versatilidade vem se mostrando ao se aprofundar o conhecimento de seus compostos bioativos, tendo-se não apenas uma matéria-prima de bebidas, mas também uma fonte de antioxidantes que podem ser empregados desde no desenvolvimento de aditivos alimentares até em produtos fármacos e cosméticos (KAHMANN et al., 2017; SCHNEIDER et al., 2018a; VIEIRA et al., 2019). O Brasil está entre os países destaques na produção e consumo da planta (SCHNEIDER et al., 2018a).

Semelhante a outros produtos, como alguns queijos e vinhos, percebeu-se que a erva-mate também adquire características sensoriais e nutricionais específicas vinculadas às regiões onde foram cultivadas, sejam por aspectos climáticos e de composição de solo da região, seja pela natureza do cultivar (nativo ou plantado), pelas técnicas de plantio ou mesmo de pós-colheita (MARCELO; POZEBON; FERRÃO, 2015; VIEIRA et al., 2019). Por isso, há uma crescente defesa da conferência de indicação geográfica desse produto (CHECHI, 2019). De acordo com o INPI, a indicação geográfica é um registro que atribui uma valorização ao produto inerente à sua origem, por possuir características únicas que o diferencia de outros do mesmo seguimento de mercado, ou seja, aquele produto tal qual se conhece só é possível obter se produzido em sua região de origem (MAPA, 2017).

Junto com esse apelo de mercado acerca da indicação geográfica vem uma prática ilegal por parte do comercializador, que objetiva aumentar os seus lucros em detrimento do real produto ofertado ao consumidor: a adulteração da origem da erva-mate. Os fraudadores processam a erva de outras regiões e atribuem, em suas embalagens, a informação de que são de uma localidade de maior valor agregado (ESTEKI; SHAHSAVARI; SIMAL-GANDARA, 2018; MARCELO; POZEBON; FERRÃO, 2015; VIEIRA et al., 2019).

Ainda a respeito das fraudes, é observada também a prática de adição de sacarose à erva com o intuito de mascarar aspectos sensoriais desagradáveis, os quais podem ser provenientes desde matérias-primas de qualidade inferior até aquelas que já estão sofrendo degradação microbiológica, neste último ocorrendo o agravante de prejuízo à segurança alimentar. Outra problemática em torno da adulteração por açúcar é que seu conteúdo não estará quantificado nas informações nutricionais, sendo um risco à saúde dos consumidores com restrições alimentares (SCHNEIDER et al., 2018a; VIEIRA et al., 2019).

Diante do exposto, fica clara a necessidade de análises laboratoriais que caracterizem o produto de modo a apontar se ele sofreu adição de açúcar e se sua procedência está de acordo com o informado.

Para tais caracterizações costuma-se aplicar métodos cromatográficos, principalmente HPLC e GC, os quais são relativamente demorados, necessitam de preparo prévio da amostra, são destrutivos e empregam, em geral, reagentes de alta toxicidade, mostrando-se um grande problema ao meio ambiente (FRIZON et al., 2015). Assim, técnicas alternativas que demandem menos tempo de análise, redução ou eliminação do preparo de amostra e, o principal, não utilizem reagentes poluentes vêm sendo estudadas, constituindo a base da Química Verde (GAŁUSZKA; MIGASZEWSKI; NAMIEŚNIK, 2013). Para isso, a espectroscopia em conjunto com a quimiometria (aplicação de ferramentas estatísticas no entendimento de dados químicos) se mostra como opção promissora na análise de alimentos (ROHMAN; MAN, 2010; STUART, 2004; ZHANG et al., 2012).

ALEXANDRE MARCELO e colaboradores (2014), FRIZON e colaboradores (2015) e VIEIRA e colaboradores (2019) aplicaram espectroscopia de infravermelho próximo (NIR) na classificação quanto à região de origem das amostras de erva-mate. ALEXANDRE MARCELO e colaboradores (2014) trabalharam com ervas-mate de quatro países sul-americanos diferentes, enquanto FRIZON e colaboradores (2015) com amostras de três regiões distintas do Paraná – estado brasileiro, e VIEIRA e colaboradores (2019) com ervas-mate dos três estados sul-brasileiros. SCHNEIDER e colaboradores (2018a), diferente dos demais autores, aplicaram a espectroscopia FTIR com a técnica ATR de leitura, para amostras de diferentes localidades do estado do Rio Grande do Sul.

Os espectros por si só não fornecem informações diretas, além da presença de determinados grupos funcionais. Por isso, ferramentas estatísticas são imprescindíveis para a extração de outras informações, por meio da comparação com dados de referência vindos de metodologias padrão (ROHMAN; MAN, 2010). A extração de informação física desses dados eletrônicos se faz pelo uso de métodos quimiométricos. De acordo com o que se objetiva, haverá ferramentas cujas aplicações são mais adequadas. Para os problemas de classificação nos quais se deseja verificar se um produto atende ou não às características de determinado grupo, como é o caso da verificação da autenticidade e adição de sacarose na erva-mate, recomendam-se os métodos supervisionados (CALLAO; RUISÁNCHEZ, 2018; FERREIRA, 2015).

Assim, o presente trabalho teve por objetivo realizar análise por espectroscopia no infravermelho médio de amostras de erva-mate, e classifica-las de acordo com seu local de origem e presença de sacarose através de técnica quimiométrica supervisionada.

1.1. OBJETIVOS

1.1.1. OBJETIVO GERAL

Classificar amostras de erva-mate de acordo com a região na qual foi cultivada (Paraná, Santa Catarina, Rio Grande do Sul) e quanto à presença de açúcar a partir da técnica instrumental de espectroscopia FTIR em conjunto com o método quimiométrico PLS-DA.

1.1.2. OBJETIVOS ESPECÍFICOS

- Verificar a capacidade do método supervisionado baseado na regressão linear e, portanto, com funcionamento simples, de reconhecer padrões complexos existentes em dados multivariados.
- Avaliar se os modelos preditivos são eficientes na discriminação para a qual foram propostos a partir de suas figuras de mérito.
- Avaliar, a partir de comparação com trabalho da literatura, se a natureza das informações obtidas por infravermelho médio possui maior conteúdo discriminante útil na construção dos modelos classificatórios do que àquelas obtidas por infravermelho próximo.

2. REVISÃO BIBLIOGRÁFICA

2.1. ERVA-MATE

Cientificamente conhecida por *Ilex paraguariensis*, a erva-mate (Figura 1) é uma planta nativa da América do Sul, região na qual se encontram seus principais produtores e consumidores mundiais, fazendo o Brasil parte de ambos devido aos seguintes estados expressivos no setor: Paraná, Santa Catarina e Rio Grande do Sul (CALLAO; RUISÁNCHEZ, 2018; KAHMANN et al., 2017; SCHNEIDER et al., 2018a; VIEIRA et al., 2019).

Figura 1: Imagens ilustrativas (a) plantio de erva-mate (b) erva-mate processada.

(a) (b)

Fonte: CAMPANÁRIO (2019); EPAGRI (2019).

Nas regiões das quais a erva-mate é originária, o elevado consumo se deve principalmente por razões culturais. Todavia, seu mercado consumidor vem expandindo em decorrência, sobretudo, do volume de estudos científicos que atribuem benefícios à saúde vindos da ingestão da planta, devido à presença de variados compostos bioativos, em especial os antioxidantes (RIACHI et al., 2018; SCHNEIDER et al., 2018a; VIEIRA et al., 2019).

Como alimento, a erva-mate costuma ser consumida na forma de bebidas através da sua infusão em água. Quando o beneficiamento da erva é finalizado na moagem, tendo-se a erva verde, é possível obter dois produtos distintos, cuja diferença está na granulometria e bebida destino: o primeiro é a erva-mate para chimarrão e o segundo a erva-mate para tererê. O chimarrão é a bebida originária da infusão a quente da erva de granulometria mais fina, enquanto o tererê vem da infusão a frio de granulometrias grossas. Há ainda uma terceira bebida obtida por infusão a quente, como o chimarrão, porém a partir da erva que passa pela etapa de torrefação no beneficiamento, tornando-se escura e com características organolépticas próprias: o chá mate (ANVISA, 2002; MARCELO; POZEBON; FERRÃO, 2015; VIEIRA et al., 2019).

Entretanto, a produção de mate não tem destino restrito ao setor de bebidas, abrangendo, por exemplo, a fabricação de fármacos, cosméticos e aditivos alimentares como corantes e conservantes (CAMOTTI BASTOS et al., 2018; SCHNEIDER et al., 2018b).

Sabe-se que semelhante a outros produtos de origem vegetal, a erva-mate sofre variações sensoriais e nutricionais de acordo com as condições a que foi exposta durante o plantio e pós-colheita. Temperatura, composição do solo e pesticidas empregados são alguns fatores que influenciam o desenvolvimento da planta e suas características quando adultas relativas à concentração de compostos químicos, por exemplo, os antioxidantes. A técnica empregada durante o sapeco e as condições de armazenagem são fatores pós-colheita também determinantes, os quais quando não adequados podem conferir aspecto sensorial desagradável e favorecer a deterioração microbiológica do produto (RIACHI et al., 2018; VIEIRA et al., 2019).

Assim, a erva-mate também está entre os produtos que vem sofrendo com fraude e adulterações para aumentar o lucro pelo fabricante. As mais comuns são a atribuição de origem geográfica inverídica e a adição de açúcar para mascarar características organolépticas indesejáveis, as quais podem ser provenientes de ação microbiológica. Além de lesar o consumidor por mascarar o produto, o açúcar ainda pode ser um risco a saúde daquele consumidor com restrições na dieta, visto que esse ingrediente não estará quantificado nas informações nutricionais (SCHNEIDER et al., 2018a; VIEIRA et al., 2019). A legislação atual determinante do Padrão de Identidade e Qualidade não estabelece limites desse ingrediente na erva-mate, facilitando a prática de adulteração (ANVISA, 2002).

2.2. ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO

Diferentemente da espectrofotometria UV-Vis, cuja análise da estrutura molecular se baseia no fenômeno da transição eletrônica resultante da alta quantidade de energia incidida, as técnicas de espectroscopia no infravermelho fazem uso de menores frequências de ondas eletromagnéticas e, consequentemente, com energia reduzida, a qual é suficiente apenas para causar as transições vibracionais de ligações químicas covalentes, limitando essa análise às moléculas que possuem esse tipo de ligação (EWING, 1972; STUART, 2004).

Sabe-se que as moléculas estão constantemente em movimento. Para que ocorra a transição vibracional, isto é, a amplificação desse movimento natural de vibração devido à absorção da energia fornecida pela radiação de infravermelho, é preciso que haja variação do momento dipolar da molécula causado pela própria vibração molecular e que tal vibração ocorra em frequência igual à onda de infravermelho (DIAS et al., 2016; STUART, 2004).

Em estruturas moleculares com momento dipolar, a vibração dos átomos da molécula altera a distância entre polos positivo e negativo. Ao incidir radiação infravermelha, a componente de frequência igual à vibração resultará no efeito de ressonância, no qual a amplitude da vibração aumenta. Dessa forma, é a interação da molécula com o campo elétrico constituinte da radiação que fará com que essa seja absorvida (DIAS et al., 2016; EWING, 1972; OLIVEIRA, 2009; STUART, 2004).

É importante ressaltar que alguns compostos cujo estado vibracional fundamental tem momento dipolar nulo, sendo inativos no infravermelho, podem se tornar detectáveis ao ocorrer um deslocamento intramolecular momentâneo causado pela própria vibração da molécula, isto é, o momento dipolar se torna não nulo (DIAS et al., 2016; OLIVEIRA, 2009; STUART, 2004).

Devido a sua especificidade de interação com as estruturas químicas presentes nos constituintes da matriz analisada, as técnicas de espectroscopia são excelentes coletoras de *fingerprints*, que são marcadores específicos capazes de possibilitar a diferenciação dos inúmeros componentes presentes na amostra. Nesse caso, esses marcadores são os picos referentes a cada tipo de vibração de determinada ligação química (ESTEKI; SHAHSAVARI; SIMAL-GANDARA, 2018; MEDINA et al., 2019).

As técnicas de espectroscopia IV se diferenciam basicamente pela faixa de comprimento de onda escolhida para análise, podendo-se optar pelo infravermelho perto, médio e distante. A região do infravermelho médio compreende o intervalo de número de onda de 4000 cm^{-1} a 400 cm^{-1}, pertencendo a essa subclasse a *Espectroscopia no Infravermelho Médio por Transformada de Fourier* (FTIR), cuja seção 1500 cm^{-1} a 600 cm^{-1} é atribuída à região de *fingerprint* dessa faixa de radiação de trabalho (ESTEKI; SHAHSAVARI; SIMAL-GANDARA, 2018; STUART, 2004). Enquanto no infravermelho médio as frequências da radiação permitem identificar as vibrações fundamentais (aquelas cuja maioria das moléculas se encontram em seu estado não excitado) com boa intensidade, no infravermelho próximo os espectros são compostos por vibrações combinadas e sobretons (vibração resultante da passagem de um modo excitado para outro de maior energia), de modo que esses espectros são menos intensos (tais vibrações não ocorrem com grande população de moléculas) e resultam em bandas alargadas (ligações de natureza diferente absorvendo na mesma região), fatores que tornam difícil a interpretação química dos espectros (SILVA, 2018; STUART, 2004; TIBOLA et al., 2018).

O preparo de amostra tem relação com o método adotado para obtenção dos espectros. Comumente se utiliza o método de transmissão, no qual se faz necessário um meio

transparente que permita a passagem da radiação de trabalho, uma vez que as energias absorvidas pela a amostra serão identificadas através da diferença entre a energia aplicada e detectada (transmitida). Quando se trabalha com amostras sólidas, costuma-se utilizar o pastilhamento como preparo pela relativa facilidade oferecida; a amostra é homogeneizada com uma substância transmissora, sendo largamente empregado o brometo de potássio como veículo, o qual não absorve na região do infravermelho médio (DIAS et al., 2016; STUART, 2004).

2.3. QUIMIOMETRIA

Um conceito usado pela primeira vez na década de 70, a quimiometria consiste em aliar recursos matemáticos, estatísticos e computacionais na tradução de dados sob a forma de sinais de detecção, provenientes de instrumentos de análise, em informações de natureza química. As possibilidades de aplicação são vastas, não estando restritas aos resultados vindos dos métodos instrumentais, podendo-se avaliar padrões e otimizar condições de trabalho a partir de respostas provenientes dos chamados métodos clássicos (FERREIRA, 2015).

A extração de conteúdo útil dos dados obtidos só é realizável por meio da identificação de padrões de comportamento das respostas frente às condições às quais as amostras foram submetidas durante o experimento. Sendo assim, o uso de ferramentas de reconhecimento de padrões se faz necessário, as quais podem ser supervisionadas ou não supervisionadas (CALLAO; RUISÁNCHEZ, 2018; FERREIRA, 2015).

Os métodos supervisionados utilizam classes pré-determinadas para reconhecer o padrão, verificando se aquela amostra se encaixa nas características daquela classe. Já nos ditos não supervisionados, classes já conhecidas não são usadas, as quais se formam durante o reconhecimento das similaridades entre amostras. Por isso, os primeiros se destinam à classificação das amostras, enquanto os últimos a uma análise exploratória para verificar se há ou não padrões de comportamento (CALLAO; RUISÁNCHEZ, 2018; FERREIRA, 2015).

2.3.1. PRÉ-TRATAMENTO DE DADOS

Diz respeito às ferramentas matemáticas usadas antes da análise dos dados experimentais com a finalidade de reduzir contribuições que não tenham natureza física, mas sim decorrentes do instrumento empregado ou de falhas por parte do laboratorista na realização da técnica. Essas informações indesejáveis podem ter impacto nos padrões reconhecidos, podendo causar interpretações equivocadas e reduzir a capacidade preditiva dos modelos construídos visto seu comportamento aleatório. Por isso, é aconselhada a aplicação

dessas ferramentas previamente ao estudo quimiométrico, contudo, essa deve ocorrer de modo a equilibrar o "custo-benefício", isto é, a distorção de informações físicas e a redução daquelas aleatórias (BRERETON, 2003; FERREIRA, 2015).

Os métodos de alisamento, como o nome sugere, tornam o sinal instrumental melhor definido pela diminuição do ruído existente, sendo por vezes responsáveis por tornar visíveis picos de baixa amplitude que se confundiam com informação ruidosa. Pertence a esses métodos o alisamento por Savitzky-Golay que necessita de uma janela e um polinômio definidos de acordo com os dados. A janela determina quantos pontos que compõe o sinal serão usados no ajuste do polinômio de grau n; a resposta do polinômio aplicado a esses pontos é um ponto alisado cuja localização é o centro da janela; então a mesma é deslocada um ponto à frente e se repete o processo. Janelas grandes fornecem maiores reduções nos ruídos dos espectros (muita informação dificulta ajustar uma curva com pouca perda de informação), assim como polinômios de menor grau (a curva por eles descrita não contempla todo o comportamento do pico). Entretanto, ambos ocasionam maiores distorções nos picos, principalmente na sua amplitude, o que influenciará na interpretação final da análise (BRERETON, 2003; FERREIRA, 2015).

O princípio de Savitzky-Golay também tem aplicação nas correções de linha de base por derivadas, sendo a janela e o polinômio usados nos cálculos das 1^a e 2^a derivadas. Enquanto a primeira derivada corrige apenas o deslocamento da linha de base, a segunda derivada ajusta seu deslocamento e inclinação, contribuindo ainda no aumento da resolução de picos sobrepostos (separação) (FERREIRA, 2015; VIEIRA et al., 2019).

A correção da linha de base pode ser feita ainda pelo chamado ajuste de função, no qual se escolhem pontos do espectro pelos quais se determinará uma curva que será a linha de base. Os demais pontos do espectro são trazidos para essa linha de base verificando a diferença entre ambas as curvas. É comum a utilização de funções lineares, as quais originam seguimentos de retas que serão unidos resultando na linha de base (FERREIRA, 2015).

Outro pré-tratamento é a normalização dos dados, cujo uso é indispensável em estudos qualitativos, uma vez que essa técnica permite eliminar a contribuição quantitativa da amostra sobre o sinal, sendo a única forma para tal quando as condições do experimento não permitem ao analista garantir o controle desses fatores por si só (tem-se, por exemplo, o uso de amostras de massas muito pequenas de difícil precisão de pesagem). Nesse método, cada sinal que compõe o espectro da amostra será dividido pela norma de amostra, sendo a norma euclidiana uma das formas de cálculo, a qual consiste na retirada da raiz quadrada da soma de todos os dados, referentes à amostra, elevados ao quadrado (FERREIRA, 2015).

2.3.2. ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais (PCA) pertence aos métodos não supervisionados, atribuindo-se sua maior funcionalidade à capacidade de simplificação de problemas de elevada complexidade através da redução do espaço multivariado que o descreve. Para isso, são realizadas combinações lineares entre as variáveis para as quais o comportamento da amostra está sendo estudado, de modo que aquelas mais correlacionadas são substituídas por uma única, chamada componente principal (PC), sem perda de seus efeitos sobre as amostras (BRERETON, 2003; FERREIRA, 2015; GRANATO et al., 2018).

Essas componentes principais são independentes entre si, não ocorrendo atribuição de informação repetida durante o reconhecimento dos padrões de comportamento. Cada PC é definida de modo a maximizar a representação da variância dos dados experimentais, sendo assim são estabelecidas de maneira que a próxima explique a variância que a anterior não abordou. As PC's formam o novo espaço multivariado reduzido, cujos atributos escores e pesos revelam, respectivamente, a contribuição de cada PC nos dados e a contribuição de cada variável original na obtenção da PC (BRERETON, 2003; FERREIRA, 2015; GRANATO et al., 2018).

2.3.3. ANÁLISE DE AGRUPAMENTOS HIERÁRQUICOS

Semelhante a PCA, a análise de agrupamentos hierárquicos (HCA) também é um método não supervisionado, porém ele verifica quão parecida cada amostra é entre si, atribuindo níveis se semelhança (FERREIRA, 2015; GRANATO et al., 2018). Parte-se do princípio de que cada amostra constitui um grupo com características próprias e uniformes, sendo os grupos os mais distintos possíveis entre si (maximização da homogeneidade interna e da heterogeneidade entre grupos); então, iterativamente ocorrerá verificação da semelhança entre dois grupos (duas amostras), cujo resultado gráfico é o dendrograma que une os grupos em níveis partindo da maior similaridade ou dissimilaridade, dependendo do método de agrupamento adotado (FERREIRA, 2015).

O parâmetro utilizado para avaliar a similaridade entre duas amostras é a distância entre elas no espaço multivariado, sendo seu uso coerente visto que "coisas" parecidas estarão distribuídas próximas no espaço por terem valores próximos para aquelas características que constituem esse espaço. A forma matemática mais conhecida e utilizada no cálculo da distância entre dois pontos é usar suas coordenadas no Método Euclidiano que consiste em somar as diferenças entre as suas coordenadas no espaço multivariado e extrair a raiz

quadrada. Entretanto, nas análises quimiométricas é mais indicada a distância de Mahalanobis, dada sua vantagem por normalizar os dados, eliminando o efeito de escala que pode causar interpretações equivocadas, o que é negligenciado pela Métrica Euclidiana, na qual a distância possui unidade de medida. Portanto, a distância de Mahalanobis é uma distância adimensional e elíptica, por isso, qualquer amostra localizada sobre a elipse está a uma mesma distância daquela no seu ponto médio (FERREIRA, 2015).

Por fim, escolhe-se o método para agrupar sendo o mais simples o Método do Vizinho Mais Próximo, no qual o agrupamento ocorre considerando os grupos mais próximos no espaço multivariado através do cálculo da distância média, isto é, considera-se no calculo a distância de todos os membros de um grupo com outro. Outra técnica de maior complexidade é o Método Ward, no qual a variância entre grupos é considerada no agrupamento, de modo que se visa minimizá-la, ou seja, encontrar a máxima similaridade entre os grupos (FERREIRA, 2015; GRANATO; KARNOPP; VAN RUTH, 2015).

2.3.4. ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS

Esta é uma metodologia supervisionada que possibilita a separação das amostras em grupos bem definidos a partir da regressão linear por mínimos quadrados parciais, assim, há adequação de uma técnica de avaliação quantitativa para estudos qualitativos. Semelhante à PCA, o método procura um novo espaço multivariado através das melhores combinações entre as variáveis originais, o qual será descrito pelas chamadas variáveis latentes. Cada variável latente maximiza a variância contida nos dados experimentais e na previsão da variável dependente (a classe à qual pertence a amostra), objetivando tornar o modelo com maior poder discriminante (FERREIRA, 2015; GROMSKI et al., 2015).

Para a regressão, há construção de um vetor no qual para cada linha da matriz dados há uma linha correspondente à classe que essa amostra pertence. Logo, esse é um vetor resposta, ou seja, onde estão as soluções que se têm interesse de prever, com o mínimo de erro, a partir do modelo matemático delineado. O problema a ser contornado está em como comparar o valor predito com o real, pois o primeiro é um número real enquanto o segundo não, podendo ser até mesmo uma *string*. Utilizam-se então métodos para determinação de valores de corte, assumindo-se que a amostra pertence a uma classe quando o valor predito é superior ao de corte e a outra quanto é inferior (FERREIRA, 2015; GROMSKI et al., 2015).

A capacidade de previsão do modelo é avaliada por meio da aplicação de figuras de mérito adequadas a parâmetros qualitativos. É comum o uso das figuras de mérito a cerca do erro existente entre os valores previstos e observados, como o RMSEP e RMSEC (raiz

quadrada do erro quadrático médio de previsão e calibração, respectivamente) em métodos de regressão, contudo, esses não são recomendados na avaliação de modelos PLS-DA, visto que o valor previsto é quantitativo enquanto o valor real é qualitativo, tendo-se assim inconsistência (CALLAO; RUISÁNCHEZ, 2018; FERREIRA, 2015; GROMSKI et al., 2015). Os RMSE são calculados com base na diferença entre a variável resposta prevista pelo modelo e seu valor real (aquele obtido em laboratório) de cada conjunto de variáveis, calibração ou previsão, conforme Equação 1 (MABOOD et al., 2017). Como para realizar a regressão se utilizou de codificações numéricas que atribuíram caráter quantitativo à resposta qualitativa, são esses os valores quantitativos usados nos cálculos dos erros padrões médios. Assim esses parâmetros acabam por dar uma aproximação de quão bem delimitada estão as classes no modelo: quanto mais definidas, menos alocações em classes erradas o modelo faz (MABOOD et al., 2017; VALDERRAMA; VALDERRAMA, 2016).

$$\text{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}} \qquad (1)$$

Onde n é o número de variáveis do conjunto avaliado; $\hat{y}_i$ é o valor previsto da variável resposta e y_i o valor real da variável resposta.

2.3.5. FIGURAS DE MÉRITO

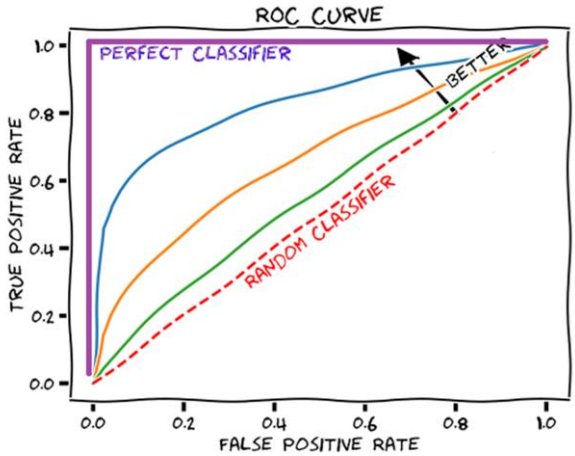
As figuras de mérito são parâmetros de natureza estatística usados na verificação do desempenho do modelo de regressão arquitetado, isto é, se o mesmo cumpre com a sua função de maneira confiável (FERREIRA, 2015).

Para essa validação do modelo preditivo é preciso eleger de forma criteriosa os parâmetros que serão avaliados, sendo essencial considerar se este tem função quantitativa ou qualitativa. Dentre as figuras de mérito comumente usadas na validação de modelos qualitativos estão a sensibilidade, seletividade (especificidade) e a área abaixo da *ROC curve* (MENDES et al., 2019).

A sensibilidade expressa a capacidade do modelo de identificar as amostras de uma determinada classe como pertencendo a mesma, enquanto a seletividade reflete sua capacidade de reconhecer as amostras que não são daquela classe e não atribuí-las a mesma (CALLAO; RUISÁNCHEZ, 2018; MENDES et al., 2019). A combinação de ambas resulta em uma terceira figura de mérito, a área abaixo da *ROC curve*, a qual provém da plotagem da

seletividade *vs* sensibilidade. Esse parâmetro varia de 0 a 1, sendo que quanto mais próximo de 1 é a área, mais confiável é o modelo, pois significará que para cada amostra classificada corretamente, não houve amostras alocadas na classe errada (Figura 2). Logo, quando a área é nula, tem-se que para cada amostra classificada corretamente, outra foi atribuída à classe errada, fazendo-se do modelo uma ferramenta inútil (MENDES et al., 2019).

Figura 2: Demonstração visual do parâmetro AUC para entendimento de seu significado. A área abaixo da curva é delimitada pela bissetriz, a qual corresponde à situação em que o modelo não é indicado para classificação, uma vez que o percentual de amostras classificadas correta e incorretamente é o mesmo.



Fonte: DRAELOS (2019).

3. METODOLOGIA

3.1. PREPARO DE AMOSTRAS

Foram avaliadas 69 amostras de erva-mate para chimarrão produzidas nos três estados brasileiros da região Sul: Paraná, Santa Catarina e Rio Grande do Sul. Marcas comerciais foram adquiridas por VIEIRA et al. (2019) em mercados dessas localidades, incluindo produtos sem adição de açúcar e com adição de acordo com o informado nas embalagens. Ainda com base na rotulagem, obteve-se a origem geográfica da erva utilizada no produto em questão. A amostragem realizada está disposta na Tabela 1.

Tabela 1: Relação quantitativa das amostras de erva-mate de acordo com o estado de origem e presença de açúcar.

Distribuição das Amostras por Classe			
	PR	SC	RS
Com Açúcar	4	3	11
Sem Açúcar	17	14	20
Total	**21**	**17**	**31**

PR = Paraná; SC = Santa Catarina; RS = Rio Grande do Sul.
Fonte: Elaborada pelo autor.

As amostras permaneceram congeladas e embaladas em filmes de polietileno selados, por isso antes de realizar o preparo das pastilhas de KBr, separou-se cerca de 1g das amostras em béqueres de 5mL, os quais foram mantidos por 24h em dessecador com sílica gel a fim de estarem livre de umidade, uma vez que esta interfere de modo considerável no espectro obtido (DIAS et al., 2016; STUART, 2004).

3.2. ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO POR TRANSFORMADA DE FOURIER (FTIR)

Primeiramente, realizou-se a secagem do brometo de potássio (Sigma – Aldrich, grau de pureza maior que 99% para espectroscopia FTIR) em estufa para eliminação da umidade ambiente absorvida por esse sal devido a sua elevada higroscopicidade. Para isso, reservou-se uma pequena fração do pó (suficiente para trabalho) em um béquer 50mL, tampando-o com papel alumínio no qual se fez alguns furos para possibilitar a saída do vapor de água. Então, o béquer seguiu para a estufa de secagem, sob temperatura de $105^{o}C \pm 5^{o}C$ por em média 12h (DIAS et al., 2016; STUART, 2004). Quando se notavam grandes aglomerados do sólido, este era macerado em almofariz de ágata antes de seguir para secagem, a fim de aumentar a

superfície de contato com o ar quente garantindo a maior eliminação de água do interior do sal.

Feito isso, pesaram-se 200mg de KBr em balança analítica sobre uma base de folha de papel alumínio devidamente dobrada, cuja escolha se deu pela facilidade de transferência da massa de trabalho com reduzida quantidade aderida ao suporte de pesagem (STUART, 2004).

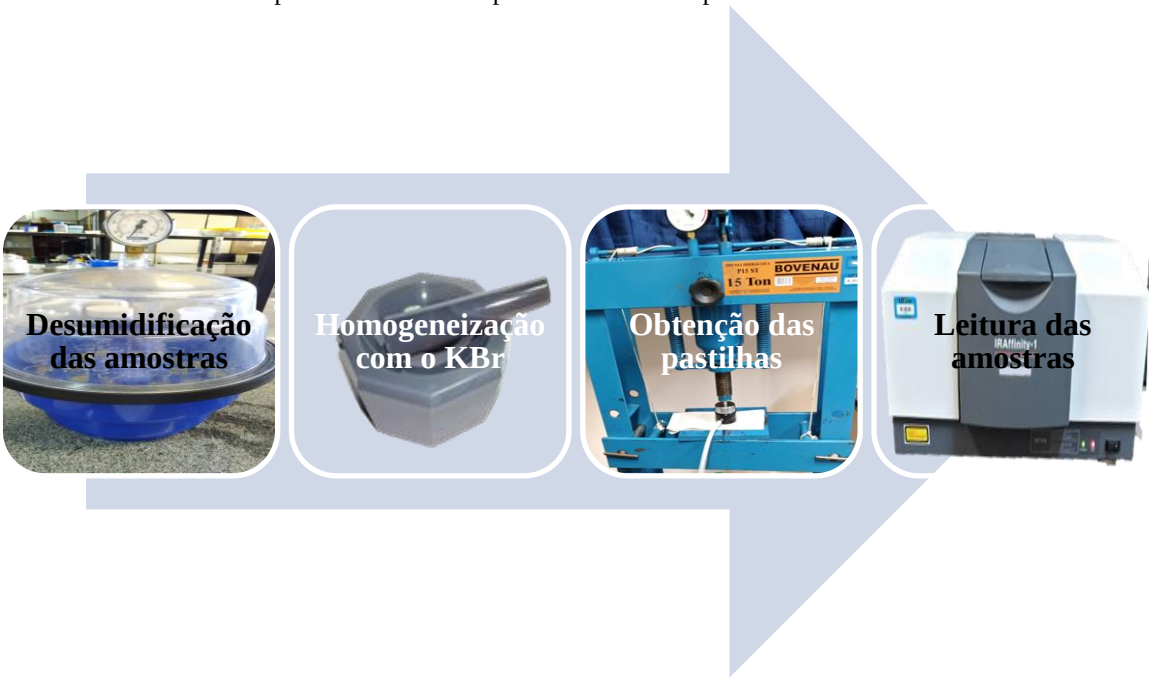
Com algodão embebido em clorofórmio limparam-se o pistilo e almofariz de ágata, passando para este os 200mg de KBr. Na sequência, colocou-se uma ponta de espátula da amostra, algo próximo a 1% m/m, e se realizou maceração com auxílio do pistilo para homogeneização da amostra no KBr e redução da granulometria de ambos (DIAS et al., 2016). Vale ressaltar que não se pesou a amostra agregada ao KBr, pois o intuito dessa análise espectroscópica foi qualitativa, ou seja, verificar padrões de picos nos espectros, e não quantitativa, na qual se correlaciona a altura do pico com a concentração daquela estrutura na amostra (DIAS et al., 2016; STUART, 2004).

A partir do auxílio da espátula, limpa com clorofórmio, houve separação do macerado em duas partes, sendo uma dessas transferida para o pastilhador. Neste, a amostra é alocada entre duas anilhas, assim, a fração foi espalhada com o pistão do conjunto de pastilhar sobre a anilha base, colocando-se então a segunda anilha. Acoplado a uma bomba de vácuo Prismatec, o conjunto seguiu para prensa Bovenau P15 ST sobre o qual se aplicaram 356,3kgf/cm^2 de pressão por 32s para obtenção de uma pastilha translúcida (DIAS et al., 2016; STUART, 2004).

Concomitantemente, configurou-se o espectrômetro Shimadzu IRAffinity-1 para realizar 32 varreduras com resolução 1cm^{-1}, função de apodização Happ-Genzel e faixa de número de onda de 4000cm^{-1} a 400cm^{-1} (STUART, 2004). Uma leitura sem amostra foi realizada para aquisição do *background*, que é o espectro referente aos constituintes do ar do local de trabalho, incluindo sua umidade, cujo efeito é descontado do espectro referente à amostra sendo, portanto, retirado tal efeito (ruído) (DIAS et al., 2016; OLIVEIRA, 2009; STUART, 2004).

Extraída a pastilha com aplicação de força moderada sobre o pistão do pastilhador, esta foi repousada sobre o suporte com ímãs (acessório do espectrômetro) e inserida no FTIR. Fez-se a leitura em percentual de absorbância e obteve-se o espectro. Todo o processo foi repetido para a segunda porção de macerado, tendo-se, portanto, análises realizadas em duplicata.

Figura 3: Esquematização das etapas descritas para o preparo e leitura das amostras de erva-mate no espectrômetro FTIR a partir da técnica de pastilhamento.



Fonte: Autoria própria (2019).

3.3. APLICAÇÃO DAS TÉCNICAS QUIMIOMÉTRICAS

Todas as análises quimiométricas foram realizadas no programa Matlab R2008b utilizando o pacote de rotinas GAMMA desenvolvido para análises quimiométricas em matrizes alimentares (BONA, 2019).

3.3.1. PRÉ-TRATAMENTO DOS ESPECTROS

Os dados contidos nos espectros foram dispostos em uma matriz 138x3735, cujas linhas correspondiam às amostras (seguidas de suas duplicatas), enquanto as colunas às frequências de onda nas quais se mediu a transmitância. Por inspeção visual, a região de *fingerprint* foi determinada como sendo $1900cm^{-1}$ – $500cm^{-1}$, o que reduziu a quantidade de informação a ser analisada resultando em uma nova matriz 138x1453.

O primeiro pré-tratamento aplicado foi o alisamento dos espectros pelo método de Savitzky-Golay empregando uma janela de tamanho 19 e um polinômio de 1^{o} grau, uma vez que essa se mostrou a melhor combinação para reduzir o ruído sem perder informação dos picos para esses dados (FERREIRA, 2015).

Depois, os espectros alisados passaram por normalização para a qual se usou a norma Euclidiana. Eliminadas as contribuições quantitativas, realizou-se a correção da linha de base através do ajuste de função (FERREIRA, 2015).

3.3.2. HCA

A análise por agrupamentos hierárquicos se deu pelo método Ward aplicado à distância euclidiana, para a qual se usou a matriz contendo as repetições (138x1453). As distâncias foram normalizadas em relação à maior distância observada, obtendo-se uma escala de dissimilaridade indo de 0 (menor diferença) a 1 (maior diferença) no dendrograma.

Sua utilização se deu a fim de identificar as replicatas que não apresentaram espectros similares, as quais se tornaram candidatas à nova análise, confirmando-se essa necessidade através de inspeção gráfica da plotagem de ambos os espectros.

A análise foi repetida para os espectros médios a fim de observar as similaridades entre as amostras e verificar se as pertencentes a um mesmo grupo teriam maior semelhança no dendrograma.

3.3.3. PCA

Uma rotina desenvolvida no Matlab R2008b foi usada para obter os espectros médios das duplicatas, tendo-se agora uma matriz 69x1453 sobre a qual aplicou a análise exploratória por componentes principais, cujo intuito foi verificar se esse método de reconhecimento de padrões já seria capaz de separar as amostras em classes, ou seja, se haveria formação das classes *origem geográfica* “Paraná”, “Santa Catarina”, “Rio Grande do Sul” e *conteúdo de açúcar* “com açúcar” e “sem açúcar”.

3.3.4. PLS-DA

Foram criados dois vetores 69x1 que continham em cada linha a classe cuja respectiva amostra pertencia: um referente à origem geográfica contendo a informação ‘p’, ‘s’ e ‘r’, correspondente a “Paraná”, “Santa Catarina” e “Rio Grande do Sul”, e outro à presença de açúcar, contendo ‘c’ e ‘s’ que representavam “com” e “sem açúcar”, nessa ordem. Aplicado o PLS-DA, teve-se como resultado dois modelos de classificação: um para origem da erva-mate e outra para o conteúdo de açúcar.

O algoritmo de Kennard-Stone foi aplicado para seleção das variáveis que iriam compor os conjuntos de calibração e previsão usados na obtenção dos modelos classificatórios e suas validações (FERREIRA, 2015; VIEIRA et al., 2019). Cada classe teve ¾ de seus representantes destinados ao conjunto de calibração, estando o restante no conjunto de previsão (Tabela 2).

Para validar os modelos foram empregadas validação cruzada e as figuras de mérito sensibilidade (SEN), seletividade (SEL), área abaixo da *ROC curve* (AUC).

Tabela 2: Quantificação das amostras selecionadas pelo algoritmo de Kennard-Stone para os conjuntos de calibração de previsão de ambos os modelos classificatórios.

Origem			
	PR	SC	RS
Quantidade de amostra para calibração	16	13	23
Quantidade de amostra para previsão	5	4	8
Conteúdo de açúcar			
	Com açúcar		Sem açúcar
Quantidade de amostra para calibração	14		38
Quantidade de amostra para previsão	4		13

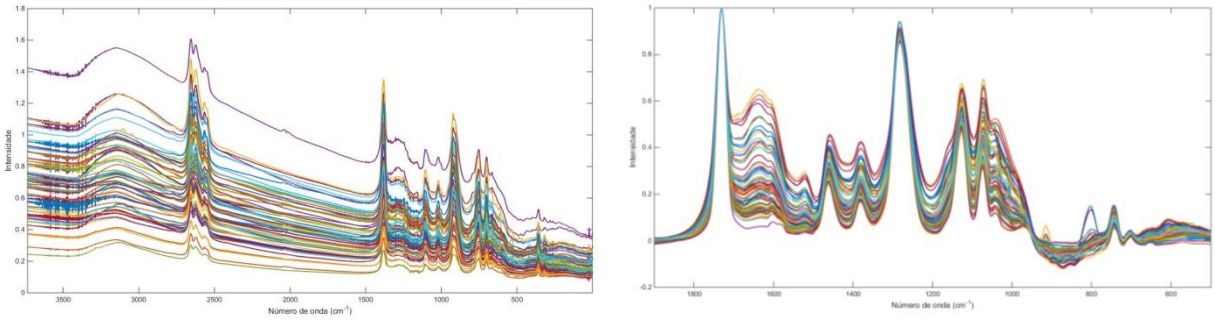
Fonte: Autoria própria (2019).

Na sequência, a matriz das variáveis independentes passou por outro pré-tratamento, a correção da linha de base dos espectros pela 2ª derivada, e então se repetiram as etapas de seleção dos conjuntos de calibração e previsão, obtenção de modelos discriminativos por PLS-DA e validação dos mesmos, a fim de verificar se a adição desse passo aprimoraria os modelos de classificação. Novamente se fez uso do método de deslocamento de janela de Savitzky-Golay, agora aplicado no cálculo das 2ª derivadas. A combinação adotada foi janela de tamanho 29 e polinômio de grau 3, por ter se mostrado a mais equilibrada na redução de ruído e manutenção do perfil dos picos (FERREIRA, 2015).

4. RESULTADOS E DISCUSSÃO

A Figura 4-a apresenta o espectro médio das 69 amostras para toda faixa de comprimento de onda utilizada sem aplicação de pré-tratamento, enquanto que na Figura 4-b se tem os espectros médios seccionados na faixa que continha informação pertinente (*fingerprint*) e pré-tratados por alisamento, normalização e correção da linha de linha por ajuste de função.

Figura 4: Espectros FTIR das 69 amostras de erva-mate estudadas. (a) Espectro completo de todas as amostras e suas replicatas sem pré-tratamento. (b) Espectros médios cortados na região de interesse – fingerprint – e pré-tratados, porém sem aplicação da segunda derivada.

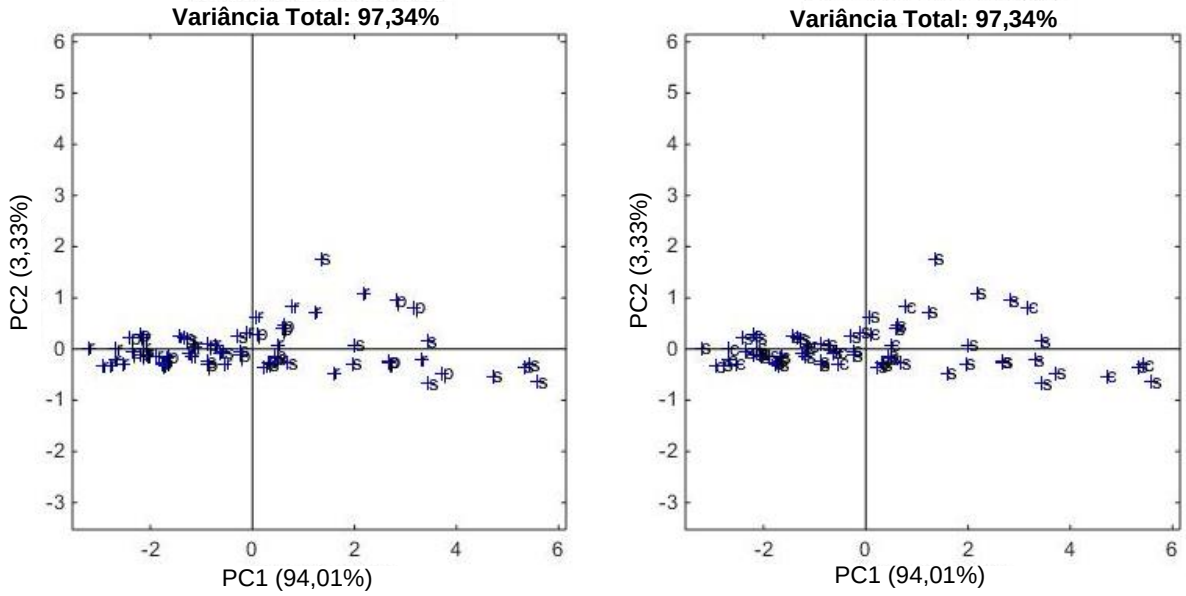


(a) (b)

Fonte: Autoria própria (2019).

Sob os dados transformados, a aplicação da análise por componentes principais mostrou a necessidade de apenas duas componentes para explicar toda a variância dos dados, sendo a primeira PC a responsável pela maior parte dessa, indicando vínculo com a real natureza dos dados, enquanto que a segunda PC está, provavelmente, indicando variações vindas da execução do experimento (Figura 5). Entretanto, nota-se pela Figura 5 que a análise exploratória não foi eficiente na indicação dos padrões de interesse (origem geográfica e presença de açúcar), uma vez que não se observa separação das amostras de mesma classe em nenhuma das PCs. Uma vez que os métodos não supervisionados criam as classes a partir de suas semelhanças, as amostras foram identificadas e legendadas nos gráficos para verificar se houve ou não o agrupamento que se esperava.

Figura 5: Gráfico de escores das duas PCs necessárias para descrever a variância total entre as amostras. (a) Amostras identificadas pela região de origem (b) Amostras identificadas pela presença de açúcar.

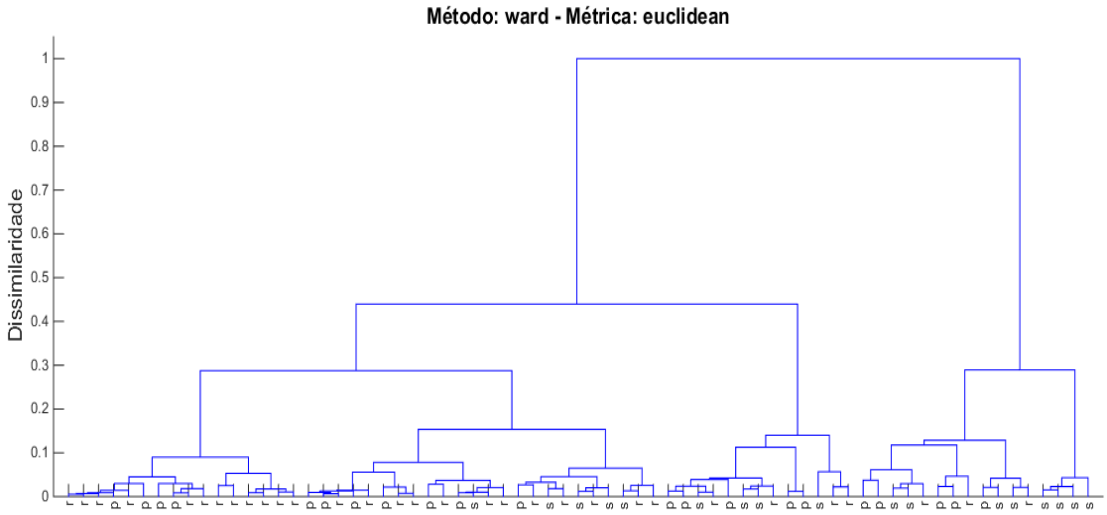


(a) (b)

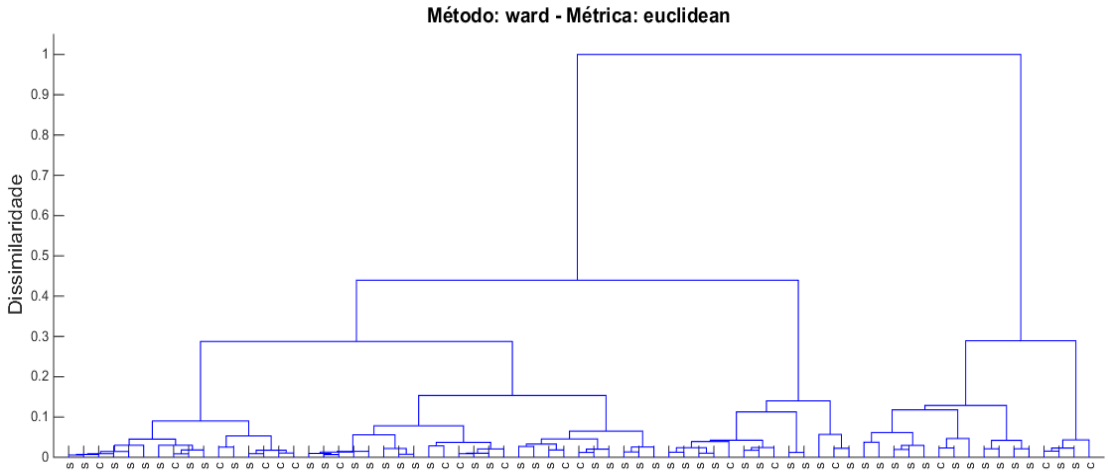
(a) P = Paraná; S = Santa Catarina; R = Rio Grande do Sul (b) C = com açúcar; S = sem açúcar.
Fonte: Autoria própria (2019).

Pela análise por agrupamentos hierárquicos, pode-se visualizar a semelhança entre as amostras, reforçando a incapacidade da PCA em separá-las. É notório na Figura 6 como ervas pertencentes a grupos diferentes estão sendo reconhecidas com alto grau de similaridade.

Figura 6: Dendrogramas dos espectros médios identificando os agrupamentos pelas classes (a) origem geográfica (b) presença de açúcar.



(a)

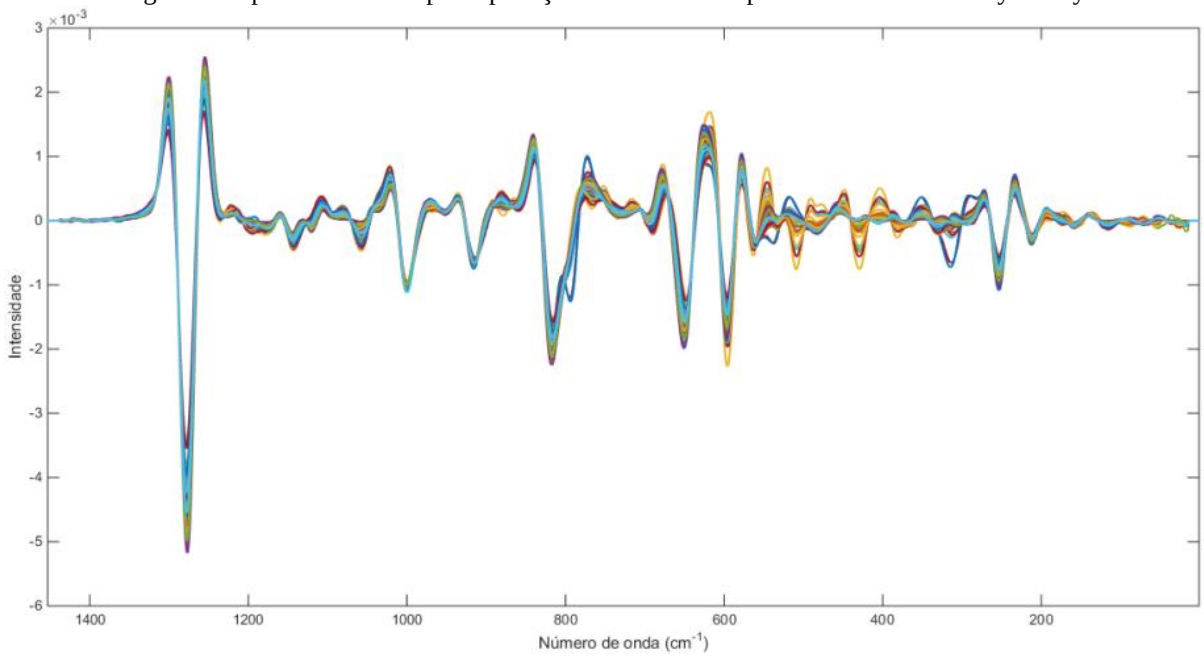


(b)

Fonte: Autoria própria (2019).

Na sequência foram obtidos os espectros com a linha de base corrigida pela 2ª derivada (Figura 7), nos quais se observa a real diferença de intensidade dos picos entre as amostras já que os mesmos variam em torno de uma mesma reta (linha de base).

Figura 7: Espectros tratados pela aplicação da 2a derivada pelo método de Savitzky-Golay.

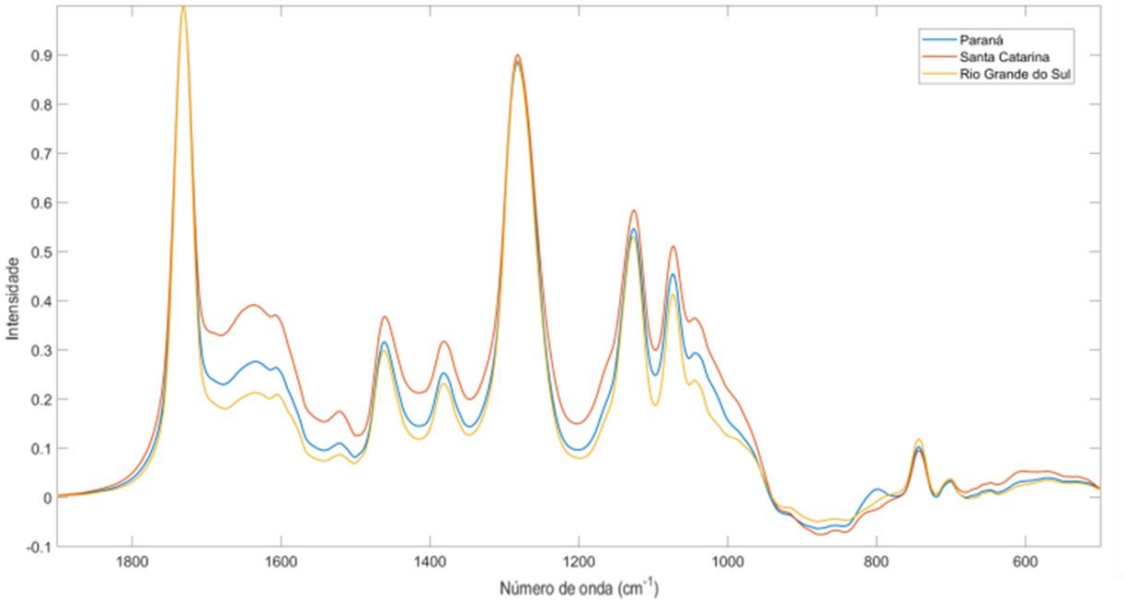


Fonte: Autoria própria (2019).

4.1. DISCRIMINAÇÃO GEOGRÁFICA

A Figura 8 traz os espectros médios dos estados de origem das ervas. Nota-se que o que varia entre os espectros são as intensidades dos picos, enquanto que a forma descrita é mantida. Esse comportamento também foi observado por VIEIRA et al. (2019) para os espectros NIR.

Figura 8: Espectros médios por estado de origem.



Fonte: Autoria própria (2019).

Os modelos de classificação por estado pelo método PLS-DA a partir dos dados com pré-tratamento parcial (sem aplicação da 2ª derivada) e completo (ajuste de linha de base pela 2ª derivada), mostraram diferentes complexidades. Os resultados de sua validação estão dispostos na Tabela 3. Para os dados não derivados, foram necessárias 20 variáveis latentes para descrever a máxima variância intrínseca no perfil dos espectros das amostras e dentro das classes já delimitadas e com os membros conhecidos. Em contrapartida, reduziu-se para 15 variáveis latentes o espaço multivariado do modelo quando os dados estavam derivados à segunda, caracterizando simplificação do problema pela maior redução de informação de baixo impacto.

Com base nas figuras de mérito, conclui-se que ambos os modelos possuem boa capacidade de classificação com sensibilidades acima de 60% e seletividades acima de 80% para os três estados. Esses parâmetros reafirmam a superioridade do modelo composto por 15 VL frente ao de 20 VL, visto que mesmo tendo reduzido as informações utilizadas para a previsão conseguiu manter a sensibilidade e seletividade preditivas para o Paraná e Rio

Grande do Sul e elevá-las para as amostras de Santa Catarina, o que também proporcionou leve aumento do parâmetro AUCP, o qual se mostrou próximo a 1 (0,89).

O estudo realizado por VIEIRA et al (2019) a cerca dessas amostras, mas aplicando a técnica NIR, resultou em um modelo classificatório mais enxuto, com 12 VL, mas com menor capacidade discriminante, apresentando sensibilidade de previsão 50% para Paraná, 80% para Santa Catarina e 100% Rio Grande do Sul, e seletividade 77% para Paraná, 93% Santa Catarina e 73% Rio Grande do Sul. O RMSEP foi superior ao do presente estudo: 0,4560.

A partir dos resultados observados, esses autores sugeriram com base em outras análises, que a dificuldade em classificar amostras do Paraná se deu pela possível adulteração das ervas de Santa Catarina e Rio Grande do Sul com esse produto (VIEIRA et al., 2019). Assim, o modelo diferenciava muito bem essas últimas mesmo adulteradas, pois as características intrínsecas de seus estados possibilitaram agrupá-las, enquanto que para as amostras do Paraná, determinar o padrão é mais complexo, pois ele se confunde às características que, teoricamente, também seriam dos demais estados.

Porém, de acordo com a Tabela 3, para o conjunto de dados FTIR foram as amostras do Rio Grande do Sul que se mostraram as mais complexas de terem seu padrão identificado, e não o Paraná.

Tem-se um bom modelo de classificação regional ao considerar a área abaixo da curva ROC próximo a 0,9 e os altos valores de sensibilidade e seletividade em geral. O que mais deixou a desejar foi a capacidade de identificação correta das amostras do Rio Grande do Sul, contudo, esta ainda é maior que 50% o que não foi alcançado por VIEIRA et al (2019) para a sua classe menos sensível.

Esperava-se que o modelo seria descrito por menos variáveis latentes que as utilizadas por VIEIRA et al (2019), isso porque a quantidade de informação contida nos espectros FTIR é muito maior e mais específica do que a provinda da análise NIR, uma vez que a frequência do infravermelho médio e por consequência a quantidade de energia carregada pelos comprimentos de onda é superior, de modo que resulta em transições vibracionais próprias e bem delimitadas (DIAS et al., 2016; OLIVEIRA, 2009). Contudo, é preciso lembrar das limitações do método de regressão linear PLS utilizado na análise discriminante, o que justifica a dificuldade de simplificação do modelo.

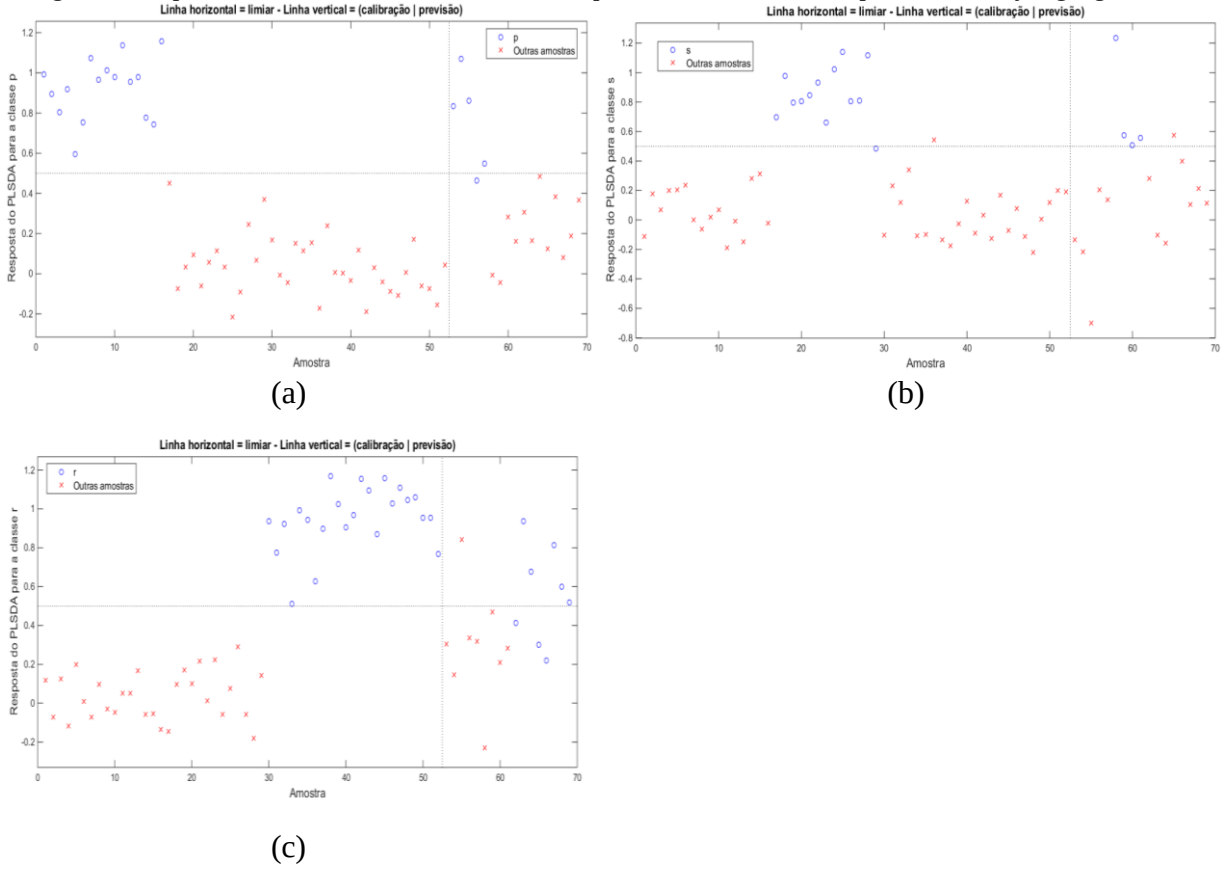
Tabela 3. Modelos PLS-DA obtidos para classificação das amostras de erva-mate quanto a origem geográfica.

Pré-tratamento	LV	Classe	RMSEC	RMSEP	AUCP	Calibração		Previsão	
						Sensibilidade	Seletividade	Sensibilidade	Seletividade
Alisamento + Normalização + Correção da Linha de Base	20	PR	0.1576	0.4158	0.8801	1.00	1.00	0.80	1.00
		SC				0.92	1.00	0.75	0.85
		RS				1.00	1.00	0.62	0.89
lisamento + Normalização + Correção da Linha de Base + 2ª Derivada	15	PR	0.1657	0.3633	0.8965	1.00	1.00	0.80	1.00
		SC				0.92	0.97	1.00	0.92
		RS				1.00	1.00	0.62	0.89

LV: Variáveis Latentes; C: “com açúcar”; S: “sem açúcar”; RMSEC: raiz do erro quadrático médio de calibração; RMSEP: raiz do erro quadrático médio de previsão; AUCP: área abaixo da curva ROC.
Fonte: Autoria própria (2019).

As respostas dos modelos PLS-DA com 15 variáveis latentes, isto é, os agrupamentos realizados pelo método supervisionado, podem ser vistos na Figura 9. A já comentada dificuldade de classificação das amostras do Rio Grande do Sul fica evidenciada na Figura **9**-c, na qual se observa que apenas 1 amostra não pertencente a esse estado é atribuída ao mesmo, em contrapartida 3/8 das amostras do conjunto de previsão que deveriam ter sido colocadas na classe Rio Grande do Sul não o foram.

Figura 9: Respostas dos modelos PLS-DA definidos por 15 variáveis latentes para classificação geográfica.



p = Paraná; s = Santa Catarina; r = Rio Grande do Sul.
Fonte: Autoria própria (2019).

4.2. DISCRIMINAÇÃO QUANTO A PRESENÇA DE AÇÚCAR

Tal qual a classificação por região de origem, os parâmetros dos modelos selecionados para discriminação com base na presença de açúcar estão tabelados (Tabela 4). Para os dados não derivados elegeu-se o modelo com 8 variáveis latentes, por ter apresentado menores RMSE e maiores sensibilidade, seletividade e AUCP, enquanto para as informações pós derivação isso se evidenciou no modelo descrito com 10 variáveis.

Contrapondo as figuras de mérito de ambos os modelos, percebe-se que o uso da segunda derivada nos dados experimentais não teve efeito positivo esperado, como observado na discriminação regional, isto é, sua aplicação resultou em um modelo preditivo de maior complexidade (necessidade de adição de 2 VL), porém com queda de qualidade preditiva, visto que a sensibilidade para a classe "sem açúcar" e a seletividade para a classe "com açúcar" sofreram redução, o que justifica a redução em AUCP.

É interessante notar a baixa sensibilidade de calibração para a classe "com açúcar" e maior sensibilidade de previsão no modelo com 8 VL, o que indica possível adulteração de amostras por adição de açúcar: apenas 57% das amostras do conjunto de calibração que se sabiam conter açúcar foram classificadas corretamente durante a construção do modelo, logo, ao restante se atribuiu a identificação "não pertence à classe" que é o mesmo, em uma classificação binária, que dizer que a amostra pertence à classe oposta, nesse caso à classe sem açúcar. Se às amostras cuja existência de sacarose era conhecida, o modelo a atribuiu a classificação "sem açúcar", isso aconteceu porque havia amostras com características similares, mas conhecidas como "sem açúcar".

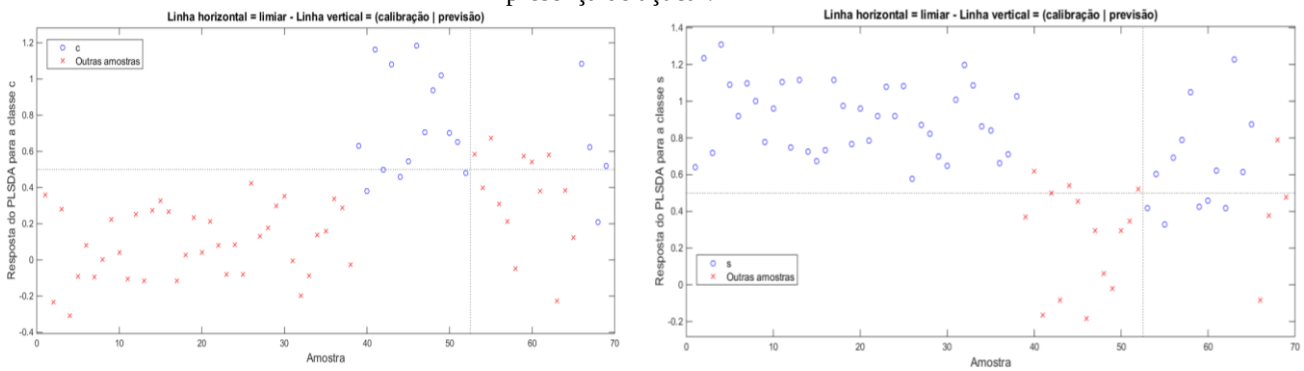
Ainda a respeito das figuras de mérito da calibração dos modelos, para o segundo caso, nota-se que os dados derivados resultaram em aumento da sensibilidade da classe "com açúcar". Contudo, isso ocorreu em detrimento da sensibilidade da classe "sem açúcar" que passou de 77% do primeiro modelo para 62%. Como a mesma seleção de calibração e previsão foram empregadas em ambos os modelos, sugere-se que a correção de linha de base por 2^{a} derivada resultou em um modelo sobre-ajustado.

VIEIRA e colaboradores (2019) obtiveram resultados mais expressivos para a identificação de açúcar nas amostras de erva-mate, contando com modelos com sensibilidade e seletividade de previsão de 100% para ambas as classes e RMSEP abaixo de 0,2000 delimitados por apenas 5 variáveis latentes. Assim, levanta-se a hipótese de que as informações experimentais contidas nos espectros NIR estejam mais relacionadas às interações resultantes da adição da componente sacarose nos demais constituintes da matriz da erva-mate.

O modelo classificatório relativo à presença de açúcar delimitado apresentou resultados insatisfatórios em relação aos alcançados pelos autores. Independente da aplicação da 2ª derivada, a quantidade de variáveis latentes permaneceu maior do que a utilizada pelos autores, o que poderia ser aceito se o modelo apresentasse a mesma sensibilidade e seletividade de 100%. Dessa forma, obteve-se apenas um modelo de maior complexidade e menor capacidade preditiva, não importando os pré-tratamentos utilizados. Sendo assim, a faixa experimental do infravermelho médio talvez não seja representativa das interações do açúcar com a matriz alimentícia, o que impossibilitou o delineamento de um modelo de melhor qualidade discriminante.

A Figura 10 traz graficamente as respostas dos modelos PLS-DA delineados por 10 variáveis latentes.

Figura 10: Respostas dos modelos PLS-DA descritos por 10 variáveis latentes para classificação quanto a presença de açúcar.



(a) (b)

c = Com açúcar; s = Sem açúcar.
Fonte: Autoria própria (2019).

Tabela 4. Modelos PLS-DA obtidos para classificação das amostras de erva-mate quanto a presença de açúcar.

Pré-tratamento	LV	Classe	RMSEC	RMSEP	AUCP	Calibração		Previsão	
						Sensibilidade	Seletividade	Sensibilidade	Seletividade
Alisamento + Normalização + Correção da Linha de Base	8	C	0.3481	0.4553	0.7115	0.57	0.97	0.75	0.77
		S				0.97	0.57	0.77	0.75
Alisamento + Normalização + Correção da Linha de Base + 2ª Derivada	10	C	0.2615	0.4479	0.6923	0.71	1.00	0.75	0.62
		S				1.00	0.71	0.62	0.75

LV: Variáveis Latentes; C: “com açúcar”; S: “sem açúcar”; RMSEC: raiz do erro quadrático médio de calibração; RMSEP: raiz do erro quadrático médio de previsão; AUCP: área abaixo da curva ROC.

Fonte: Autoria própria (2019).

5. CONCLUSÃO

A análise de ervas-mate pela técnica instrumental de espectroscopia de infravermelho médio FTIR e interpretação dos dados experimentais pelo método PLS-DA (análise discriminante por mínimos quadrados parciais) possibilitou a obtenção de bons modelos de previsão ao avaliar suas figuras de mérito. A quantidade de variáveis latentes utilizadas nos modelos se justifica pela lógica simplificada do método supervisionado PLS-DA, a qual impossibilita maior redução do conteúdo de informação original sem maiores perdas de capacidade de predição dos modelos.

Para a discriminação geográfica, o modelo mais simples e com melhor qualidade preditiva foi alcançado empregando 15 variáveis latentes no novo espaço amostral e utilizando-se os dados tratados pela aplicação da segunda derivada. Contudo, este ainda apresentou reduzida capacidade de identificação de amostras do Rio Grande do Sul.

Já para a classificação referente à presença de açúcar, a metodologia também se mostrou possível de ser realizada através dos modelos obtidos. Entretanto, diferentemente do modelo para origem geográfica, os modelos para presença de açúcar se apresentaram mais simplificados e de melhor qualidade ao empregar os dados sem aplicação da derivada a segunda, o que sugere que esse tratamento tem maior inserção de ruídos nos dados do que eliminação dos mesmos nas regiões de importância para o modelo classificatório.

Percebeu-se que as transições vibracionais fundamentais, detectadas pela metodologia FTIR mediante radiação de comprimento de onda no infravermelho médio, são importantes na classificação geográfica, visto que, em comparação com o encontrado na literatura, resultaram em modelos preditivos mais confiáveis que aqueles embasados em dados experimentais provenientes da técnica instrumental NIR. Em contrapartida, o mesmo não se repetiu para os modelos classificatórios de presença de açúcar, para os quais as vibrações sutis identificadas pelo infravermelho próximo possibilitaram a obtenção de modelos de alta qualidade preditiva.

6. REFERÊNCIAS

ALEXANDRE MARCELO, M. C. et al. Methods of multivariate analysis of NIR reflectance spectra for classification of yerba mate. **Analytical Methods**, v. 6, n. 19, p. 7621–7627, 2014.

ANVISA. **Resolução RDC n 303, de 07 de novembro de 2002**Brasil, 2002. Disponível em: <http://portal.anvisa.gov.br/documents/33916/394219/RDC_303_2002.pdf/6acf6086-1086-4bfc-af80-1b74a213a346>

BONA, E. **GAMMA - Ferramentas de Análise Multivariada para Alimentos**. 2019.

BRERETON, R. G. **Chemometrics Data Analysis for the Laboratory and Chemical Plant**. [s.l.] Wiley, 2003.

CALLAO, M. P.; RUISÁNCHEZ, I. An overview of multivariate qualitative methods for food fraud detection. **Food Control**, v. 86, p. 283–293, 2018.

CAMOTTI BASTOS, M. et al. Yerba mate: Nutrient levels and quality of the beverage depending on the harvest season. **Journal of Food Composition and Analysis**, v. 69, p. 1–6, 2018.

CAMPANÁRIO. **Propriedades da Erva-Mate**. Disponível em: <http://campanario.ind.br/blog/propriedades-da-erva-mate/>.

CHECHI, L. A. **Erva-mate: história, tradição e mercado no sul do Brasil**. Disponível em: <http://fidamercosur.org/claeh/experiencias/experiencias-en-la-región/894-erva-mate-história,-tradição-e-mercado-no-sul-do-brasil>.

DIAS, S. L. P. et al. **Química analítica : teoria e prática essenciais**. 1. ed. Porto Alegre: Bookman, 2016.

DRAELOS, R. **Measuring Performance: AUC (AUROC)**. Disponível em: <https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc/>.

EPAGRI. **Dia do Chimarrão: Epagri pesquisa erva-mate há mais de 30 anos**. Disponível em: <https://www.epagri.sc.gov.br/index.php/2019/04/24/dia-do-chimarrao-epagri-pesquisa-erva-mate-ha-mais-de-30-anos/>.

ESTEKI, M.; SHAHSAVARI, Z.; SIMAL-GANDARA, J. Use of spectroscopic methods in combination with linear discriminant analysis for authentication of food products. **Food**

Control, v. 91, p. 100–112, 2018.

EWING, G. W. **Métodos Instrumentais de Análise Química**. São Paulo: Edgard Blucher, 1972.

FERREIRA, M. M. C. **Quimiometria - Conceitos, Métodos e Aplicações**. Campinas: Editora da Unicamp, 2015.

FRIZON, C. N. T. et al. Determination of total phenolic compounds in yerba mate (Ilex paraguariensis) combining near infrared spectroscopy (NIR) and multivariate analysis. **LWT - Food Science and Technology**, v. 60, n. 2, Part 1, p. 795–801, 2015.

GAŁUSZKA, A.; MIGASZEWSKI, Z.; NAMIEŚNIK, J. The 12 principles of green analytical chemistry and the SIGNIFICANCE mnemonic of green analytical practices. **TrAC Trends in Analytical Chemistry**, v. 50, p. 78–84, 2013.

GRANATO, D. et al. Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. **Trends in Food Science & Technology**, v. 72, p. 83–90, 2018.

GRANATO, D.; KARNOPP, A. R.; VAN RUTH, S. M. Characterization and comparison of phenolic composition, antioxidant capacity and instrumental taste profile of juices from different botanical origins. **Journal of the Science of Food and Agriculture**, v. 95, n. 10, p. 1997–2006, 15 ago. 2015.

GROMSKI, P. S. et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. **Analytica Chimica Acta**, v. 879, p. 10–23, 2015.

KAHMANN, A. et al. Near infrared spectroscopy and element concentration analysis for assessing yerba mate (Ilex paraguariensis) samples according to the country of origin. **Computers and Electronics in Agriculture**, v. 140, p. 348–360, 2017.

MABOOD, F. et al. FT-NIRS coupled with chemometric methods as a rapid alternative tool for the detection & quantification of cow milk adulteration in camel milk samples. **Vibrational Spectroscopy**, v. 92, p. 245–250, 2017.

MAPA. **O que é Indicação Geográfica (IG)?** Disponível em:

<http://www.agricultura.gov.br/assuntos/sustentabilidade/indicacao-geografica/o-que-e-indicacao-geografica-ig>.

MARCELO, M. C. A.; POZEBON, D.; FERRÃO, M. F. Authentication of yerba mate according to the country of origin by using Fourier transform infrared (FTIR) associated with chemometrics. **Food Additives & Contaminants: Part A**, v. 32, n. 8, p. 1215–1222, 3 ago. 2015.

MEDINA, S. et al. Food fingerprints – A valuable tool to monitor food authenticity and safety. **Food Chemistry**, v. 278, p. 144–162, 2019.

MENDES, T. DE O. et al. Discrimination between conventional and omega-3 fatty acids enriched eggs by FT-Raman spectroscopy and chemometric tools. **Food Chemistry**, v. 273, p. 144–150, 2019.

OLIVEIRA, G. M. DE. **Simetria de Moléculas e Cristais : Fundamentos da espectroscopia vibracional**. 1. ed. Porto Alegre: Bookman, 2009.

RIACHI, L. G. et al. Effect of light intensity and processing conditions on bioactive compounds in maté extracted from yerba mate (Ilex paraguariensis A. St.-Hil.). **Food Chemistry**, v. 266, p. 317–322, 2018.

ROHMAN, A.; MAN, Y. B. C. Fourier transform infrared (FTIR) spectroscopy for analysis of extra virgin olive oil adulterated with palm oil. **Food Research International**, v. 43, n. 3, p. 886–892, 1 abr. 2010.

SCHNEIDER, M. et al. Exploratory Analysis Applied for the Evaluation of Yerba Mate Adulteration (Ilex paraguariensis). **Food Analytical Methods**, v. 11, n. 7, p. 2035–2041, 2018a.

SCHNEIDER, M. et al. Exploratory Analysis Applied for the Evaluation of Yerba Mate Adulteration (Ilex paraguariensis). **Food Analytical Methods**, 2018b.

SILVA, V. H. DA. **Uso da espectroscopia vibracional na análise de polimorfos.** 2018. 75 f. Tese (Doutorado em Química) - Universidade Federal de Pernambuco, Recife, 2018.

STUART, B. **Infrared Spectroscopy: Fundamentals and Applications**. 1. ed. [s.l.] Wiley, 2004.

TIBOLA, C. S. et al. (EDS.). **Espectroscopia no Infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos**. Brasília: Embrapa, 2018.

VALDERRAMA, L.; VALDERRAMA, P. Nondestructive identification of blue pen inks for documentoscopy purpose using iPhone and digital image analysis including an approach for interval confidence estimation in PLS-DA models validation. **Chemometrics and Intelligent Laboratory Systems**, v. 156, p. 188–195, 2016.

VIEIRA, T. F. et al. Chemometric Approach Using ComDim and PLS-DA for Discrimination and Classification of Commercial Yerba Mate (Ilex paraguariensis St. Hil.). **Food Analytical Methods**, 22 maio 2019.

ZHANG, Q. et al. Authentication of edible vegetable oils adulterated with used frying oil by Fourier Transform Infrared Spectroscopy. **Food Chemistry**, v. 132, n. 3, p. 1607–1613, 1 jun. 2012.