Universidade de São Paulo
Escola de Agricultura “Luiz de Queiroz”

Ressonância magnética nuclear no domínio do tempo, infravermelho
próximo e ciência de dados para análise de citros

Douglas William Menezes Flores

Tese apresentada para obtenção do título de Doutor
em Ciências. Área: Ciência e Tecnologia de
Alimentos

Piracicaba
2020

Douglas William Menezes Flores
Tecnólogo em Alimentos

Ressonância magnética nuclear no domínio do tempo, infravermelho próximo e
ciência de dados para análise de citros

Orientadora:
Prof. Dra. **MARTA HELENA FILLET SPOTO**

Tese apresentada para obtenção do título de Doutor
em Ciências. Área: Ciência e Tecnologia de Alimentos

Piracicaba
2020

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Flores, Douglas William Menezes

Ressonância magnética nuclear no domínio do tempo, infravermelho próximo e ciência de dados para análise de citros / Douglas William Menezes Flores - - Piracicaba, 2020.

107 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura "Luiz de Queiroz".

1. Análise de alimentos via RMN e NIR 2. Ciência de dados e quimiometria 3. Qualidade de citros 4. Python. I. Título

AGRADECIMENTOS

A Deus por mostrar, iluminar e guiar os caminhos.

Aos meus pais, Valdir Machado Flores e Eva Flores por nunca desistirem de me incentivar a estudar, por todo sacrifício, incentivo e confiança em mim.

A minha esposa e colega de doutorado Marilia, por estar ao meu lado e me apoiar em todos os momentos, bons e ruins e pelo incentivo a nunca desistir nestes 10 anos desde a graduação até o final da pós-graduação.

A minha orientadora Marta, pela confiança e oportunidade. Por me ensinar como ser um pesquisador independente, com princípios e humildade.

A todos amigos, colegas, colaboradores, analistas, estagiários e pesquisadores da Embrapa Instrumentação de São Carlos e do GEFH da ESALQ que me ajudaram e ensinaram muitas coisas além da ciência multidisciplinar de ponta.

A empresa Fine Instrument Technology – FIT de São Carlos pela oportunidade do meu primeiro emprego na área de P&D e poder levar junto o doutorado. Sempre levarei comigo todos os desafios superados, conhecimento adquirido e soluções desenvolvidas que hoje facilitam a rotina de muitos laboratórios de análises em empresas de alimentos.

A Daniela e equipe, Diego, Adriano, Luís e Claudio da John Bean Food Tech. Inc, unidade de Araraquara São Paulo, que apoiou todo o desenvolvimento experimental e parceria durante o projeto. Cedendo a planta industrial de citros para diversos testes, análises físico químicas e frutas.

A empresa Astro34 de São Paulo, pela oportunidade do meu segundo emprego em P&D, também apoiando meu doutorado e poder aplicar todo conhecimento adquirido com muito estudo para desenvolver novas soluções com tecnologias de ponta e retornar isto a sociedade.

A todas agências de fomento, FAPESP, CAPES, CNPq, FAPERGS por tanto auxílio estrutural, instrumental e de bolsas para a ciência.

Aos mais de milhares de anônimos do StackOverFlow, YouTube e GitHub por compartilhar de forma gratuita e imensurável diversos conhecimentos em programação Java, C#, Python e Android.

EPÍGRAFE

“Nunca houve uma só metodologia nem uma abordagem única.”

Herman Wold

SUMÁRIO

RESUMO	7
ABSTRACT	8
1. INTRODUÇÃO	9
1.1. REVISÃO DE LITERATURA	13
REFERENCIAS	22
2. RMN-DT APLICADA AO CONTROLE DE PROCESSO DE CITROS: FILTRAÇÃO DE SUCO	31
RESUMO	31
ABSTRACT	32
2.1. INTRODUÇÃO	33
2.2. MATERIAL E MÉTODOS.....	35
2.2.1. AMOSTRAS	35
2.2.2. TESTE DE QUICK FIBER (QF)	35
2.2.3. RESSONÂNCIA MAGNÉTICA NUCLEAR NO DOMÍNIO DO TEMPO (RMN-DT).....	35
2.2.4. ANÁLISE DE DADOS.....	36
2.3. RESULTADOS E DISCUSSÃO	36
2.3.1. CALIBRAÇÃO DE RMN-DT PARA LIMÃO SICILIANO E TAITI	36
2.3.2. CALIBRAÇÃO DE RMN-DT PARA LARANJAS VALÊNCIA, PERA RIO E HAMLIN.....	41
2.3.3. VALIDAÇÃO DA ANÁLISE DE QUICK FIBER POR RMN-DT	44
2.4. CONCLUSÃO	47
REFERÊNCIAS	49
3. RMN-DT APLICADA AO CONTROLE DE PROCESSO DE CITROS: RECUPERAÇÃO DE ÓLEO.	53
RESUMO	53
ABSTRACT	54
3.1. INTRODUÇÃO	55
3.2. MATERIAIS E MÉTODOS.....	56
3.2.1. AMOSTRAS	56
3.2.2. RMN-DT	57
3.2.3. ÓLEO RECUPERÁVEL EM EMULSÃO (%).....	57
3.2.4. ANÁLISE DESCRIMINANTE POR MÍNIMOS QUADRADOS (PLS-DA).....	57
3.3. RESULTADOS	58
3.3.1. PROCESSAMENTO DOS SINAIS DE RMN-DT	58
3.3.2. CONSTRUÇÃO DO MODELO DE PLS-DA	59
3.4. CONCLUSÕES	65
REFERÊNCIAS	67
4. CIÊNCIA DE DADOS: ANÁLISE AUTÔNOMA DE DADOS DE INFRAVERMELHO PRÓXIMO (NIR) PORTÁTIL PARA ANÁLISE DA QUALIDADE DE LARANJAS NO CAMPO	71
RESUMO	71
ABSTRACT	72
4.1. INTRODUÇÃO	73

4.2. METODOLOGIAS	75
4.2.1. ANÁLISES DOS PARÂMETROS DE QUALIDADE DAS LARANJAS	75
4.2.2. ANÁLISE VIA INFRAVERMELHO PRÓXIMO PORTÁTIL.....	76
4.2.3. PYTHON E MÓDULOS.....	77
4.2.4. ESTRUTURA DOS DADOS DE ENTRADA.....	77
4.2.5. FLUXOGRAMA DA ANÁLISE AUTOMATIZADA EM PYTHON	78
4.2.6. CONJUNTO DE DADOS	79
4.2.7. AVALIAÇÃO DOS RESULTADOS.....	80
4.3. RESULTADOS	80
4.4. CONCLUSÃO.....	86
REFERÊNCIAS	87
ANEXOS.....	93

RESUMO

Ressonância magnética nuclear no domínio do tempo, infravermelho próximo e ciência de dados para análise de citros

A citricultura está entre as culturas mais importantes do mundo. A colheita de laranja da Flórida, maior estado produtor de suco de laranja dos EUA foi estimada em 74 milhões de caixas, a qual corresponde a um crescimento de produção. A exportação de suco de laranja também apresenta uma alta no Brasil de 33% em 2019. Além da produção de suco, a produção de frutas cítricas envolve a produção de óleos essenciais. Desde a colheita dos citros até o processamento, alguns parâmetros de qualidade são controlados e monitorados, como, teor de sólidos solúveis, acidez, índice de maturação, pH, óleo essencial, viscosidade, cor, entre outros. Além dos parâmetros físico-químicos analisados, o monitoramento dos processos como extração, filtração do suco ou a extração e centrifugação para recuperação de óleo essencial da casca são parâmetros fundamentais na indústria cítrica. Neste contexto, aplicações de métodos não destrutivos para monitoramento da qualidade dos alimentos têm sido amplamente estudadas. A técnica de ressonância magnética nuclear no domínio do tempo (RMN-DT) vem sendo estudada como alternativa de análise não destrutiva em frutas. A análise é baseada na medição dos tempos de relaxação longitudinal (T_1) e transversal (T_2) e que estão diretamente relacionadas à mobilidade da água nos tecidos dos frutos. Por outro lado, a espectroscopia de NIR é uma técnica que se baseia na interação das ondas eletromagnéticas em um determinado comprimento de onda com os constituintes da amostra e possui inúmeras aplicações não destrutivas em frutas reportadas na literatura. Entretanto, a complexidade dos espectros de NIR e RMN-DT dificulta a interpretação dos resultados, o que torna abordagens estatísticas ou análises multivariadas ferramentas fundamentais na interpretação dos dados. Com o avanço da tecnologia os sensores produzem cada vez uma informação maior de dados e necessita de ferramentas como o Python que é uma linguagem de alto nível e de código aberto que possui diversos módulos de computação científica, visualização e processamento de dados. Desta forma, a quimiometria ou ciência de dados aliada a RMN-DT neste projeto possibilitou o desenvolvimento de uma aplicação para controle de filtração de sucos cítricos e monitoramento de faixas ótimas de recuperação de óleo essencial de citros. Ainda, a tecnologia NIR associada ao Python, possibilitou a criação de modelos de calibração multivariados automatizados para uso em grande escala para prever a qualidade de laranjas inteiras diretamente no campo.

Palavras-chave: Citros, Análise de alimentos por RMN-DT E NIR, Análise automática de outliers, Análise não destrutiva, Quimiometria e ciência de dados via PCA E PLS.

ABSTRACT

Time domain nuclear magnetic resonance, near infrared and data science for citrus analysis

Citriculture is among the most important cultures in the world. The orange harvest from the largest state producing orange juice in the US was estimated at 74 million boxes, which corresponds to a growth in production. Orange juice exports are also up 33% in Brazil in 2019. In addition to juice production, citrus production involves the production of essential oils. These are present in the peel, which represent approximately 45% of the total volume of the fruit. From citrus harvesting to processing, some quality parameters are controlled and monitored, such as soluble solids content, acidity, ripeness index, pH, essential oil, viscosity and color. In addition to the physicochemical parameters, processes monitoring such as extraction, juice filtration or extraction and centrifugation for recovery essential oil of peel are very important parameters in the citrus industry. In this context, applications of non-destructive methods for food quality monitoring have been widely studied. The TD-NMR technique has been studied as an alternative for non-destructive analysis in fruits, since it observes the nuclear spin transition when the sample is placed under the influence of a static magnetic field. On the other hand, NIR spectroscopy is an optical technique that uses the near infrared region of the electromagnetic spectrum and measures the radiation that is absorbed by the compounds to be analyzed and has numerous non-destructive applications in fruits reported in the literature. However, the complexity of NIR and TD-NMR spectra makes it difficult to interpret the results, which makes statistical approaches or multivariate analysis fundamental tools for retrieve data insights. The advancement of technology sensors producing more data information and need tools like Python which is a high level and open source language that has several modules of scientific computing, visualization and data processing. Thus, chemometrics or data science coupled with DT-NMR in this project made it possible to develop an application for citrus juice filtration control and monitoring of optimum citrus essential oil recovery ranges. In addition, Python-associated NIR technology has enabled the creation of automated multivariate calibration models for large-scale use to predict the quality of whole oranges directly in the field.

Keywords: Citrus, Food analysis by TD-NMR and NIR, Automatic outlier analysis, Non-destructive analysis, Chemometrics and data science using PCA and PLS.

1. INTRODUÇÃO

A citricultura está entre as culturas mais importantes do mundo. Segundo a primeira estimativa divulgada pelo Fundecitrus, a safra de laranja de 2019/20 do cinturão citrícola de São Paulo e Triângulo/Sudoeste Mineiro será de 388,42 milhões de caixas. No cenário mundial, a colheita de laranja da Flórida, maior estado produtor de suco de laranja dos EUA foi estimada em 74 milhões de caixas, a qual corresponde a um aumento de 3% ao da safra de 2018/2019 (Fundecitrus, 2019).

Neste contexto, a exportação de suco de laranja está em alta. No primeiro trimestre da safra 2019/2020 os embarques de suco de laranja brasileiro apresentaram alta de 33% em relação à última safra. Tanto o volume de suco concentrado congelado FCOJ (*Frozen Concentrate Orange Juice*) quanto do suco integral NFC (*Not From Concentrate*) apresentaram aumento nas exportações para todos os destinos nos três primeiros meses da safra 2019/2020 (Fundecitrus, 2019).

Além da produção de suco, a produção de frutas cítricas envolve a produção de óleos essenciais. Entre os compostos presentes na composição dos óleos essenciais cítricos estão, compostos voláteis como, terpenos, aldeídos, alcoóis e ésteres e seu principal componente o D-limoneno (Shaw, 1979). Devido suas propriedades físico-químicas, os óleos essenciais provenientes de frutas cítricas são as fragrâncias mais populares no comércio (Ciriminna, Lomell-Rodriguez, Cara, Lopez-Sanchez, & Pagliaro, 2014).

O suco de laranja, embora seja uma bebida sensível aos tratamentos de conservação, é um produto que conserva grande parte das propriedades sensoriais e nutricionais da fruta fresca (Lanza, 2004). Seu processamento se inicia com etapa de extração, onde são utilizados sistemas ‘*squeezing*’. Nesse tipo de extração, trabalha-se com extratora FMC, constituída de dois copos em forma de dedos que se interpenetram, comprimindo a laranja inteira. O copo inferior é dotado de um tubo através do qual escoo o suco. Esse tubo, construído com tela perfurada, retém parte da polpa e demais partes sólidas, impedindo que estas venham a se incorporar ao suco. Um dispositivo comprime a parte da laranja contida dentro do tubo para completar a extração.

Em geral, a extratora FMC extrai 2.200 litros/hora de suco de laranja, dependendo da variedade e tamanho do fruto. Essas máquinas conseguem recuperar até 75 % do óleo, pois este é expelido quando a fruta é esmagada, e como a fruta é

lavada com jato de água na hora da extração, se cria uma emulsão para separar o óleo a posteriori.

O suco de laranja sai do fruto com 15 a 25% de polpa, que é reduzida na filtração (centrífugas), onde também se reduz o teor de óleo em excesso. O suco é em geral comercializado com 10% (concentrado congelado). O suco é, então, encaminhado para os tanques de espera, antes da pasteurização e evaporação. Após a concentração o suco é levado aos tanques de homogeneização ou "*blenders*" onde são devolvidas as essências (parte do óleo essencial) e feitos os ajustes de "*Ratio*" (relação sólidos solúveis/acidez titulável), sendo, então, encaminhado para os "*Tanks Farms*" onde são mantidos a -14°C (Lanza, 2004).

Algumas bactérias como, *Lactobacillus sp.* e *Leuconostoc sp.* são causadoras da deterioração do suco de laranja, pois produzem compostos como diacetil (2,3-Butanodiona e 2,3-Pentanodiona) que são considerados off flavours no suco de laranja. Ainda, outros responsáveis pela alteração do sabor do suco de laranja são as leveduras *Saccharomyces cerevisiae*, *Rhodotorula sp.* e *Zygosaccharomyces sp.* que conferem sabor fermentado ao produto (Deák, 2003). Por este motivo, o suco deve ser clarificado e centrifugado para remoção de óleo (Lanza, 2004). Em sequência, o suco é submetido ao tratamento térmico para após esta etapa poder percorrer diferentes caminhos, que levam a diferentes produtos. O suco NFC (*Not From Concentrate*) corresponde a um dos produtos. Este tipo de suco é processado e armazenado congelado ou pasteurizado rapidamente e depois armazenado de forma asséptica. Diferentemente, o suco FCOJ (*Frozen concentrate Orange juice*) utiliza em sua produção um evaporador de curta duração termicamente acelerado para produzir uma concentração de até 65/75° Brix em um tempo muito curto (James & James, 2003).

Após o processamento, o controle de alguns parâmetros físico-químicos é realizado, como teor de sólidos solúveis (SST), acidez, cor, pH e viscosidade. Usualmente, podem ser consideradas variações de 9 a 14 ° Brix e de 1,5 a 0,5% de acidez em suco de laranja processado. Em casos de sucos de laranja comerciais contendo mistura de diferentes variedades e diferentes graus de maturação, são encontradas variações de 11 a 12° Brix e de 0,8 a 0,9% de acidez (Teixeira & Ferreira, 2003). Além dos parâmetros físico-químicos analisados, o monitoramento do equipamento de filtração do suco e a porcentagem de óleo recuperável são parâmetros fundamentais na qualidade dos sucos cítricos.

Entretanto, os atuais métodos de análise de *Quick Fiber*, a qual consiste na análise para controlar o equipamento *finisher* ou filtração de suco, assim para quantificar a porcentagem de óleo recuperável são dispendiosos, pois requerem preparação da amostra e diversas etapas de análise. Portanto, um método rápido, com pouca preparação de amostra e não invasivo para análise de *Quick Fiber* e de porcentagem de óleo recuperável em citros pode representar uma alternativa eficiente para avaliação da qualidade durante o processamento.

Neste contexto, aplicações de métodos não destrutivos para monitoramento da qualidade dos alimentos têm sido amplamente estudadas. Métodos baseados em propriedades acústicas, de ultrassom, elétricas, ópticas e ondas de rádio vem sendo explorados para medição não destrutiva da qualidade de produtos agrícolas (Abasi, Minaei, Jamshidi, & Fathi, 2018). Entre os métodos não destrutivos estudados, a Espectroscopia de infravermelho próximo (NIR) e Ressonância magnética nuclear no domínio do tempo (RMN-DT) estão entre os métodos mais frequentes (Abasi et al., 2018; Pereira, Carvalho, Cabeça, & Colnago, 2013; F. Z. Ribeiro et al., 2010; Viegas, Mata, Duarte, & Lima, 2016).

A técnica de RMN-DT observa a transição de spin nuclear quando o analito é colocado sobre a influência de um campo magnético estático. A análise é baseada na medição dos tempos de relaxação longitudinal (T_1) e transversal (T_2) e o coeficiente de autodifusão (D). A difusão pode estar diretamente relacionada à mobilidade da água nos tecidos (F. Z. Ribeiro et al., 2010). Por outro lado, a espectroscopia de NIR é uma técnica a qual utiliza a região do infravermelho próximo do espectro eletromagnético e mede a radiação eletromagnética que é absorvida pelos compostos a serem analisados (Sirisomboon, 2018). A complexidade dos espectros de NIR e RMN-DT dificulta a interpretação dos resultados, o que torna abordagens estatísticas ou análises multivariadas ferramentas aliadas na interpretação dos resultados finais (Farres, Srata, Fethi, & Kadaoui, 2019). Os diferentes métodos quimiométricos têm objetivo de extrair informações úteis de conjuntos de dados grandes e complexos (Medina, Perestrelo, Silva, Pereira, & Câmara, 2019).

Com o avanço da tecnologia e uso cada vez mais de sensores, são produzidas uma maior quantidade de dados e informação. Desta forma, a quimiometria ou ciência de dados vem para complementar as tecnologias/sensores como o NIR e a RMN-DT. A análise multivariada de dados se desenvolveu após os anos 70 e é um dos pilares das aplicações em alimentos utilizando espectrofotômetros NIR e espectrômetros de

RMN-DT. As análises de componentes principais (PCA), regressão por componentes principais (PCR) e regressão por mínimos quadrados parciais (PLS) conseguiram elevar a extração de informações químicas relevantes nos espectros de NIR e RMN-DT a um novo patamar. O mundo é multivariado, onde, todos os dados são analisados ao mesmo tempo e com as tecnologias NIR e RMN-DT não é diferente. Os sinais são processados por computadores cada vez mais rápidos e o volume de resultados ou *'insights'* é gerado quase que de forma instantânea. Como resultado, inúmeras aplicações utilizando PLS em dados de NIR e RMN-DT dominam as revistas de aplicações com espectroscopia (Bizzani, Flores, Colnago, & Ferreira, 2017; Guillemain, Dégardin, & Roggo, 2017; Pereira, Bertelli Pflanzner, et al., 2013; Pereira, Carvalho, et al., 2013; Physics, 2017; Xu, Qi, Sun, Fu, & Ying, 2012).

Entretanto, os softwares e a ciência de dados precisam ser compreendidos por usuários cada vez mais avançados para tratar o grande volume de resultados obtidos. A ferramenta Python é uma linguagem de alto nível e código aberto. Proven diversos módulos para leitura e escrita de arquivos, computação científica, visualização de dados, entre outros. Python está se estabelecendo como uma das linguagens mais populares para computação científica. Graças à sua natureza interativa de alto nível e ao seu ecossistema maduro de bibliotecas científicas, é uma opção atraente para o desenvolvimento algorítmico e análise exploratória de dados (Mendes & Gontijo, 2017; Pedregosa, Weiss, & Brucher, 2011; Varoquaux et al., 2015). *Scikit-learn* é um módulo do Python que aproveita esse ambiente rico para fornecer funções de ponta de muitos algoritmos de aprendizado de máquina conhecidos, mantendo uma interface fácil de usar totalmente integrada à linguagem Python. Desta forma, utilizar sensores como NIR e RMN-DT acoplados a tecnologias como o Python para desenvolver sistemas autônomos para análise robusta da informação e criação de aplicações robustas com menor interferência humana é o próximo passo na evolução da tecnologia NIR e de RMN-DT (Doupe, Faghmous, & Basu, 2019; Himanen et al., 2019; Moraes & Gradvohl, 2013).

Portanto, o objetivo geral dessa tese foi desenvolver inovações em análise de alimentos como aplicações utilizando RMN-DT para controlar o processamento de citros (laranja e limão) e ainda utilizar a tecnologia NIR acoplada à ciência de dados para análise autônoma da qualidade de laranjas de forma não destrutiva com auxílio da ferramenta Python. A tese foi desenvolvida em 4 capítulos:

Capítulo 1. Revisão de literatura

Capítulo 2. RMN-DT aplicada ao controle de processo de citros: filtração de suco

Capítulo 3. RMN-DT aplicada ao controle de processo de citros: recuperação de óleo

Capítulo 4. Ciência de dados: Análise autônoma de dados de infravermelho próximo (NIR) portátil para análise da qualidade de laranjas no campo.

1.1. Revisão de literatura

Ressonância Magnética Nuclear no domínio do tempo (RMN-DT)

A ressonância magnética nuclear (RMN) é uma técnica analítica que se destaca com inúmeras aplicações, como, na determinação estrutural de compostos orgânicos e inorgânicos e quantificação de metabólitos em extratos, tecidos intactos, água, gordura em alimentos entre outros sistemas biológicos (J.van Duynhoven, A.Voda, M.Witek, H.Van, 2010). As aplicações espectroscópicas, a RMN em baixa resolução em baixo campo é amplamente utilizada no controle de qualidade de indústrias de alimentos há mais de 54 anos. Nas últimas duas décadas passou também a ser usada como método não invasivo na avaliação de injúrias mecânicas, atributos sensoriais e estudos fisiológicos em frutas (Nascimento, Carvalho, Júnior, Pereira, & Teixeira, 2016; Pereira, Pflanzner, et al., 2013; F. Z. Ribeiro et al., 2010; P. M. Santos, Pereira-Filho, & Colnago, 2016b).

Juntamente com a descoberta do deslocamento químico, o desenvolvimento da instrumentação de equipamentos, possibilitou a aplicação de novas técnicas experimentais de RMN (R. de O. R. Ribeiro et al., 2014). Este método de análise, baseia-se na medida da absorção da radiação eletromagnética na região da radio frequência (24-1200MHz), com núcleos atômicos que possuem momento magnético não nulo (spin diferente de zero) quando inseridos sobre a influência de um campo magnético externo (Blümich, Perlo, & Casanova, 2008; Deshmukh, Simpson, & Hatcher, 2003). Quando os momentos magnéticos de tais núcleos atômicos não se encontram na presença de um campo magnético, eles não possuem orientação preferencial, sendo suas direções aleatórias. Entretanto, quando são submetidos a um campo externo B_0 estático estes núcleos orientam-se a favor do campo magnético

em um movimento giratório oscilante denominado de precessão angular, representando uma condição de equilíbrio térmico.

Para obtenção do sinal de RMN é necessário perturbar a condição de equilíbrio térmico. Para isso um campo magnético oscilante B_1 é aplicado gerando um torque sobre a magnetização que se deslocará da direção Z (condição de equilíbrio térmico) para o plano $x'y'$. Em seguida, após o campo B_1 ser desligado a condição de equilíbrio térmico é restabelecida com o retorno da magnetização na direção Z em um fenômeno conhecido como relaxação. Nesta etapa, quando a magnetização volta a precessionar a favor de B_0 é feita a aquisição do sinal de RMN. Pois quando o sistema nuclear de spin retorna ao estado de equilíbrio térmico é redistribuído o excesso de energia devido aplicação de B_1 (Amaral, 2005). Os processos de relaxação (retorno do estado excitado ao de equilíbrio térmico) observados em RMN são denominados relaxação longitudinal (T_1) e relaxação transversal (T_2) (Cristina et al., 2017). A relaxação longitudinal (T_1) está relacionada ao retorno da magnetização ao equilíbrio térmico, e a relaxação transversal (T_2) com o desaparecimento do sinal no plano $x'y'$. Ambos os processos de relaxação quando relacionados à análise de alimentos são capazes de determinar a proporção de hidrogênio na fase sólida e na fase líquida de um alimento (Gourvénec, Fernández Pierna, Massart, & Rutledge, 2003).

O mecanismo de relaxação longitudinal, spin-rede ou T_1 , está relacionado ao retorno da magnetização ao estado de equilíbrio térmico, M_0 . Essa relaxação é resultante da existência de momentos magnéticos transitórios produzidos pelos movimentos rotacionais e translacionais de moléculas da vizinhança, oscilando com frequências apropriadas em torno da frequência de *Larmor*. A taxa de relaxação $1/T_1$ é máxima quando a frequência do campo magnético flutuante, provocado pela rotação da molécula (τ_c), equivale à frequência de Larmor ($\tau_c \approx 1/\nu_0$). Fora dessa condição, a relaxação longitudinal tende a ser menos eficiente. O T_1 é responsável pela recuperação exponencial do módulo da magnetização, ao longo do eixo z (M_z).

$$M_z(t) = M_0[1 - \exp(-\frac{t}{T_1})]$$

Assim, para análises quantitativas, o T_1 deve ser conhecido ou estimado para evitar a saturação de um ou mais sinais da amostra. Normalmente, é necessário um tempo de espera (T_r) de pelo menos $5T_1$, dependendo da magnitude do campo

magnético, entre cada varredura do espectro de RMN para que a magnetização M_z retorne aproximadamente 99% de M_0 . A sequência de pulsos mais usada para medir T_1 é a sequência de inversão-recuperação (IR). Essa sequência consiste em um pulso de 180° , para inverter a magnetização, seguido por um tempo t (variável) e um pulso de 90° para medir a intensidade do sinal. Essa sequência também é representada esquematicamente por $(180^\circ-t-90^\circ)$ n. Outro processo que ocorre na RMN é o tempo de relaxação transversal, spin-spin ou T_2 , está relacionado à perda de coerência da magnetização no plano $x'y'$. Essa perda de coerência é atribuída às interações diretas entre os momentos magnéticos individuais dos spins em uma amostra. A relaxação transversal provoca um decaimento exponencial da magnetização transversal.

$$M_y(t) = M_{xy} \cos(\omega_0 t) \exp\left(-\frac{t}{T_2}\right)$$

A relaxação transversal, ao contrário da longitudinal, não apresenta um valor máximo de relaxação quando $t_c \approx 1/\nu_0$. Ela é maior em moléculas com alta mobilidade e decresce com a redução da mobilidade molecular. Isso justifica porque, na maioria das amostras líquidas ou em solução, $T_1 \approx T_2$, enquanto nos sólidos $T_1 > T_2$. O T_2 pode ser utilizado diretamente em correlações univariados com propriedades químicas de interesse como viscosidade, teor de sólidos solúveis e concentração de compostos que atuem diretamente na relaxação dos spins de hidrogênio observados (Bizzani et al., 2017; Pereira, Carvalho, et al., 2013; P. M. Santos et al., 2016b). De forma alternativa, os sinais de RMN-DT podem ser avaliados por quimiometria para extrair com mais detalhes as informações químicas das amostras analisadas (De Andrade, Netto, & Colnago, 2011; Pereira, Bertelli Pflanzner, et al., 2013)(Carvalho, Moraes, Lima, & Teixeira, 2019; Coimbra et al., 2020; Romanel, Cunha, Castro, & Barbosa, 2018).

Espectroscopia de infravermelho próximo (NIR)

A região do espectro eletromagnético é dividida em três faixas, onde a radiação no infravermelho próximo (NIR) corresponde à faixa de comprimento de onda de 13.300 a 4.000cm^{-1} ou $750-2500\text{nm}$, a radiação no infravermelho médio (MIR) de 4.000 a 200cm^{-1} , e a radiação no infravermelho distante, de 200 a 10cm^{-1} (C. Pasquini, 2003). Na faixa da radiação de NIR, a luz infravermelha tem energia suficiente para provocar vibrações atômicas, apenas 1J ou $0,24$ cal é necessário para aumentar

0,24°C 1g de H₂O, assim, a energia necessária é alcançada rapidamente. O que permite que a espectroscopia NIR detecte vibrações fundamentais entre dois átomos diferentes. O espectro se origina na observação da transferência da energia da radiação para os átomos, a qual estes a dissipam na forma de energia mecânica, como oscilações (Amodio, Ceglie, Chaudhry, Piazzolla, & Colelli, 2017).

As formas de aquisição dos espectros de NIR são diversas, como por exemplo, transmitância usadas em amostra transparentes e caminho ótico definido, transfectância usadas em amostras transparentes, porém muito densas e com auxílio de um espelho atrás da amostra para refletir a luz emitida, reflectância difusa para amostras sólidas e reflectivas. Entre elas, a reflectância difusa possibilita maior versatilidade das técnicas, pois minimizam o preparo de amostra, aumentando assim a rapidez das análises (Amodio et al., 2017; Biancolillo, Marini, & D'Archivio, 2020; Malegori et al., 2017a). Recentemente, espectrômetros portáteis de NIR e MIR tornaram-se disponíveis comercialmente (Li, Qian, Shi, Medlicott, & East, 2018; Ma, Babu, & Amamcharla, 2019; VIAVI, 2019; Wiedemair & Huck, 2018). Estes instrumentos são robustos e apresentam desempenho analítico similar aos de bancada, mas com objetivos focados em análises rápidas para aplicações já conhecidas. Além disso, estes equipamentos são apropriados para medidas no próprio local de amostras. A informação espectral que dá suporte ao desenvolvimento de análises qualitativas e quantitativas é obtida da informação entre a interação da radiação eletromagnética entre 750-2500 nm compreendidas no infravermelho próximo. A absorção das bandas de NIR diminui quando a massa dos átomos conectados aumenta em cada modo de vibração anarmônica e diminui em função do aumento da força das ligações químicas. Instrumentos miniaturizados geralmente operam entre 900-1700 nm. Nestas faixas de comprimento de onda é possível obter bandas vibracionais de moléculas C-H, R-OH, N-H, S-H o que permite desenvolver uma ampla gama de aplicações em alimentos como em frutas, grãos e cereais, carnes e bebidas de forma rápida e não destrutiva (Celio Pasquini, 2003)(Celio Pasquini, 2018)(Biancolillo et al., 2020; Chapman et al., 2019; Farres et al., 2019; Richter, Rurik, Gurk, Kohlbacher, & Fischer, 2019; Sirisomboon, 2018; Yuan, Mao, Chen, Li, & Huang, 2020).

Devido a estas características a espectroscopia de infravermelho próximo (NIR) tem sido amplamente explorada em diversas áreas, como por exemplo, na análise de alimentos (Alamar, Caramês, Poppi, & Pallone, 2016; Liu, Sun, & Ouyang,

2010; Nascimento et al., 2016; Sirisomboon, 2018; Viegas et al., 2016). Além do desenvolvimento de equipamentos portáteis que não necessitam de preparo de amostra, estes métodos de análise vem sendo favorecidos pelo auxílio de ferramentas de análise multivariada, pois desta forma é possível o desenvolvimento de modelos qualitativos e quantitativos (Borba & Dos Santos, 2017)(Li et al., 2018; Ma et al., 2019; Wiedemair & Huck, 2018).

Quimiometria e ciência de dados para análise de alimentos

Muitos estatísticos ou cientistas de dados nunca ouviram falar de quimiometria. Por outro lado, a maioria dos quimiometristas são familiares com a estatística, mas poucos conhecem a ciência de dados. A multidisciplinariedade em ciência de dados é grande e varia entre as formações de cientistas da computação, estatísticos, economistas, biólogos, engenheiros, tecnólogos, cientistas de alimentos, assim como, as diferentes aplicações nessas áreas (Frolov, Nascimento, Fenner, & Mirkin, 2019; C. Wang, Liu, Gao, & Fu, 2019; Wijewardhana, Meyer, & Jayawardana, 2020).

Atualmente apenas conhecer todas as teorias estatísticas e matemáticas não é suficiente, é necessário principalmente compreender o contexto ou lógica de negócio em que se está inserido. Neste cenário da multidisciplinariedade a atuação das ferramentas matemáticas está focada em problemas cada vez mais específicos, como é o caso da área de análises de alimentos, que cada vez mais vem empregando técnicas e tecnologias avançadas para desenvolver soluções inovadoras, verdes e rápidas para serem implementadas no dia a dia das indústrias (Biancolillo et al., 2020; Blumich, 2016; Magwaza, Opara, et al., 2014; Medina et al., 2019; Wiedemair & Huck, 2018; Yuan et al., 2020). Os quimiometristas estão acostumados a utilizar softwares estatísticos para resolverem problemas no contexto químico, aproveitando toda base de conhecimento das interações intermoleculares e os sensores como espectrofotômetros e espectrômetros.

O aprendizado de máquina, inteligência artificial, big data e quimiometria se tornaram um assunto quente na mídia e na mídia científica (Cacciola, Conti, & Tomasello, 2020; Doupe et al., 2019; Montáns, Chinesta, Gómez-Bombarelli, & Kutz, 2019; Szymańska, 2018). Dentre todas áreas e nomenclaturas, as ferramentas matemáticas utilizadas pela ciência de dados ou quimiometria as técnicas são muitas

vezes as mesmas. O fluxograma das análises dos dados segue uma lógica comum para todos, a preparação dos dados é a tarefa mais importante e a que consome maior tempo de quem trabalha com informação (Montáns et al., 2019; Szymańska, 2018).

Quando trabalhamos com ciência de dados químicos de espectrômetros ou espectrofotômetros é necessário realizar etapas de pré-processamento dos dados, onde, se não realizada corretamente invalida todas as etapas a seguir. O objetivo de pré-processar os dados consiste em remover fontes irrelevantes de informação como artefatos, ruídos, desníveis de linha de base, efeitos globais e deslocamentos de bandas. Quando trabalhamos com dados de infravermelho próximo (NIR) é necessário corrigir a linha de base utilizando primeira/segunda derivada e reduzir o espalhamento da luz que geralmente ocorre em dados de NIR com alguma normalização por variação normal padrão (SNV) ou correção de dispersão múltipla (MSC) (Celio Pasquini, 2018)(Maniwara et al., 2019; Szostak & Mazurek, 2009)(Hegazi, Taha, & Selim, 2019).

Quando trabalhamos com dados de ressonância magnética nuclear no domínio do tempo, é necessário normalizar os sinais pelo máximo quando desejamos maximizar a informação dos tempos de relaxação transversal e também utilizar algum filtro de média móvel como o de *Savitsky-Golay* que utiliza um polinômio para remover ruído que geralmente está presente em sinais de RMN-DT. Após os dados estarem processados, vem a etapa de exploração dos dados que tem objetivo de minimizar a redundância e reduzir os dados ao nível de informação que realmente possui importância em grandes conjuntos. A redução da dimensionalidade de espectros NIR ou de RMN-DT é importante pois estes sensores possuem dados altamente correlacionados entre si, com muita informação repetida ao longo do vetor. Selecionar os intervalos importantes ao longo do vetor e selecionar as amostras com alta variabilidade é essencial para os modelos matemáticos que serão desenvolvidos (Szymańska, 2018).

A análise de dados ou treinamento de modelos em ciência de dados químicos é tratada com etapas de análises não supervisionadas e supervisionadas, com testes univariados ou multivariados. As análises não supervisionadas são compostas por métodos de testes univariados como estatística descritiva, análise de correlação, análise de componentes principais (PCA), análise hierárquica (HCA) entre outros. Estas análises não têm influência do analista de dados que por sua vez pode

interpretar, sumarizar e visualizar o conjunto de dados de forma rápida, clara e objetiva.

A análise descritiva e PCA são as primeiras análises não supervisionadas utilizadas na ciência de dados químicos para identificar e inspecionar as fontes de variação mais importantes no conjunto de dados e assim compreender como as variáveis se relacionam entre si e observar a formação de grupos, tendências ou padrões específicos de comportamento. A PCA utiliza dentre as opções existentes o algoritmo de decomposição por valores singulares (SVD) que é uma tradicional função algébrica que mostra quantos vetores numa matriz são linearmente independentes, ou seja, quantos vetores não são combinações lineares dos outros vetores dentro da mesma matriz. O SVD é uma decomposição de solução única, onde, a partir de uma matriz X obtém-se o auto vetor direito (U), valor singular (S) e o auto vetor esquerdo (V). Após obter os valores U , S e V , para calcularmos a PCA basicamente precisamos obter os *scores* que é o resultado da multiplicação de U por S e os *loadings* são compostos pelo auto vetor esquerdo (V). Os *scores* podem ser interpretados como sendo a localização de cada amostra no novo espaço (espaço das componentes principais). As amostras que possuírem informações químicas semelhantes, irão possuir *scores* semelhantes, e irão estar próximas no espaço das componentes principais (PC's). Logo, os *scores* fornecem a intensidade de contribuição de cada *loading*, para cada amostra. Os *loadings* nos mostram quais foram as variáveis que continham informações em cada componente principal e que foram responsáveis pela formação ou separação de grupos no espaço das componentes principais. Os valores singulares nos informam qual a porcentagem da variação é independente dentro do conjunto de dados, ou seja, a contribuição das variáveis para as informações nas componentes principais. Desta forma é possível obter a distribuição de milhares de dados no novo espaço das componentes principais, observar tendências e padrões nas amostras, obter um ótimo sumário da relação entre as variáveis e prover a conexão entre as observações e variáveis. Com a PCA ainda é possível filtrar informações não relevantes e reduzir o conjunto de dados inicial. Em análises de alimentos a PCA é empregada para estudar dados complexos de análises químicas e de sensores de instrumentação (Anyango, de Kock, & Taylor, 2011; Fasciotti & Pereira Netto, 2010; Gerhardt, Schwolow, Rohn, & Pérez-cacho, 2019; Magwaza, Landahl, et al., 2014; Mesquita & Monteiro, 2018).

De outro lado estão as análises supervisionadas como regressão multivariada linear (MLR), regressão de componente principal (PCR), regressão por mínimos quadrados parciais (PLS), regressão de máquinas de vetores de suporte (SVR), regressão via redes neurais multicamada Perceptron (MLP) e muitos outros (Chen & Wang, 2019; Ciccoritti, Paliotta, Amoriello, & Carbone, 2019; Tsouvaltzis, Babellahi, Amodio, & Colelli, 2020). O diferencial das análises supervisionadas é que não apenas uma matriz de vetores X é decomposta, mas sim um conjunto de dados de um sensor como por exemplo NIR ou RMN-DT é associado com uma nova matriz Y de dados, geralmente em análise de alimentos, dados de análises químicas e físicas realizadas por análises destrutivas e com reagentes em laboratório (Acidez, pH, teor de sólidos solúveis, proteína, umidade, etc.).

O PLS é de longe o mais utilizado pela comunidade científica para modelar dados de sensores instrumentais e dados químicos em laboratório (Bizzani et al., 2017; Ciccoritti et al., 2019; Farres et al., 2019; Medina et al., 2019; Tsouvaltzis et al., 2020). Isto se deve ao fato de que o PLS lida muito bem com o problema de colinearidade existentes em dados de NIR ou RMN-DT fornecendo resultados melhores quando em comparação com modelos de PCR por exemplo. Outro benefício do PLS é a transparência para os analistas que conseguem entender todo o processo de construção da curva de calibração entre os dados dos sensores com os dados do laboratório. Assim como na PCA, o PLS extraí a informação da matriz X , a diferença é que o algoritmo visa extrair a informação que melhor representa os dados da matriz Y , ou seja, os fatores extraídos do PLS possuem maior correlação com os valores de Y do que as componentes principais extraídas pela PCA.

O PLS precisa ser calibrado e validado para poder realizar previsões em novos dados e este é o principal ponto de interesse. Em análise de alimentos a complexidade da composição química das amostras lidas por sensores como NIR ou RMN-DT é enorme. Ainda, existem variações de safra em frutas ou novas misturas complexas que pode ocorrer com variações durante o processamento do alimento. Para calibrarmos modelos de PLS utilizando sinais de NIR ou RMN-DT é necessário um conhecimento prévio da matriz que se está trabalhando além dos cálculos que geram os modelos. É necessário um planejamento experimental sólido, com início, meio e fim, que consiga retratar bem as condições necessárias para um bom desenvolvimento de uma curva de calibração utilizando PLS e sensores NIR ou RMN-DT (Celio Pasquini, 2018).

Além do planejamento experimental, é necessário separar uma fração dos dados para etapa de calibração (70%) e validação (30%) seguido de uma etapa de validação em outro momento como teste. Geralmente algoritmos randômicos de seleção auxiliam nesta tarefa criando blocos casualizados (a casualização aumenta as chances de que possíveis fatores interferentes desconhecidos fiquem igualmente distribuídos nos grupos experimentais) com amostras que não participaram da calibração para aferirmos o desempenho do modelo de PLS na validação possibilitando assim criar uma equação de uma forma mais robusta. Além da separação dos dados, outra tarefa extremamente delicada é a identificação de amostras válidas para realização da calibração e validação dos modelos de PLS. Da mesma forma que não é possível calibrar uma balança com pesos não certificados por algum órgão de medição e aferição, não é possível calibrar um modelo de PLS com dados inválidos de sensores com defeito, ou ainda, com respostas de um laboratório em que não se sabe o desvio padrão e variabilidade das análises de referências fornecidas (Celio Pasquini, 2018).

A identificação de amostras anômalas ou *outliers* é a etapa mais difícil em uma calibração de modelos de PLS. Quando em ambientes industriais esse problema é ainda maior, pois o número de dados gerados é incrivelmente alto, com diversas fontes diferentes, analistas, sensores etc. Neste cenário, as ferramentas estatísticas entram em ação para que seja possível trabalhar com margem de confiabilidade visando reduzir a influência destas amostras dentro dos conjuntos de dados em etapas de calibração de modelos de PLS.

Portanto, a atual análise de alimentos está inserida em um contexto tecnológico na presença de novos sensores e ferramentas analíticas cada dia mais robustas. Assim, a ciência de dados, quimiometria e análise de alimentos começam a fazer parte do mesmo conjunto de ferramentas que visam o mesmo objetivo, criar soluções para a academia, indústria e contexto social a qual se encaixam.

REFERÊNCIAS

Abasi, S., Minaei, S., Jamshidi, B., & Fathi, D. (2018). Dedicated non-destructive devices for food quality measurement: A review. *Trends in Food Science and Technology*, 78(May), 197–205. <https://doi.org/10.1016/j.tifs.2018.05.009>

Alamar, P. D., Caramês, E. T. S., Poppi, R. J., & Pallone, J. A. L. (2016). Quality evaluation of frozen guava and yellow passion fruit pulps by NIR spectroscopy and chemometrics. *Food Research International*, 85, 209–214. <https://doi.org/10.1016/j.foodres.2016.04.027>

Amodio, M. L., Ceglie, F., Chaudhry, M. M. A., Piazzolla, F., & Colelli, G. (2017). Potential of NIR spectroscopy for predicting internal quality and discriminating among strawberry fruits from different production systems. *Postharvest Biology and Technology*, 125, 112–121. <https://doi.org/10.1016/j.postharvbio.2016.11.013>

Anyango, J. O., de Kock, H. L., & Taylor, J. R. N. (2011). Evaluation of the functional quality of cowpea-fortified traditional African sorghum foods using instrumental and descriptive sensory analysis. *LWT - Food Science and Technology*, 44(10), 2126–2133. <https://doi.org/10.1016/j.lwt.2011.07.010>

Biancolillo, A., Marini, F., & D'Archivio, A. A. (2020). Geographical discrimination of red garlic (*Allium sativum* L.) using fast and non-invasive Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy combined with chemometrics. *Journal of Food Composition and Analysis*. <https://doi.org/10.1016/j.jfca.2019.103351>

Bizzani, M., Flores, D. W. M., Colnago, L. A., & Ferreira, M. D. (2017). Non-invasive spectroscopic methods to estimate orange firmness, peel thickness, and total pectin content. *Microchemical Journal*, 133, 168–174. <https://doi.org/10.1016/j.microc.2017.03.039>

Blumich, B. (2016). Introduction to compact NMR: A review of methods. *TrAC - Trends in Analytical Chemistry*, 83, 2–11. <https://doi.org/10.1016/j.trac.2015.12.012>

Blümich, B., Perlo, J., & Casanova, F. (2008). Mobile single-sided NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 52(4), 197–269. <https://doi.org/10.1016/j.pnmrs.2007.10.002>

Borba, K. R., & Dos Santos, P. M. (2017). Espectroscopia de infravermelho para análise de qualidade em frutas e hortaliças- Instrumentação pós-colheita em

frutas e hortaliças. In M. D. Ferreira (Ed.), *Instrumentação pós-colheita em frutas e hortaliças*.

Cacciola, A., Conti, A., & Tomasello, F. (2020). Big Data Analysis: The Leap into a New Science Methodology. *World Neurosurgery*. <https://doi.org/10.1016/j.wneu.2019.04.268>

Carvalho, L. C. de, Morais, C. de L. M. de, Lima, K. M. G. de, & Teixeira, G. H. de A. (2019). Assessment of macadamia kernel quality defects by means of near infrared spectroscopy (NIRS) and nuclear magnetic resonance (NMR). *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.06.021>

Chapman, J., Elbourne, A., Truong, V. K., Newman, L., Gangadoo, S., Rajapaksha Pathirannahalage, P., ... Cozzolino, D. (2019). Sensomics - From conventional to functional NIR spectroscopy - Shining light over the aroma and taste of foods. *Trends in Food Science and Technology*. <https://doi.org/10.1016/j.tifs.2019.07.013>

Chen, Y. yuan, & Wang, Z. bin. (2019). Feature selection based convolutional neural network pruning and its application in calibration modeling for NIR spectroscopy. *Chemometrics and Intelligent Laboratory Systems*. <https://doi.org/10.1016/j.chemolab.2019.06.004>

Ciccoritti, R., Paliotta, M., Amoriello, T., & Carbone, K. (2019). FT-NIR spectroscopy and multivariate classification strategies for the postharvest quality of green-fleshed kiwifruit varieties. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108622>

Ciriminna, R., Lomell-Rodriguez, M., Cara, P. D., Lopez-Sanchez, J. A., & Pagliaro, M. (2014). Limonene: a versatile chemical of the bioeconomy. *Chem. Commun.*, 50, 15288–15296.

Coimbra, P. T., Bathazar, C. F., Guimarães, J. T., Coutinho, N. M., Pimentel, T. C., Neto, R. P. C., ... Cruz, A. G. (2020). Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.107006>

Cristina, L., Picinin, A., Toaldo, I. M., Hoff, R. B., Souza, F. N., Leite, M. O., ... Pinho, O. (2017). *Milk quality parameters associated with the occurrence of veterinary drug residues in bulk tank milk*. (June), 195–202.

De Andrade, F. D., Netto, A. M., & Colnago, L. A. (2011). Qualitative analysis by online nuclear magnetic resonance using Carr-Purcell-Meiboom-Gill sequence with

low refocusing flip angles. *Talanta*, 84(1), 84–88.
<https://doi.org/10.1016/j.talanta.2010.12.033>

Deák, T. (2003). *Encyclopedia of Food Sciences and Nutrition (Second Edition)*. Budapest.

Deshmukh, A. P., Simpson, A. J., & Hatcher, P. G. (2003). Evidence for cross-linking in tomato cutin using HR-MAS NMR spectroscopy. *Phytochemistry*, 64(6), 1163–1170. [https://doi.org/10.1016/S0031-9422\(03\)00505-3](https://doi.org/10.1016/S0031-9422(03)00505-3)

Doupe, P., Faghmous, J., & Basu, S. (2019). Machine Learning for Health Services Researchers. *Value in Health*. <https://doi.org/10.1016/j.jval.2019.02.012>

Farres, S., Srata, L., Fethi, F., & Kadaoui, A. (2019). Argan oil authentication using visible/near infrared spectroscopy combined to chemometrics tools. *Vibrational Spectroscopy*, 102(September 2018), 79–84.
<https://doi.org/10.1016/j.vibspec.2019.04.003>

Fasciotti, M., & Pereira Netto, A. D. (2010). Optimization and application of methods of triacylglycerol evaluation for characterization of olive oil adulteration by soybean oil with HPLC–APCI–MS–MS. *Talanta*, 81(3), 1116–1125.
<https://doi.org/10.1016/j.talanta.2010.02.006>

Frolov, D., Nascimento, S., Fenner, T., & Mirkin, B. (2019). Parsimonious generalization of fuzzy thematic sets in taxonomies applied to the analysis of tendencies of research in data science. *Information Sciences*.
<https://doi.org/10.1016/j.ins.2019.09.082>

Fundecitrus. (2019). Reestimativa da safra de laranja 2018/19 do cinturão citrícola de São Paulo e triângulo/sudoeste mineiro cenário em fevereiro de 2019. Retrieved December 9, 2019, from <https://www.fundecitrus.com.br/pes/estimativa>

Gerhardt, N., Schwolow, S., Rohn, S., & Pérez-cacho, P. R. (2019). Quality assessment of olive oils based on temperature-ramped HS-GC-IMS and sensory evaluation: Comparison of different processing approaches by. *Food Chemistry*, 278(November 2018), 720–728. <https://doi.org/10.1016/j.foodchem.2018.11.095>

Gourvénec, S., Fernández Pierna, J. A., Massart, D. L., & Rutledge, D. N. (2003). An evaluation of the PoLiSh smoothed regression and the Monte Carlo Cross-Validation for the determination of the complexity of a PLS model. *Chemometrics and Intelligent Laboratory Systems*. [https://doi.org/10.1016/S0169-7439\(03\)00086-8](https://doi.org/10.1016/S0169-7439(03)00086-8)

Guillemain, A., Dégardin, K., & Roggo, Y. (2017). Talanta Performance of NIR handheld spectrometers for the detection of counterfeit tablets. *Talanta*, 165(November 2016), 632–640. <https://doi.org/10.1016/j.talanta.2016.12.063>

Hegazi, A., Taha, A., & Selim, M. M. (2019). An improved copy-move forgery detection based on density-based clustering and guaranteed outlier removal. *Journal of King Saud University - Computer and Information Sciences*. <https://doi.org/10.1016/j.jksuci.2019.07.007>

Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., ... Foster, A. S. (2019). Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*. <https://doi.org/10.1016/j.cpc.2019.106949>

J.van Duynhoven*A.Voda*M.Witek†H.Van As†. (2010). Chapter 3 - Time-Domain NMR Applied to Food Products. *Annual Reports on NMR Spectroscopy*, 69, 145–197.

James, C., & James, S. (2003). HEAT TREATMENT | Chemical and Microbiological Changes. *Encyclopedia of Food Sciences and Nutrition*, (1987), 3039–3044. <https://doi.org/10.1016/b0-12-227055-x/00586-1>

Lanza, C. M. (2004). CITRUS FRUITS | Processed and Derived Products of Oranges. *Encyclopedia of Food Sciences and Nutrition*, 1346–1354. <https://doi.org/10.1016/b0-12-227055-x/00243-1>

Li, M., Qian, Z., Shi, B., Medlicott, J., & East, A. (2018). Evaluating the performance of a consumer scale SCiO™ molecular sensor to predict quality of horticultural products. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2018.07.009>

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible-NIR spectrometric technique with PLSR and PCA-BPNN. *LWT - Food Science and Technology*, 43(4), 602–607. <https://doi.org/10.1016/j.lwt.2009.10.008>

Ma, Y. B., Babu, K. S., & Amamcharla, J. K. (2019). Prediction of total protein and intact casein in cheddar cheese using a low-cost handheld short-wave near-infrared spectrometer. *LWT*. <https://doi.org/10.1016/j.lwt.2019.04.039>

Magwaza, L. S., Landahl, S., Cronje, P. J. R., Nieuwoudt, H. H., Mouazen, A. M., Nicolaï, B. M., ... Opara, U. L. (2014). The use of Vis/NIRS and chemometric analysis to predict fruit defects and postharvest behaviour of “Nules Clementine”

mandarin fruit. *Food Chemistry*, 163, 267–274.
<https://doi.org/10.1016/j.foodchem.2014.04.085>

Magwaza, L. S., Opara, U. L., Cronje, P. J. R., Landahl, S., Nieuwoudt, H. H., Mouazen, A. M., ... Terry, L. a. (2014). Assessment of rind quality of “Nules Clementine” mandarin fruit during postharvest storage: 2. Robust Vis/NIRS PLS models for prediction of physico-chemical attributes. *Scientia Horticulturae*, 165, 421–432. <https://doi.org/10.1016/j.scienta.2013.09.050>

Malegori, C., Nascimento Marques, E. J., de Freitas, S. T., Pimentel, M. F., Pasquini, C., & Casiraghi, E. (2017). Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. *Talanta*, 165(November 2016), 112–116. <https://doi.org/10.1016/j.talanta.2016.12.035>

Maniwaru, P., Nakano, K., Ohashi, S., Boonyakiat, D., Seehanam, P., Theanjumpol, P., & Poonlarp, P. (2019). Evaluation of NIRS as non-destructive test to evaluate quality traits of purple passion fruit. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108712>

Medina, S., Perestrelo, R., Silva, P., Pereira, J. A. M., & Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science and Technology*, 85(December 2018), 163–176. <https://doi.org/10.1016/j.tifs.2019.01.017>

Mendes, F. M., & Gontijo, C. C. (2017). Kpop: A Python package for population genetics analysis. *Forensic Science International: Genetics Supplement Series*. <https://doi.org/10.1016/j.fsigss.2017.09.159>

Mesquita, E., & Monteiro, M. (2018). Simultaneous HPLC determination of flavonoids and phenolic acids profile in Pêra-Rio orange juice. *Food Research International*, 106(December 2017), 54–63. <https://doi.org/10.1016/j.foodres.2017.12.025>

Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. *Comptes Rendus - Mécanique*, 347(11), 845–855. <https://doi.org/10.1016/j.crme.2019.11.009>

Moraes, M. B. De, & Gradvohl, A. L. S. (2013). Viabilidade da Linguagem Python no Desenvolvimento de Aplicações de Alto Desempenho. *IV Escola Regional de Alto Desempenho de São Paulo*, (June 2014), 147–150. Retrieved from https://www.researchgate.net/profile/Andre_Gradvohl/publication/253357879_Viabilid

ade_da_Linguagem_Python_no_Desenvolvimento_de_Aplicacoes_de_Alto_Desempenho/links/0046351f90691297f9000000/Viabilidade-da-Linguagem-Python-no-Desenvolvimento-de-Aplicacoes-d

Nascimento, P. A. M., Carvalho, L. C. de, Júnior, L. C. C., Pereira, F. M. V., & Teixeira, G. H. de A. (2016). Robust PLS models for soluble solids content and firmness determination in low chilling peach using near-infrared spectroscopy (NIR). *Postharvest Biology and Technology*, 111, 345–351. <https://doi.org/10.1016/j.postharvbio.2015.08.006>

Pasquini, C. (2003). Review Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. *J. Braz. Chem. Soc.*, 14, 198–219.

Pasquini, Celio. (2003). Near infrared spectroscopy: Fundamentals, practical aspects and analytical applications. *Journal of the Brazilian Chemical Society*, 14(2), 198–219. <https://doi.org/10.1590/S0103-50532003000200006>

Pasquini, Celio. (2018). Analytica Chimica Acta Near infrared spectroscopy: A mature analytical technique with new perspectives e A review. *Analytica Chimica Acta*, 1026, 8–36. <https://doi.org/10.1016/j.aca.2018.04.004>

Pedregosa, F., Weiss, R., & Brucher, M. (2011). *Scikit-learn: Machine Learning in Python*. 12, 2825–2830.

Pereira, F. M. V., Bertelli Pflanzner, S., Gomig, T., Lugnani Gomes, C., De Felício, P. E., & Alberto Colnago, L. (2013). Fast determination of beef quality parameters with time-domain nuclear magnetic resonance spectroscopy and chemometrics. *Talanta*, 108, 88–91. <https://doi.org/10.1016/j.talanta.2013.02.070>

Pereira, F. M. V., Carvalho, A. de S., Cabeça, L. F., & Colnago, L. A. (2013). Classification of intact fresh plums according to sweetness using time-domain nuclear magnetic resonance and chemometrics. *Microchemical Journal*, 108, 14–17. <https://doi.org/10.1016/j.microc.2012.12.003>

Physics, C. (2017). SCIO POCKET-SIZED MICRO SPECTROMETER. Retrieved from <https://www.consumerphysics.com/business/>

Ribeiro, F. Z., Marconcini, L. V., de Toledo, I. B., de Vasconcellos Azeredo, R. B., Barbosa, L. L., & Colnago, L. A. (2010). Nuclear magnetic resonance water relaxation time changes in bananas during ripening: a new mechanism. *Journal of the Science of Food and Agriculture*, 90(12), 2052–2057. <https://doi.org/10.1002/jsfa.4051>

Ribeiro, R. de O. R., Mársico, E. T., Carneiro, C. da S., Monteiro, M. L. G., Júnior, C. C., & Jesus, E. F. O. de. (2014). Detection of honey adulteration of high

fructose corn syrup by Low Field Nuclear Magnetic Resonance (LF 1H NMR). *Journal of Food Engineering*, 135, 39–43. <https://doi.org/10.1016/j.jfoodeng.2014.03.009>

Richter, B., Rurik, M., Gurk, S., Kohlbacher, O., & Fischer, M. (2019). Food monitoring: Screening of the geographical origin of white asparagus using FT-NIR and machine learning. *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.04.032>

Romanel, S. A., Cunha, D. A., Castro, E. V. R., & Barbosa, L. L. (2018). Time domain nuclear magnetic resonance (TD-NMR): A new methodology to quantify adulteration of gasoline. *Microchemical Journal*. <https://doi.org/10.1016/j.microc.2018.03.041>

Santos, P. M., Pereira-Filho, E. R., & Colnago, L. A. (2016). Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). *Microchemical Journal*, 124, 15–19. <https://doi.org/10.1016/j.microc.2015.07.013>

Shaw, P. E. (1979). Review of qualitative analyses of citrus essential oils. *J. Agr. Food Chem.*, 27, 246–257.

Sirisomboon, P. (2018). NIR Spectroscopy for Quality Evaluation of Fruits and Vegetables. *Materials Today: Proceedings*, 5(10), 22481–22486. <https://doi.org/10.1016/j.matpr.2018.06.619>

Szostak, R., & Mazurek, S. (2009). Simple transformation of spectra to effectively reduce quantification errors in FT-Raman multivariate analysis of complex systems. *Vibrational Spectroscopy*. <https://doi.org/10.1016/j.vibspec.2008.09.001>

Szymańska, E. (2018). Modern data science for analytical chemical data – A comprehensive review. *Analytica Chimica Acta*, 1028, 1–10. <https://doi.org/10.1016/j.aca.2018.05.038>

Teixeira, A. R. N., & Ferreira, R. M. B. (2003). Ripening of Fruit. *Encyclopedia of Food Sciences and Nutrition*, (1982), 5006–5014. <https://doi.org/10.1016/b0-12-227055-x/01027-0>

Tsouvaltzis, P., Babellahi, F., Amodio, M. L., & Colelli, G. (2020). Early detection of eggplant fruit stored at chilling temperature using different non-destructive optical techniques and supervised classification algorithms. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2019.111001>

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), 29–33. <https://doi.org/10.1145/2786984.2786995>

VIAMI. (2019). MicroNIR onsite W. Retrieved December 2, 2019, from <https://www.viavisolutions.com/en-us/osp/products/micronir-onsite-w>

Viegas, T. R., Mata, A. L. M. L., Duarte, M. M. L., & Lima, K. M. G. (2016). Determination of quality attributes in wax jambu fruit using NIRS and PLS. *Food Chemistry*, 190, 1–4. <https://doi.org/10.1016/j.foodchem.2015.05.063>

Wang, C., Liu, Z., Gao, H., & Fu, Y. (2019). VOS: A new outlier detection model using virtual graph. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2019.104907>

Wiedemair, V., & Huck, C. W. (2018). Evaluation of the performance of three hand-held near-infrared spectrometer through investigation of total antioxidant capacity in gluten-free grains. *Talanta*. <https://doi.org/10.1016/j.talanta.2018.06.056>

Wijewardhana, U. A., Meyer, D., & Jayawardana, M. (2020). Statistical models for the persistence of threatened birds using citizen science data: A systematic review. *Global Ecology and Conservation*. <https://doi.org/10.1016/j.gecco.2019.e00821>

Xu, H., Qi, B., Sun, T., Fu, X., & Ying, Y. (2012). Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. *Journal of Food Engineering*, 109(1), 142–147. <https://doi.org/10.1016/j.jfoodeng.2011.09.022>

Yuan, L. M., Mao, F., Chen, X., Li, L., & Huang, G. (2020). Non-invasive measurements of ‘Yunhe’ pears by vis-NIRS technology coupled with deviation fusion modeling approach. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2019.111067>

2. RMN-DT APLICADA AO CONTROLE DE PROCESSO DE CITROS: FILTRAÇÃO DE SUCO

RESUMO

O método atual para controlar o processo de filtração de suco e evitar a perdas de produção nas instalações de processamento de suco é via análise de *Quick Fiber* (QF) ou análise da porcentagem de fibra rápida. Esse método é lento e pode causar uma falta de ajuste do sistema. Portanto, o objetivo deste estudo é utilizar um espectrômetro comercial de espectroscopia de ressonância magnética nuclear no domínio do tempo (RMN-DT) e desenvolver um método para predição rápida de valores de fibras em variedades de polpa cítrica e diferentes configurações de extração processadas em uma planta semi-industrial. Neste estudo, descobrimos que os valores de QF causam uma alta variação nas respostas do tempo de relaxamento transversal (T_2) de RMN-DT. Isto permitiu a calibração não linear entre T_2 e QF. O coeficiente de determinação encontrado explicou 98% e 97% dos dados e a raiz quadrática do erro médio de predição (RMSEP) obtido foi de 5,57 ml e 4 ml para o limão Taiti e limão Siciliano, respectivamente. A separação das curvas conforme a variedade foi indicada para limões devido a diferenças na média de sólidos solúveis e ácidos orgânicos entre as duas variedades, isto foi confirmado pela redução do RMSEC de 11,7 ml para menos de 7 ml após a separação das variedades. Para laranjas não foi encontrada variação significativa entre variedades que justificassem a separação das curvas devido ao teor de sólidos solúveis médio ser muito próximo entre as variedades. As diferentes configurações de extratora não afetaram as curvas de calibração de laranjas ou limões e a mudança das configurações podem ser detectadas pelas variações no QF que possuem alta correlação com os sinais de RMN-DT. Portanto, o método de RMN-DT associado a curvas de calibração logarítmicas entre o tempo de relaxação transversal (T_2) pôde ser aplicado para medir o valor de *Quick Fiber* (ml) na polpa de laranjas, limão Siciliano e Taiti dentro de 30 segundos de forma não destrutiva independente da configuração de extratora utilizada.

Palavras-chave: Controle de processo de citros, Análise de alimentos por RMN-DT, Qualidade de citros.

ABSTRACT

The current method for controlling the juice filtration process and avoiding production losses in juice processing facilities is Quick Fiber (QF) analysis. This method is slow and cause a lack of system adjustment. Therefore, the aim of this study is to use a commercial time domain nuclear magnetic resonance spectroscopy spectrometer (TD-NMR) to develop a method for rapid prediction of fiber values in different citrus pulp varieties and different extraction configurations process in one semi-industrial plant. In this study, we found that the QF values cause a high variation in the responses of the transverse relaxation time (T_2) of TD-NMR. This allowed a non-linear calibration between T_2 and QF. The determination coefficient found explained 98% and 97% of the data and the root mean square error of prediction (RMSEP) obtained was 5.57 ml and 4 ml for Tahiti lime and Sicilian lemon, respectively. The separation of the curves according to the variety was indicated for lemons and limes due to differences in the average of soluble solids and organic acids between the two varieties, this was confirmed by the reduction of the RMSEC from 11.7 ml to less than 7 ml after the separation of the varieties. For oranges, no significant variation was found between varieties that justified the separation of the curves due to the average soluble solids content being very close between the varieties. Also, the different extractor configurations did not affect the calibration curves of oranges or lemons and the change in settings can be detected by variations in the QF that have a high correlation with the TD-NMR signals. Therefore, the TD-NMR method associated with logarithmic calibration curves between the time of transversal relaxation (T_2) could be applied to measure the value of Quick Fiber (ml) in the oranges pulp, Sicilian lemon and Tahiti lime within 30 seconds of non-destructive way regardless of the extractor configuration used.

Keywords: Citrus process control, Food analysis by TD-NMR, Citrus quality.

2.1. INTRODUÇÃO

O citros encontra-se entre as culturas mais importantes do mundo, sendo a laranja a principal fruta produzida no Brasil, com cerca de 284,88 milhões de caixas segundo estimativa da safra de 2018/2019 do cinturão citrícola de São Paulo e Triângulo/Sudoeste Mineiro (Fundecitrus, 2019). Neste contexto, o suco de laranja destaca-se como uma bebida muito popular em todo mundo que contribui significativamente para o consumo diário de frutas (Paravisini & Peterson, 2019). A maior parte das laranjas produzidas é direcionada para produção e exportação de NFC (*Not From Concentrate*) e FCOJ (*Frozen Concentrated Orange Juice*) (Mesquita & Monteiro, 2018).

Além dos parâmetros físico-químicas analisados no suco de laranja como, cor, pH, acidez, °Brix e viscosidade é necessário o monitoramento do equipamento de filtração do suco (Baba, Acar, Önta, Sabri, & Sevdan, 2016; Cruz et al., 2020; Galant, Widmer, Luzio, & Cameron, 2014; Rojas J., Perea A., Torres Sáez R., & Ortiz-López C., 2007). Os finalizadores ou *finisher* do tipo parafuso e pá, por exemplo, dependem do suco a ser extrudado através de uma tela de filtração, que regula o tamanho da polpa que é mantida dentro do fluxo de suco. Qualquer célula que seja muito grande para ser extrusadas através dessa peneira é comprimida por centrífugas regulando assim a quantidade de células retidas ou que passam para o suco. O suco separado do concentrador de células é processado por pasteurização ou evaporação. A polpa que é avaliada a partir do concentrador de células tem vários usos. Em alguns processamentos de suco, essas células são consideradas resíduo ou é adicionado a outros resíduos de frutas e transformado em ração animal (Satari & Karimi, 2018; Yan et al., 2018). Ainda, em outras instalações de processamento de suco, essa celulose é considerada um subproduto valioso e pode ser usado para lavagem de polpa. Quando usada para lavagem de polpa, são lavadas com água para recuperar os sólidos solúveis presentes na polpa (Zema et al., 2018).

Nas instalações de processamento de suco, o método atual para controlar o equipamento *finisher* na etapa de filtração é a análise de *Quick Fiber* (QF). Onde a fibra rápida da polpa (QF) é determinada com base no líquido livre que é removido sem a aplicação de pressão. O líquido medido em mililitros ou gramas é chamado de *Quick Fiber*, e o tempo total de análise são cerca de 10 minutos mais tempo necessário para a preparação (Patent No. US 8,820,226 B2, 2014). Este método é

lento e fornece uma falta de ajuste do sistema por meio do tempo necessário para ser realizado, levando os analistas de indústrias a usar sua própria experiência para pular essa importante análise de controle e realizar ajustes empíricos podendo produzir sucos com teor de polpas indesejáveis ao produto que se está produzindo podendo causar problemas sensoriais no produto final (John, Yaragarla, Muthaiah, Ponnusamy, & Appusamy, 2017; Njikam & Schiewer, 2012; Pasandide, Khodaiyan, Mousavi, & Hosseini, 2017).

Por outro lado, alternativas são buscadas pelo setor, como por exemplo um aparelho mecânico e método de monitorar a concentração da polpa em fluxo foi desenvolvido pela indústria pela correlação não linear da densidade da polpa e pressão aplicada no sistema, entretanto o valor de fibra rápida (QF) não é fornecido apenas a concentração aproximada da polpa cítrica (Patent No. US 8,820,226 B2, 2014). O RMN-DT é uma técnica espectroscópica para observar transições de spin nuclear quando as amostras são colocadas em um campo magnético estático e então os tempos de relaxação são adquiridos, estes são conhecidos como relaxação longitudinal (T_1) e transversal (T_2) e difusão molecular que é um sensor promissor para análise rápida de alimentos e processos de produção de alimentos (F. Z. Ribeiro et al., 2010) (Azeredo et al., 2014). Por consequência, estas características permitem o aumento de estudos de RMN-DT para aplicação do método na medição de parâmetros de qualidade em amostras de alimentos (Flores, Colnago, Ferreira, & Spoto, 2016; Pereira, Carvalho, et al., 2013; P. M. Santos, Pereira-Filho, & Colnago, 2016a). Assim, o RMN-DT possui grande velocidade de análise com baixos custos quando comparado a equipamentos de RMN de alta resolução, o que leva a um aumento na frequência analítica com baixos custos operacionais e manutenção de equipamentos, podendo ser uma alternativa para indústria de citros (Pereira, Carvalho, et al., 2013).

Portanto, o objetivo principal deste estudo é usar o espectrômetro RMN-DT para prever valores de fibra rápida (QF) em variedades de polpa de laranja e limão usando diferentes configurações de máquinas de extração em uma planta de citros industrial para controle de processo de filtração de suco.

2.2. Material e métodos

2.2.1. Amostras

Um total de 79 amostras de polpas cítricas foi gerado usando 20 kg de massa de fruta colhidas em fazendas do interior de São Paulo, Brasil por amostra para alimentar a extratora. Foram utilizadas 50 amostras para calibração e 29 amostras de polpas cítricas geradas em outro período para validação. Na etapa de acabamento, foram usados diferentes regulamentos de extração para obter diferentes valores de *Quick Fiber (QF)*. Das 50 amostras utilizadas para calibração, 20 foram de polpas de limão, 10 *Taiti* e 10 *Sicilianos*. Para laranjas, 30 amostras no total, sendo, 10 *Valências*, 10 *Pêra Rio* e 10 da variedade *Hamlin*. Das 29 amostras de validação, foram, 14 laranjas, 7 limões Sicilianos e 8 limões Taiti (John Bean Technologies Copororation, 2011).

2.2.2. Teste de Quick Fiber (QF)

A metodologia de QF foi realizada por meio da coleta de amostras representativas de polpa frescas do *finisher* e seguindo as etapas do protocolo para análise do QF com resultado expresso em mililitros (ml) (John Bean Techonologies Copororation, 2011).

2.2.3. Ressonância magnética nuclear no domínio do tempo (RMN-DT)

Individualmente cada amostra foi homogeneizada, estabilizada à temperatura ambiente de $24 \pm 2,5$ ° C. Cada amostra foi replicada cinco vezes com aproximadamente 30 ± 5 g com massa independente da mesma amostra.

O espectrômetro RMN-DT utilizado foi o Specfit (Fine Instrument Technology, Brasil), que possui um campo magnético de 0,35T, equivalente a 15 MHz para ^1H . A sequência de pulsos utilizada foi a de Carr-Purcel-Meibom-Gill (CPMG) com ciclo de fases, e os parâmetros: pulso de $90^\circ = 12 \mu\text{s}$, $\tau = 1000 \mu\text{s}$ entre pulsos, tempo de aquisição = $64 \mu\text{s}$, *tempo de reciclo* = 5 s, número de eco = 1500 e 16 varreduras. Após a aquisição do sinal foi realizada a análise do tempo de relaxação transversal (T_2) via ajuste mono exponencial e os resultados expressos em segundos.

2.2.4. Análise de dados

Para compor o conjunto de calibração foram utilizadas 50 amostras de polpa geradas com 20kg de frutas cada, sendo, 30 amostras de polpa de laranjas, 10 de limão Siciliano e 10 de limão Taiti. Para o conjunto de validação, foram utilizadas amostras coletadas em outro período, sendo, 14 polpas de laranjas, 7 limões Siciliano e 8 limões Taiti. As amostras foram medidas cinco vezes, em seguida, um ajuste exponencial de primeira ordem foi aplicado a cada sinal de RMN-DT para obter as cinco variáveis constantes de decaimento, reconhecidas como tempo de relaxação transversal (T_2) em segundos. Os valores médios de T_2 (x) foram correlacionados com valores médios de *Quick Fiber* – QF (y) usando a função logarítmica $y = a * \ln(x) - b$. Ainda, coeficiente de determinação (R^2) e a raiz quadrada do erro médio (RMSE) também foram calculados, tanto para os dados de calibração (RMSEC) como validação (RMSEP) (Equação 1). Por fim, foi realizado o teste-t de Student no conjunto de validação para os dados preditos via RMN-DT *versus* QF obtido no laboratório considerando ($p \leq 0,05$).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (\hat{y}_i - y_i)^2} \quad (1)$$

2.3. Resultados e Discussão

2.3.1. Calibração de RMN-DT para limão Siciliano e Taiti

A Tabela 1 mostra os valores médios de T_2 e de *Quick Fiber* (QF) dos limões Siciliano e Taiti. Para amostras com valores de QF acima de 210 ml, o T_2 é maior que 0,22s e amostras de QF com valores abaixo de 150 ml os valores de T_2 são menores que 0,15s. Esses resultados estão associados à mobilidade da água pelas fibras da polpa cítrica em amostras com altos valores de QF. O tempo de relaxação transversal (T_2) está diretamente associado à mobilidade dos núcleos de hidrogênio da água na amostra. A diferença encontrada nos valores de T_2 para diferentes valores de QF segue a variação inversa a de bananas verdes e maduras, onde, ao aumento da

concentração de açúcares solúveis devido ao processo de amadurecimento reduz os tempos de relaxação transversal T_2 e vice versa, devido à quebra do amido em açúcares de cadeia menores alterando assim a mobilidade dos núcleos de hidrogênio, onde o hidrogênio presente nas paredes mais sólidas da banana verde para mais líquida devido à quebra das células ao amadurecer (banana madura) (Raffo, A., Gianferri, R., Barbieri, R., & Brosio, 2005; F. Z. Ribeiro et al., 2010).

Para amostras com baixos valores de QF, a ação das fibras da polpa cítrica causa uma redução da mobilidade dos núcleos de hidrogênio da água, o que reduz os valores do T_2 exponencialmente. A RMN-DT já foi proposta como um método alternativo para determinação da capacidade de retenção de água, sendo a análise de qualidade de carne umas das aplicações mais relatadas em ciência de alimentos (P. M. Santos et al., 2016a). Por fornecer informações sobre os núcleos de hidrogênio no estado de água ligada, água livre e a troca entre estes os dois estados, a RMN-DT demonstrou sensibilidade suficiente para medir o teor de água em amostras de carne (J.van Duynhoven*A.Voda*M.Witek†H.Van As†, 2010). Desta forma, a correlação dos dados obtidos entre T_2 e QF estão de acordo com o que vem sendo obtido na literatura baseado na relaxação transversal que sofre alterações conforme a amostra de polpa cítrica sofre alterações no teor de água livre e valores de fibra rápida, modificando assim, o ambiente dos núcleos de hidrogênio observados pela sequência de CPMG via RMN-DT (Carvalho et al., 2019; Coimbra et al., 2020; Flores et al., 2016; Pereira, Carvalho, et al., 2013; Romanel et al., 2018).

Tabela 1. Dados experimentais de amostras de limões medidas em RMN-DT e análise de referência de *Quick Fiber (QF)*.

Variedade	T_2 (s)	ΔT_2 (s)	Quick fiber (ml)
Siciliano	0,299	0,001	222
Siciliano	0,089	0,001	90
Siciliano	0,148	0,004	154
Siciliano	0,202	0,002	174
Siciliano	0,133	0,004	130
Siciliano	0,113	0,004	126
Siciliano	0,106	0,007	112
Siciliano	0,119	0,001	140

Siciliano	0,186	0,001	164
Siciliano	0,263	0,004	210
Taiti	0,299	0,011	248
Taiti	0,110	0,002	140
Taiti	0,152	0,001	172
Taiti	0,230	0,006	210
Taiti	0,183	0,006	192
Taiti	0,080	0,002	90
Taiti	0,108	0,001	132
Taiti	0,164	0,002	180
Taiti	0,087	0,001	112
Taiti	0,132	0,004	164

A Figura 1 mostra uma forte correlação entre os valores T_2 e QF, o que pode ser explicado por um modelo de regressão logarítmico.

Figura 1. Regressão logarítmica de T_2 vs *Quick Fiber* para limão Siciliano e Taiti.

O coeficiente de determinação R^2 explicou 92% dos dados. Embora seja um valor alto, a diferença química entre o limão Siciliano e o Taiti em termos de sólidos solúveis e ácidos orgânicos afeta negativamente o ajuste devido à influência que

essas moléculas exercem sobre o T_2 . Portanto, a separação das curvas de calibração por variedade de limão é recomendada com base na avaliação realizada observando os valores de RMSEC. Quando realizamos a calibração conjunta, obtivemos um RMSEC de 11.77ml, após a separação das curvas o RMSEC caiu para 5.06ml para Taiti e 6.80 ml para Siciliano. Portanto a separação das curvas de calibração de QF em polpa de limão por variedade pode aumentar o valor do coeficiente de determinação e reduzir o resíduo de previsão, melhorando a precisão das curvas de calibração devido otimizar para a faixa de teor de sólidos solúveis (TSS) das amostras uma vez que o TSS afeta os tempos de relaxação T_2 obtidos por RMN-DT (Figura 2 e 3) (De Andrade et al., 2011; Deng et al., 2014; Medina et al., 2019).

Figura 2. Correlação entre T_2 e *Quick Fiber* para amostras de limão Taiti.

Figura 3. Correlação entre T₂ e Quick Fiber para amostras de limão Siciliano

Utilizando as variedades de limão individualmente para as calibrações foram alcançados RMSEC abaixo de 7ml e coeficientes de determinação que explicaram 98% e 97% dos dados para o Taiti e Siciliano respectivamente. Portanto, se houver diferença significativa na composição química entre as variedades/frutos, recomenda-se ter uma calibração dedicada/separada que melhor descreva as variações de resposta para a variedade/cultivar observada devido ao T₂ sofrer influência dos sólidos solúveis (TSS)(Ciccoritti et al., 2019; Firmani, La Piscopia, Bucci, Marini, & Biancolillo, 2020; Pell, 2000). Outro fator importante é que a temperatura das amostras não deve variar mais do que $\pm 3^{\circ}\text{C}$ da temperatura das amostras durante a etapa de calibração da equação. Isso garante a funcionalidade das curvas de calibração e as aquisições dos sinais do espectrômetro RMN-DT bem como os valores de T₂ medidos devido a temperatura afetar as medições.

2.3.2. Calibração de RMN-DT para Laranjas Valência, Pera Rio e Hamlin

A Tabela 2 mostra os dados experimentais de T_2 e QF das laranjas Valência, Pera Rio e Hamlin com diferentes configurações de extratora. Utilizando a configuração 62 da máquina, foi alcançado o maior valor de QF e usando a configuração padrão da máquina, o valor QF obtido foi mais baixo. Assim, observou-se que a configuração de máquinas extratoras e variedades de laranjas afetou o T_2 final medido no RMN-DT bem como os valores de QF conforme a amostra, mostrando a correlação entre T_2 e QF durante variação do processo de extração com diferentes configurações da extratora. Desta forma, as variações nas configurações da extratora refletem no QF obtido no laboratório que se correlacionam com os valores de T_2 (Figura 4).

Tabela 2. Dados experimentais de amostras de laranjas medidas em RMN-DT e análise de referência de *Quick Fiber* em diferentes configurações de extratora.

Variedade	Extratora	T_2 (s)	ΔT_2 (s)	Quick Fiber (ml)
Valencia	40	0,22	0,016	218
Valencia	40	0,14	0,004	164
Valencia	40	0,13	0,002	146
Valencia	40	0,10	0,001	118
Valencia	40	0,07	0,001	90
Valencia	62	0,30	0,007	258
Valencia	62	0,14	0,004	162
Valencia	62	0,13	0,001	150
Valencia	62	0,09	0,002	110
Valencia	62	0,07	0,002	90
Pera Rio	40	0,24	0,005	226
Pera Rio	40	0,15	0,002	172
Pera Rio	40	0,12	0,001	150
Pera Rio	40	0,10	0,001	122
Pera Rio	40	0,08	0,002	90
Pera Rio	62	0,32	0,018	254
Pera Rio	62	0,14	0,003	160

Pera Rio	62	0,11	0,001	134
Pera Rio	62	0,09	0,001	112
Pera Rio	62	0,08	0,002	96
Hamlin	Default	0,06	0,001	62
Hamlin	Default	0,09	0,002	116
Hamlin	Default	0,13	0,001	150
Hamlin	Default	0,15	0,003	172
Hamlin	Default	0,22	0,010	220
Hamlin	Plus	0,08	0,001	80
Hamlin	Plus	0,09	0,000	120
Hamlin	Plus	0,11	0,002	142
Hamlin	Plus	0,13	0,002	162
Hamlin	Plus	0,22	0,004	224

O coeficiente de determinação R^2 de 0,99 mostra que 99% dos dados foram explicados pelo modelo logarítmico, mantendo a correlação entre os valores de QF e T_2 extremamente forte, indicando que todas as variedades de laranjas utilizadas neste experimento podem ser previstas pelo mesmo modelo de calibração. Este resultado mostra também que o equipamento de RMN-DT com a curva desenvolvida pode monitorar os valores de QF até mesmo em mudanças no processo de extração, como a configuração da extratora utilizada e pode ser uma ferramenta mais ágil na escolha do modelo de configuração a ser utilizado para obter o suco com as propriedades de qualidade que mais atendam a produção da linha em determinado momento (Belton, Colquhoun, & Hills, 1993; Reguera-Alvarado, Blanco-Oliver, & Martín-Ruiz, 2016; Zude, Pflanz, Kaprielian, & Aivazian, 2008).

Figura 4. Curva de calibração de laranjas das variedades Valencia, Pera Rio e Hamlin.

Além disso, a redução do coeficiente de determinação encontrada em múltiplas variedades de limão não foi observada com múltiplas variedades de laranjas. Este comportamento é explicado baseado em testes preliminares onde o grande número de frutos usados para produzir polpas de laranja tem um conteúdo de sólidos solúveis (°Brix) e ácidos orgânicos medianos muito próximos, o que não afetou os valores de T_2 de forma significativa como nos limões (Ladaniya, 2008; Raddatz-Mota et al., 2019; Torres, Sánchez, de la Haba, & Pérez-Marín, 2019). Esses dados indicam que é possível usar uma curva de calibração para prever o valor de *Quick Fiber* em múltiplas variedades de laranjas e configurações de máquina simultaneamente, sem a necessidade de curvas específicas como para os limões, com um RMSEC de 5.98 ml em um range de variação de 50 a 300 ml, sendo, suficiente para monitorar o processo de forma ágil.

2.3.3. Validação da análise de Quick Fiber por RMN-DT

As curvas de calibração de RMN-DT para laranja, limões *Siciliano* e *Taiti* foram aplicadas para prever o valor de polpa por *Quick Fiber* de diferentes amostras originadas pelos extratores da indústria obtidos em outro período para validação, os resultados são mostrados abaixo (Tabela 3 e Figura 5). O RMSEP geral de predição encontrado foi de 4.87 ml, considerando as curvas de calibração das laranjas, limões Siciliano e Taiti juntos. Quando avaliadas em separado, obtivemos para laranjas um RMSEP de 4.69 ml, para limão Siciliano RMSEP de 4 ml e para limão Taiti um RMSEP de 5.57 ml. Todas as curvas de calibração quando avaliadas no conjunto externo de validação mostraram estar próximo ou abaixo do RMSEC obtidos, confirmando assim a validade das calibrações de QF via RMN-DT.

Tabela 3. Validação das curvas de calibração para predição via RMN-DT dos valores de Quick Fiber (QF) medidos pelo laboratório para polpa de laranjas, limão Siciliano e Taiti.

Amostra	QF Medido (ml)	QF Predito (ml)	Δ QF (ml)	Calibração
1	168	168	0	Laranjas
2	173	164	9	Laranjas
3	178	179	1	Laranjas
4	180	186	6	Laranjas
5	184	189	5	Laranjas
6	178	177	1	Laranjas
7	174	174	0	Laranjas
8	165	163	2	Laranjas
9	178	183	5	Laranjas
10	165	156	9	Laranjas
11	168	169	1	Laranjas
12	162	163	1	Laranjas
13	167	164	3	Laranjas
14	164	156	8	Laranjas
15	198	192	6	Siciliano
16	202	201	1	Siciliano
17	200	196	4	Siciliano
18	182	175	7	Siciliano

19	186	186	0	Siciliano
20	188	187	1	Siciliano
21	168	165	3	Siciliano
22	162	162	0	Taiti
23	160	164	4	Taiti
24	160	167	7	Taiti
25	169	161	8	Taiti
26	168	166	2	Taiti
27	162	172	10	Taiti
28	168	164	4	Taiti
29	168	168	0	Taiti

Os valores de R^2 permanecem altos na validação (Figuras 5, 6 e 7), exceto para limão Taiti, onde no período de amostragem para validação os valores de QF obtidos pelo laboratório ficaram entre o mínimo de 161 ml até um máximo de 172 ml, ou seja, estão dentro da margem de erro de RMSEC obtidos na calibração de QF. O valor de RMSEP obtido para limão Taiti na validação foi de 5.57 ml. Quando não se tem uma variabilidade de range da análise de QF de mínimo e máxima alta suficiente para ser superior ao RMSEC da curva, os valores de R^2 tendem a não ser elevados. Ainda, se o RMSE das previsões do conjunto de testes/validação é baixo mesmo com R^2 baixo o modelo pode ser útil (Alexander, Tropsha, & Winkler, 2015). Desta forma, o parâmetro mais importante para avaliação da curva de calibração são os valores de RMSEP obtidos, pois é o valor do desvio entre o laboratório e a predição feita pela curva de calibração que melhor retrata a calibração e o que dita se o modelo pode ser utilizado em produção pelos especialistas da área.

Figura 5. Validação da curva de calibração para predição via RMN-DT dos valores de Quick Fiber (QF) medidos pelo laboratório para polpa de laranjas.

Figura 6. Validação da curva de calibração para predição via RMN-DT dos valores de Quick Fiber (QF) medidos pelo laboratório para polpa de limão Siciliano.

Figura 7. Validação da curva de calibração para predição via RMN-DT dos valores de Quick Fiber (QF) medidos pelo laboratório para polpa de limão Taiti.

Os resultados da validação confirmaram a alta aplicabilidade deste método usando o espectrômetro RMN-DT para prever *Quick Fiber* em aproximadamente 30 segundos. Com base em um teste-t de Student realizado utilizando todas as amostras de validação preditas pelas curvas de RMN-DT em comparação com os valores de *Quick Fiber* lidos em laboratório, não houve diferença significativa entre as médias a 95% de confiança ($p < 0,05$). Desta forma, este método tem alto potencial de ser usado como controle de processo em indústrias cítricas visando a obtenção de suco com qualidade sensorial desejada pela linha de produção, podendo controlar o processo de filtração e o teor de sólidos que estarão presentes no suco.

2.4. CONCLUSÃO

Neste estudo descobrimos que a variação dos valores de *Quick Fiber* causa uma alta variação no tempo de relaxação transversal (T_2) obtidos pela sequência de pulsos CPMG via RMN-DT. Isto permitiu a calibração logarítmica entre T_2 (s) e Quick Fiber (ml) de polpas cítricas. O coeficiente de determinação encontrado explicou 98% e 97% dos dados e o RMSEP obtido foi de 5.57 ml e 4 ml para o limão Taiti e limão Siciliano, respectivamente. A separação das curvas conforme a variedade foi indicada

para limões devido a diferenças na média de sólidos solúveis e ácidos orgânicos entre as duas variedades, isto foi confirmado pela redução do RMSEC de 11.7 ml para menos de 7 ml após a separação das variedades. Para laranjas não foi encontrada variação significativa entre variedades que justificassem a separação das curvas devido ao teor de sólidos solúveis médio ser muito próximo entre as variedades. O RMSEP obtido na validação da curva de calibração das laranjas foi de 4.69 ml. As diferentes configurações de extratora não afetaram as curvas de calibração de laranjas ou limões e a mudança das configurações podem ser detectadas pelas variações no QF que possuem alta correlação com os sinais de RMN-DT. Portanto, o método de RMN-DT associado a curvas de calibração logarítmicas entre o tempo de relaxação transversal (T_2) pôde ser aplicado para medir o valor de *Quick Fiber* (ml) na polpa de laranjas dentro de 30 segundos de forma não destrutiva independente da configuração de extratora utilizada.

REFERÊNCIAS

- Alexander, D. L. J., Tropsha, A., & Winkler, D. A. (2015). Beware of R²: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of Chemical Information and Modeling*, 55(7), 1316–1322. <https://doi.org/10.1021/acs.jcim.5b00206>
- Azeredo, R. B. V., Lima, A. A., Cerioni, L. M., Pusiolo, D. J., Instrumentação, E., & Carlos, S. (2014). *Why is Inline NMR Rarely Used as Industrial Sensor ? Challenges and Opportunities*. (2), 191–203. <https://doi.org/10.1002/ceat.201300380>
- Baba, E., Acar, Ü., Önta, C., Sabri, O., & Sevdan, Y. (2016). *Evaluation of Citrus limon peels essential oil on growth performance , immune response of Mozambique tilapia Oreochromis mossambicus challenged with Edwardsiella tarda*. 465, 13–18. <https://doi.org/10.1016/j.aquaculture.2016.08.023>
- Belton, P. S., Colquhoun, I. J., & Hills, B. P. (1993). Applications of NMR to Food Science. *Annual Reports on NMR Spectroscopy*, 26(C), 1–53. [https://doi.org/10.1016/S0066-4103\(08\)60057-5](https://doi.org/10.1016/S0066-4103(08)60057-5)
- Carvalho, L. C. de, Moraes, C. de L. M. de, Lima, K. M. G. de, & Teixeira, G. H. de A. (2019). Assessment of macadamia kernel quality defects by means of near infrared spectroscopy (NIRS) and nuclear magnetic resonance (NMR). *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.06.021>
- Ciccoritti, R., Paliotta, M., Amoriello, T., & Carbone, K. (2019). FT-NIR spectroscopy and multivariate classification strategies for the postharvest quality of green-fleshed kiwifruit varieties. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108622>
- Coimbra, P. T., Bathazar, C. F., Guimarães, J. T., Coutinho, N. M., Pimentel, T. C., Neto, R. P. C., ... Cruz, A. G. (2020). Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.107006>
- Cruz, A. G., Mtz-Enríquez, A. I., Díaz-Jiménez, L., Ramos-González, R., Valdés, J. A. A., Flores, M. E. C., ... Ilyina, A. (2020). Production of fatty acid methyl esters and bioactive compounds from citrus wax. *Waste Management*. <https://doi.org/10.1016/j.wasman.2019.10.021>
- De Andrade, F. D., Netto, A. M., & Colnago, L. A. (2011). Qualitative analysis by online nuclear magnetic resonance using Carr-Purcell-Meiboom-Gill sequence with

low refocusing flip angles. *Talanta*, 84(1), 84–88.
<https://doi.org/10.1016/j.talanta.2010.12.033>

Deng, F., Xiao, L., Chen, W., Liu, H., Liao, G., Wang, M., & Xie, Q. (2014). Rapid determination of fluid viscosity using low-field two-dimensional NMR. *Journal of Magnetic Resonance*, 247, 1–8. <https://doi.org/10.1016/j.jmr.2014.08.003>

Firmani, P., La Piscopia, G., Bucci, R., Marini, F., & Biancolillo, A. (2020). Authentication of P.G.I. Gragnano pasta by near infrared (NIR) spectroscopy and chemometrics. *Microchemical Journal*. <https://doi.org/10.1016/j.microc.2019.104339>

Flores, D. W. M., Colnago, L. A., Ferreira, M. D., & Spoto, M. H. F. (2016). Prediction of Orange juice sensorial attributes from intact fruits by TD-NMR. *Microchemical Journal*, 128, 113–117. <https://doi.org/10.1016/j.microc.2016.04.009>

Fundecitrus. (2019). Reestimativa da safra de laranja 2018/19 do cinturão citrícola de São Paulo e triângulo/sudoeste mineiro cenário em fevereiro de 2019. Retrieved December 9, 2019, from <https://www.fundecitrus.com.br/pes/estimativa>

Galant, A. L., Widmer, W. W., Luzio, G. A., & Cameron, R. G. (2014). Characterization of molecular structural changes in pectin during juice cloud destabilization in frozen concentrated orange juice. *Food Hydrocolloids*, 41, 10–18. <https://doi.org/10.1016/j.foodhyd.2014.03.013>

J.van Duynhoven* A.Voda* M.Witek† H.Van As†. (2010). Chapter 3 - Time-Domain NMR Applied to Food Products. *Annual Reports on NMR Spectroscopy*, 69, 145–197.

John Bean Technologies Coporation, I. (2011). *PROCEDURES FOR ANALYSIS OF CITRUS PRODUCTS*. Lakeland: JBT FoodTech.

John, I., Yaragarla, P., Muthaiah, P., Ponnusamy, K., & Appusamy, A. (2017). Statistical optimization of acid catalyzed steam pretreatment of citrus peel waste for bioethanol production. *Resource-Efficient Technologies*. <https://doi.org/10.1016/j.reffit.2017.04.001>

Ladaniya, M. S. (2008). Fruit Biochemistry. *Citrus Fruit: Biology, Technology and Evaluation*, 150, 193–197. <https://doi.org/http://dx.doi.org/10.1016/B978-012374130-1.50008-5>

Medina, S., Perestrelo, R., Silva, P., Pereira, J. A. M., & Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science and Technology*, 85(December 2018), 163–176. <https://doi.org/10.1016/j.tifs.2019.01.017>

Mesquita, E., & Monteiro, M. (2018). Simultaneous HPLC determination of flavonoids and phenolic acids profile in Pêra-Rio orange juice. *Food Research International*, 106(December 2017), 54–63. <https://doi.org/10.1016/j.foodres.2017.12.025>

Milla, J. D., Schrader, G. W., Michael, L. S., & Danner, D. S. (2014). *Patent No. US 8,820,226 B2*.

Njikam, E., & Schiewer, S. (2012). Optimization and kinetic modeling of cadmium desorption from citrus peels: A process for biosorbent regeneration. *Journal of Hazardous Materials*. <https://doi.org/10.1016/j.jhazmat.2012.01.084>

Paravisini, L., & Peterson, D. G. (2019). Mechanisms non-enzymatic browning in orange juice during storage. *Food Chemistry*, 289(January), 320–327. <https://doi.org/10.1016/j.foodchem.2019.03.049>

Pasandide, B., Khodaiyan, F., Mousavi, Z. E., & Hosseini, S. S. (2017). Optimization of aqueous pectin extraction from Citrus medica peel. *Carbohydrate Polymers*. <https://doi.org/10.1016/j.carbpol.2017.08.098>

Pell, R. J. (2000). Multiple outlier detection for multivariate calibration using robust statistical techniques. *Chemometrics and Intelligent Laboratory Systems*. [https://doi.org/10.1016/S0169-7439\(00\)00082-4](https://doi.org/10.1016/S0169-7439(00)00082-4)

Pereira, F. M. V., Carvalho, A. de S., Cabeça, L. F., & Colnago, L. A. (2013). Classification of intact fresh plums according to sweetness using time-domain nuclear magnetic resonance and chemometrics. *Microchemical Journal*, 108, 14–17. <https://doi.org/10.1016/j.microc.2012.12.003>

Raddatz-Mota, D., Franco-Mora, O., Mendoza-Espinoza, J. A., Rodríguez-Verástegui, L. L., Díaz de León-Sánchez, F., & Rivera-Cabrera, F. (2019). Effect of different rootstocks on Persian lime (*Citrus latifolia* T.) postharvest quality. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108716>

Raffo, A., Gianferri, R., Barbieri, R., & Brosio, E. (2005). Ripening of banana fruit monitored by water relaxation and diffusion ¹H-NMR measurements. *Food Chemistry*, 149–158.

Reguera-Alvarado, N., Blanco-Oliver, A., & Martín-Ruiz, D. (2016). Testing the predictive power of PLS through cross-validation in banking. *Journal of Business Research*. <https://doi.org/10.1016/j.jbusres.2016.04.016>

Ribeiro, F. Z., Marconcini, L. V., de Toledo, I. B., de Vasconcellos Azeredo, R. B., Barbosa, L. L., & Colnago, L. A. (2010). Nuclear magnetic resonance water

relaxation time changes in bananas during ripening: a new mechanism. *Journal of the Science of Food and Agriculture*, 90(12), 2052–2057. <https://doi.org/10.1002/jsfa.4051>

Rojas J., Perea A., Torres Sáez R., & Ortiz-López C. (2007). Determination of flavonone compounds in citrus juices by high performance liquid chromatography. *Journal of Biotechnology*, 13, 130–132. <https://doi.org/10.1016/j.jbiotec.2007.07.228>

Romanel, S. A., Cunha, D. A., Castro, E. V. R., & Barbosa, L. L. (2018). Time domain nuclear magnetic resonance (TD-NMR): A new methodology to quantify adulteration of gasoline. *Microchemical Journal*. <https://doi.org/10.1016/j.microc.2018.03.041>

Santos, P. M., Pereira-Filho, E. R., & Colnago, L. a. (2016). Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). *Microchemical Journal*, 124, 15–19. <https://doi.org/10.1016/j.microc.2015.07.013>

Satari, B., & Karimi, K. (2018). Citrus processing wastes: Environmental impacts, recent advances, and future perspectives in total valorization. *Resources, Conservation and Recycling*. <https://doi.org/10.1016/j.resconrec.2017.10.032>

Torres, I., Sánchez, M. T., de la Haba, M. J., & Pérez-Marín, D. (2019). LOCAL regression applied to a citrus multispecies library to assess chemical quality parameters using near infrared spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. <https://doi.org/10.1016/j.saa.2019.03.090>

Yan, L., Ye, X., Linhardt, R. J., Chen, J., Yu, D., Huang, R., ... Chen, S. (2018). Full recovery of value-added compounds from citrus canning processing water. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.jclepro.2017.12.018>

Zema, D. A., Calabrò, P. S., Folino, A., Tamburino, V., Zappia, G., & Zimbone, S. M. (2018). Valorisation of citrus processing waste: A review. *Waste Management*. <https://doi.org/10.1016/j.wasman.2018.09.024>

Zude, M., Pflanz, M., Kaprielian, C., & Aivazian, B. L. (2008). NIRS as a tool for precision horticulture in the citrus industry. *Biosystems Engineering*, 99(3), 455–459. <https://doi.org/10.1016/j.biosystemseng.2007.10.016>

3. RMN-DT APLICADA AO CONTROLE DE PROCESSO DE CITROS: RECUPERAÇÃO DE ÓLEO.

RESUMO

O processamento de suco de laranja concentrado congelado ocasiona a remoção de vários compostos de aroma, portanto, é feita recuperação essencial destes compostos. Entretanto, os atuais métodos de referência utilizados para quantificação da porcentagem de óleo recuperável são dispendiosos, pois envolvem diversas etapas de execução. Neste contexto, tecnologias como a espectroscopia de ressonância magnética nuclear (RMN-DT) vêm sendo estudada para determinar a qualidade de frutas de forma não invasiva. Desta forma, o objetivo do estudo foi utilizar o método não invasivo de RMN-DT associado a modelos matemáticos de regressão e análise discriminante (PLS-DA) para classificar amostras de citros no processo pela porcentagem de óleo recuperável e manter o nível ótimo do rendimento industrial. Com base em uma análise prévia dos dados, foi necessário normalização das amplitudes de todas as amostras, de modo a evitar confusão aos modelos de PLS-DA. Observou-se que a classe de poucas amostras fora confundida pelo modelo, mesmo assim o desempenho do modelo indicou uma alta seletividade na identificação das classes. Avaliando o conjunto de validação externa a porcentagem de acerto do modelo de PLS-DA foi de 84,21%. Desta forma, a técnica de RMN-DT via CPMG associada a modelos de PLS-DA para classificar amostras de emulsão entre as faixas de % de óleo recuperável de citros indica um alto potencial no controle do processo para sua faixa máxima de rendimento.

Palavras-chave: RMN-DT, PLS-DA, Recuperação de óleo essencial, Classificação em processamento de citros.

ABSTRACT

The processing of frozen concentrated orange juice removes various aroma compounds, therefore, essential recovery of these compounds is made. However, the current reference methods used to quantify the percentage of recoverable oil are expensive, as they involve several stages of execution. In this context, technologies such as nuclear magnetic resonance spectroscopy (TD-NMR) have been studied to determine the quality of fruits in a non-invasive way. Thus, the aim of the study was to use the non-invasive method of TD-NMR associated with mathematical models such as partial least's square regression discriminating analysis (PLS-DA) to classify citrus samples in the process by the percentage of recoverable oil and maintain the optimal level of yield industrial. Based on a previous analysis of the data, it was necessary to normalize the amplitudes of all samples, in order to avoid confusion to the PLS-DA models. It was observed that the class of few samples was confused by the model, even so the performance of the model indicated a high selectivity in the identification of the classes. Evaluating in the external validation set, the percentage of accuracy of the PLS-DA model was 84.21%. Thus, the technique of TD-NMR using CPMG pulse sequence associated with PLS-DA models showed high potential to classify emulsion samples among the ranges of recoverable citrus oil in the process.

Keywords: TD-NMR, PLS-DA, Essential oil recovery, Classification in citrus processing.

3.1. Introdução

Os citros encontram-se entre os frutos mais cultivados do mundo. Somente o Brasil no ano de 2019 apresentou uma área de plantação de laranja correspondente a 648.138 hectares, como produção de 16.490.331 toneladas (SIDRA, 2019). Além da laranja, que o é o principal produto, a citricultura abrange a tangerina, a lima ácida e o limão a qual apresentam alto consumo em diversos países. Entre laranja, limão e tangerina no Brasil são 1,44 milhões de estabelecimentos produtores (Santos, 2019).

Entre a produção de frutas cítricas, a casca de laranja, de acordo com a cultivar, pode representar 45% do volume total (Franco-vega, Ramírez-corona, Palou, & Aurelio, 2016). A composição da casca de laranja inclui a epiderme a qual cobre o exocarpo que consiste em células parenquimatosas irregulares que recobrem diversos sacos de óleo (Farhat, Fabiano-tixier, El, & Maingonnat, 2011; Lago, Rodríguez, Arce, & Soto, 2014; Velázquez-nuñez, Avila-sosa, Palou, & López-malo, 2013). Da mesma forma, as cascas de citros como limão, representam 30-40% do peso do fruto e são fonte de óleo essencial, a qual consiste em um líquido hidrofóbico concentrado que contém compostos de aroma voláteis (Baba, Acar, Önta, Sabri, & Sevdan, 2016).

Os óleos essenciais provenientes de frutas cítricas são as fragrâncias mais populares no comércio. Em suas composições, os óleos essenciais compreendem compostos voláteis como terpenos e derivados oxigenados tais como aldeídos (citral), álcoois e ésteres, sendo seu componente principal D-limoneno (Shaw, 1979). Este componente pode constituir até 90-95% em óleo de casca de laranja e 75% em óleo de casca de limão (Kesterson, Hendrickson, & Braddock, 1971) e desempenha um papel importante no campo de sabor e fragrância devido às suas propriedades físico-químicas (Ciriminna, Lomell-Rodriguez, Cara, Lopez-Sanchez, & Pagliaro, 2014).

Neste contexto, o processamento de suco de laranja concentrado congelado possui etapas como concentração, que remove vários compostos de aroma que ocorrem naturalmente no suco, modificando o sabor e aroma do produto. Desta forma, é feita recuperação essencial destes compostos que consiste em uma coluna de fracionamento e uma série de condensadores e lavadores de gás, onde são obtidos dois produtos: a fase oleosa e a fase aquosa. Por fim, as duas misturas (fase oleosa e fase aquosa) são incorporadas de volta ao suco concentrado congelado em diferentes proporções (Haypek et al., 2000).

Entretanto, os atuais métodos de referência utilizados para quantificação da porcentagem de óleo recuperável em diversas etapas do processamento de citros são dispendiosos, pois requerem preparação da amostra, homogeneização, destilação e titulação (Scott. W.C. and M.K.Veldhuis., 1966). Portanto, um método rápido, com pouca preparação de amostra e não invasivo para medir da porcentagem de óleo recuperável em citros pode representar uma alternativa eficiente para avaliação da qualidade durante o processamento.

Métodos não invasivos, como espectroscopias de ressonância magnética nuclear (RMN-DT) tem sido utilizado para determinar a qualidade de frutas intactas (Flores, Colnago, Ferreira, & Spoto, 2016). Este método encontra-se baseado nos tempos de relaxação longitudinal T_1 , tempo de relaxação transversal T_2 e difusão molecular (Meyer, Kern, Zientek, Guthausen, & Maiwald, 2016; Ribeiro et al., 2010). Em comparação aos espectrômetros de RMN de alta resolução, a qual são conhecidos pelo alto custo de análise, os espectrômetros RMN-DT usam um ímã permanente de baixo custo, podendo ser uma alternativa para os atuais métodos de análise de alimentos (Bizzani, Flores, Colnago, & Ferreira, 2017; Castell-Palou, Rossell, Femenia, Bon, & Simal, 2011; Coimbra et al., 2020; De Andrade, Netto, & Colnago, 2011; Deng et al., 2014; Medina, Perestrelo, Silva, Pereira, & Câmara, 2019; Pereira, Bertelli Pflanzner, et al., 2013; Pereira, Carvalho, Cabeça, & Colnago, 2013).

Desta forma, o objetivo do estudo é utilizar o método não invasivo de RMN-DT associado a modelos matemáticos de regressão e análise discriminante (PLS-DA) para classificar amostras de citros no processo pela porcentagem de óleo recuperável e manter o nível ótimo do rendimento industrial.

3.2. Materiais e métodos

3.2.1. Amostras

Foram coletadas 63 amostras de óleo em emulsão do processo de entrada da centrífuga de recuperação de óleo de diferentes indústrias processadoras de suco de laranja e limão diretamente dos processos com objetivo de encontrar variação no teor de óleo recuperável presente.

3.2.2. RMN-DT

Individualmente, cada amostra foi homogeneizada, estabilizada a temperatura ambiente de $24 \pm 2,5$ ° C e retirado 25 ml de cada amostra para serem inseridas no equipamento e realizadas as leituras.

O espectrômetro de RMN-DT utilizado foi o Specfit (Fine Instrument Technology, Brasil), que tem um campo magnético 0,35T, equivalente a 15 MHz para ^1H . A sequência utilizada foi CPMG com ciclo de fase e os parâmetros foram: pulso de $90^\circ = 12 \mu\text{s}$, tau = 1000 μs entre pulsos, tempo de aquisição = 64 μs , tempo de reciclo = 5 s, número de eco = 1500 e 16 varreduras.

3.2.3. Óleo recuperável em emulsão (%)

Cada amostra foi homogeneizada individualmente e retirada uma alíquota de 25ml. Em um balão de destilação foi adicionado 25ml da amostra, 25ml de isopropanol e 25ml de água com antiespumante. A mistura é destilada completamente e o destilado é reservado. A seguir o destilado é titulado com solução 0,0247 N de bromato de potássio até viragem indicada via indicador laranja de metila. O resultado foi expresso em % de óleo (v/v) conforme (Scott. W.C. and M.K.Veldhuis., 1966).

3.2.4. Análise discriminante por mínimos quadrados (PLS-DA)

Foi desenvolvido um modelo de classificação por *Partial Least Squares Discriminant Analysis* (PLS-DA) utilizando os decaimentos de CPMG obtidos via RMN-DT normalizados pela amplitude máxima e centrados na média e as classes foram construídas baseadas na porcentagem de óleo contida nas amostras de entrada da centrífuga.

As classes foram construídas com base em experiência prévia do controle do processo, onde, se a porcentagem de óleo está entre 0,7 e 1% a classe da amostra é igual a 1 (no controle), caso a porcentagem de óleo esteja fora desta faixa a classe da amostra é igual a 0 (fora de controle).

Do total de 63 amostras, a calibração foi realizada com 44 e validação externa com 19 (separação 70-30%)(R. W. Kennard & L. A. Stone, 1969). As operações

matemáticas foram realizadas utilizando o software MATLAB, The MathWorks, Inc., Natick, Massachusetts, United States. A acurácia, seletividade e sensibilidade do modelo de PLS-DA foi calculada com base no conjunto total de amostras separados em calibração e validação.

3.3. Resultados

3.3.1. Processamento dos sinais de RMN-DT

A seguir os sinais das 63 amostras de emulsão obtidos no equipamento de RMN-DT são apresentados (Figura 1). Neles é possível observar uma variação nos tempos de relaxação transversal (T_2) e na amplitude na variável 0, o que indica a sensibilidade do equipamento em detectar a variação do volume inserido e diferentes teores de óleo ou açúcar que estão influenciando as mudanças no perfil do espectro de RMN-DT e que compõem as amostras.

Figura 1. Sinais de RMN-DT das 63 amostras de emulsão.

As variações observadas nos sinais da figura 1 podem ser um fator positivo para auxiliar as calibrações do equipamento de RMN-DT com o método de referência, entretanto, variações externas como mudanças na amplitude e açúcares podem ser fatores negativos e podem causar confusão aos modelos de PLS-DA. Desta forma, com base nesta análise previa dos dados, ficou claro a necessidade de utilizar uma normalização das amplitudes de todas as amostras (Filzmoser, 2004; Gan & Ng, 2017; Luo, Bao, & Peng, 2019). A normalização que geralmente é utilizada é a normalização pelo máximo, onde, cada espectro tem suas variáveis divididas pela variável que carrega o maior valor, assim, todos os espectros estarão normalizados entre 1 e 0, isto remove o efeito de cada espectro iniciar em uma amplitude diferente, que geralmente está correlacionado com o volume da amostra. Esta normalização auxilia a extração de informação pelo modelo de PLS-DA já que às alterações dos tempos de relaxação transversal são enfatizadas. Ainda, o uso da centralização na média dos sinais de RMN-DT após a normalização da amplitude garante que todos os sinais tenham média zero, ou seja, as diferenças dos tempos de relaxação transversal serão ainda mais enfatizadas, visando assim, a melhoria das calibrações dos modelos de PLS-DA.

3.3.2. Construção do modelo de PLS-DA

Diferentemente do PLS, o PLS-DA tem a capacidade de prever as classes e não quantificar a porcentagem de óleo diretamente. Desta forma os dados da porcentagem de óleo foram utilizados para a construção das classes para serem utilizadas nos modelos de PLS-DA (Tabela 1)(Pasandide, Khodaiyan, Mousavi, & Hosseini, 2017; Ren et al., 2020; Tsouvaltzis, Babellahi, Amodio, & Colelli, 2020).

Tabela 1. Matriz de dados de referência.

Amostra	% óleo	Classe estipulada
1	1,35	0
2	1,37	0
3	0,91	1
4	0,86	1
5	1,03	0

6	1,77	0
7	1,56	0
8	1,3	0
9	1,54	0
10	0,55	0
11	0,50	0
12	0,60	0
13	0,52	0
14	0,92	1
15	0,71	1
16	0,80	1
17	1,15	0
18	1,10	0
19	1,11	0
20	0,87	1
21	1,47	0
22	1,36	0
23	3,29	0
24	2,87	0
25	1,83	0
26	1,43	0
27	1,4	0
28	1,68	0
29	1,76	0
30	0,34	0
31	0,94	1
32	0,81	1
33	0,97	1
34	1,09	0
35	0,95	1
36	1,51	0
37	1,64	0
38	1,47	0
39	1,5	0
40	1,43	0
41	1,6	0

42	1,51	0
43	1,7	0
44	0,57	0
45	0,65	0
46	0,61	0
47	0,55	0
48	0,68	0
49	1,35	0
50	1,80	0
51	0,89	1
52	1,32	0
53	0,71	1
54	1,95	0
55	1,63	0
56	1,82	0
57	1,65	0
58	1,29	0
59	1,52	0
60	1,88	0
61	1,79	0
62	2,04	0

As faixas estipuladas para as criações das classes tiveram como base a expertise de profissionais que atuam diretamente na indústria de citros e conhecem os processos de recuperação de óleo em centrífugas. Desta forma, foi estabelecido que uma faixa entre 0,7 e 1% de óleo é a porcentagem ideal para uma amostra de emulsão entrar no equipamento de recuperação de óleo, onde, se sabe que a eficiência da recuperação será máxima. Assim, se as amostras estiverem fora desta faixa serão necessárias regulares etapas anteriores de processos para que a faixa da porcentagem de óleo seja atendida e assim controlar o processo em formato ideal (Raddatz-Mota et al., 2019; Wu, Cronjé, Verboven, & Defraeye, 2019).

Na tabela 2, podem-se avaliar os parâmetros alcançados pela calibração e validação do modelo de PLS-DA utilizando os sinais de RMN-DT para prever as classes com diferentes teores de óleo nas amostras. Os valores de sensibilidade e seletividade, quanto mais próximos a 1 indicam modelos muito assertivos e os erros entre as classes indicam qual é a margem de segurança para discriminar a amostra

entre duas classes (Amodio, Ceglie, Chaudhry, Piazzolla, & Colelli, 2017; Biancolillo, Marini, & D'Archivio, 2020; Medina et al., 2019; Reguera-Alvarado, Blanco-Oliver, & Martín-Ruiz, 2016).

Tabela 2. Avaliação do modelo de PLS-DA utilizando sinais de RMN-DT para prever as classes com diferentes % de óleo em emulsão.

Parâmetro	Calibração	Validação cruzada	Validação externa
		70%	30%
Sensitividade	1	0,87	1
Seletividade	1	0,806	0,6
Erro entre classes	0	0,1597	0,2

A figura 2 mostra os valores preditos das classes via modelo de PLS-DA utilizando os sinais de RMN-DT normalizados e centrados na média. Pode-se verificar que poucas amostras (dentro do círculo) onde houve algumas confusões pelo modelo, mas que mesmo assim indicam uma alta seletividade na identificação das classes. Ainda, como pode-se verificar que próximo de 1 no eixo y estão as amostras com formato de triângulo de classe 1 e próximo a zero no eixo y estão as amostras em formato de círculo de classe zero.

Figura 2. PLS-DA com as classes previstas das amostras via sinais de RMN-DT.

A porcentagem de acerto do modelo de PLS-DA construído, avaliada no conjunto de validação externa (30%) foi de 84,21% (16 amostras corretas em 19) e pode ser visualizada abaixo (Tabela 3).

Tabela 3. Aplicação do modelo de PLS-DA utilizando sinais de RMN-DT para prever as classes com diferentes % de óleo em emulsão previamente estipuladas.

Classe estipulada	Classe predita	Classe predita final	Acertos	Soma	% Acerto
0	0,26	0	1	16	84,21
0	0,32	0	1		
1	0,52	1	1		
0	0,63	1	0		
0	0,64	1	0		
0	0,35	0	1		
1	0,56	1	1		
0	-0,27	0	1		
0	0,20	0	1		
1	0,65	1	1		
0	-0,11	0	1		
1	0,59	1	1		
0	-0,30	0	1		
0	0,19	0	1		
0	0,29	0	1		
0	0,65	1	0		
0	0,07	0	1		
0	0,12	0	1		
0	-0,03	0	1		

*Classe predita final utilizou a lógica $>0,5$ para classe 1 e $<0,5$ para classe 0.

Os resultados obtidos para a validação externa foram satisfatórios e indicam um alto potencial da técnica de RMN-DT via CPMG associada a modelos de PLS-DA para classificar amostras de emulsão entre as faixas de porcentagem de óleo recuperável de citros. Desta forma é possível obter o controle do processo para sua faixa máxima de rendimento esperado na etapa de centrifugação e recuperação de óleos cítricos em apenas 30s sem a necessidade de uso de solventes e preparo de amostras.

Neste contexto, a técnica de RMN-DT apesar de não apresentar resolução suficiente nos sinais para os tipos de acoplamentos de hidrogênio como no método de RMN de alto campo, é capaz de adquirir informações de certas regiões do sinal que podem estar associadas a porções específicas de ácidos graxos e compostos menores (McDowell, Defernez, Kemsley, Elliott, & Koidis, 2019). E ainda, quando acoplado à quimiometria estes sinais de RMN-DT podem ser utilizado para análise qualitativa de alimentos (Jakes et al., 2015).

No estudo da qualidade de óleos, a técnica de RMN-DT também vem sendo explorada para rastrear as alterações na qualidade do óleo de amendoim adulterado com óleo de soja, óleo de colza ou óleo de palma em proporções de 0 a 100% (Zhu, Wang, & Chen, 2017). Além da análise de T_2 também foram analisadas áreas dos picos, onde se observou que à medida que a proporção de adulteração aumentava o T_2 e as áreas dos picos mudavam linearmente. Neste estudo, foi constatado que a amplitude dos picos para os diferentes tipos de óleos puros são proporcionais ao número de prótons de hidrogênio estáveis e sua frequência de movimento (Zhu et al., 2017). Ainda, no que diz respeito à autenticação de óleos comestíveis, a técnica de RMN-DT demonstrou-se capaz de detectar adulteração de óleo de avelã em azeite até um limite de 11,2% m/m (Parker et al., 2014). Neste estudo, a razão entre as áreas dos picos foi suficiente para detectar a adulteração no nível de 13% m/m para o conjunto de óleos indicado, mostrando alto potencial na leitura de óleos.

3.4. Conclusões

O uso de RMN-DT para medir amostras de emulsão de óleo/água do processo de citros via sequência de CPMG com amplitude normalizada pelo máximo e centrada na média apresentou informações suficientes para realizar a calibração de modelos de PLS-DA. Foi possível definir faixas ótimas em porcentagem de óleo necessárias para alcançar o rendimento máximo no processo de centrifugação e recuperação de óleo. O acerto encontrado por este método foi de 84.21% em uma análise de 30s sem a necessidade do preparo da amostra e uso de solventes. Portanto, o uso de RMN-DT associada a quimiometria e modelo de PLS-DA demonstrou alto potencial de aplicabilidade em monitorar faixas que podem ser estipuladas por especialistas do

processo para controlar de forma rápida e não destrutiva parâmetros importantes na rotina industrial.

REFERÊNCIAS

- Amodio, M. L., Ceglie, F., Chaudhry, M. M. A., Piazzolla, F., & Colelli, G. (2017). Potential of NIR spectroscopy for predicting internal quality and discriminating among strawberry fruits from different production systems. *Postharvest Biology and Technology*, 125, 112–121. <https://doi.org/10.1016/j.postharvbio.2016.11.013>
- Baba, E., Acar, Ü., Önta, C., Sabri, O., & Sevdan, Y. (2016). *Evaluation of Citrus limon peels essential oil on growth performance , immune response of Mozambique tilapia Oreochromis mossambicus challenged with Edwardsiella tarda*. 465, 13–18. <https://doi.org/10.1016/j.aquaculture.2016.08.023>
- Biancolillo, A., Marini, F., & D'Archivio, A. A. (2020). Geographical discrimination of red garlic (*Allium sativum* L.) using fast and non-invasive Attenuated Total Reflectance-Fourier Transformed Infrared (ATR-FTIR) spectroscopy combined with chemometrics. *Journal of Food Composition and Analysis*. <https://doi.org/10.1016/j.jfca.2019.103351>
- Bizzani, M., Flores, D. W. M., Colnago, L. A., & Ferreira, M. D. (2017). Non-invasive spectroscopic methods to estimate orange firmness, peel thickness, and total pectin content. *Microchemical Journal*, 133, 168–174. <https://doi.org/10.1016/j.microc.2017.03.039>
- Castell-Palou, A., Rossell??, C., Femenia, A., Bon, J., & Simal, S. (2011). Moisture profiles in cheese drying determined by TD-NMR: Mathematical modeling of mass transfer. *Journal of Food Engineering*, 104(4), 525–531. <https://doi.org/10.1016/j.jfoodeng.2011.01.010>
- Ciriminna, R., Lomell-Rodriguez, M., Cara, P. D., Lopez-Sanchez, J. A., & Pagliaro, M. (2014). Limonene: a versatile chemical of the bioeconomy. *Chem. Commun.*, 50, 15288–15296.
- Coimbra, P. T., Bathazar, C. F., Guimarães, J. T., Coutinho, N. M., Pimentel, T. C., Neto, R. P. C., ... Cruz, A. G. (2020). Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*. <https://doi.org/10.1016/j.foodcont.2019.107006>
- De Andrade, F. D., Netto, A. M., & Colnago, L. A. (2011). Qualitative analysis by online nuclear magnetic resonance using Carr-Purcell-Meiboom-Gill sequence with low refocusing flip angles. *Talanta*, 84(1), 84–88. <https://doi.org/10.1016/j.talanta.2010.12.033>

Deng, F., Xiao, L., Chen, W., Liu, H., Liao, G., Wang, M., & Xie, Q. (2014). Rapid determination of fluid viscosity using low-field two-dimensional NMR. *Journal of Magnetic Resonance*, 247, 1–8. <https://doi.org/10.1016/j.jmr.2014.08.003>

Farhat, A., Fabiano-tixier, A., El, M., & Maingonnat, J. (2011). Microwave steam diffusion for extraction of essential oil from orange peel : Kinetic data , extract ' s global yield and mechanism. *Food Chemistry*, 125(1), 255–261. <https://doi.org/10.1016/j.foodchem.2010.07.110>

Filzmoser, P. (2004). A multivariate outlier detection method. *Seventh International Conference on Computer Data Analysis and Modeling*, 1(1989), 18–22. Retrieved from <http://computerwranglers.com/com531/handouts/mahalanobis.pdf>

Flores, D. W. M., Colnago, L. A., Ferreira, M. D., & Spoto, M. H. F. (2016). Prediction of Orange juice sensorial attributes from intact fruits by TD-NMR. *Microchemical Journal*, 128, 113–117. <https://doi.org/10.1016/j.microc.2016.04.009>

Franco-vega, A., Ramírez-corona, N., Palou, E., & Aurelio, L. (2016). *Estimation of mass transfer coef fi cients of the extraction process of essential oil from orange peel using microwave assisted extraction*. 170, 136–143.

Gan, G., & Ng, M. K. P. (2017). k-means clustering with outlier removal. *Pattern Recognition Letters*. <https://doi.org/10.1016/j.patrec.2017.03.008>

Haypek, E., Silva, L. H. ., Batista, E., Marques, D. S., Meirelles, A. J. A., & Meireles, M. A. A. (2000). RECOVERY OF AROMA COMPOUNDS FROM ORANGE ESSENTIAL OIL. *Brazilian Journal of Chemical Engineering*, 17.

Jakes, W., Gerdova, A., Defernez, M., Watson, A. D., McCallum, C., Limer, E., ... Kemsley, E. K. (2015). Authentication of beef versus horse meat using 60 MHz ¹H NMR spectroscopy. *Food Chemistry*, 175, 1–9. <https://doi.org/10.1016/j.foodchem.2014.11.110>

Kesterson, R., Hendrickson, R., & Braddock, R. J. (1971). *Florida Citrus Oils*. Gainesville.

Lago, S., Rodríguez, H., Arce, A., & Soto, A. (2014). Fluid Phase Equilibria Improved concentration of citrus essential oil by solvent extraction with acetate ionic liquids. *Fluid Phase Equilibria*, 361, 37–44. <https://doi.org/10.1016/j.fluid.2013.10.036>

Luo, L., Bao, S., & Peng, X. (2019). Robust monitoring of industrial processes using process data with outliers and missing values. *Chemometrics and Intelligent Laboratory Systems*. <https://doi.org/10.1016/j.chemolab.2019.103827>

McDowell, D., Defernez, M., Kemsley, E. K., Elliott, C. T., & Koidis, A. (2019). Low vs high field 1h Nmr spectroscopy for the detection of adulteration of cold pressed rapeseed oil with refined oils. *Lwt*, 111(May), 490–499. <https://doi.org/10.1016/j.lwt.2019.05.065>

Medina, S., Perestrelo, R., Silva, P., Pereira, J. A. M., & Câmara, J. S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science and Technology*, 85(December 2018), 163–176. <https://doi.org/10.1016/j.tifs.2019.01.017>

Meyer, K., Kern, S., Zientek, N., Guthausen, G., & Maiwald, M. (2016). *Process control with compact NMR*. <https://doi.org/10.1016/j.trac.2016.03.016>

Parker, T., Limer, E., Watson, A. D., Defernez, M., Williamson, D., & Kemsley, E. K. (2014). 60MHz 1H NMR spectroscopy for the analysis of edible oils. *TrAC - Trends in Analytical Chemistry*, 57, 147–158. <https://doi.org/10.1016/j.trac.2014.02.006>

Pasandide, B., Khodaiyan, F., Mousavi, Z. E., & Hosseini, S. S. (2017). Optimization of aqueous pectin extraction from Citrus medica peel. *Carbohydrate Polymers*. <https://doi.org/10.1016/j.carbpol.2017.08.098>

Pereira, F. M. V., Bertelli Pflanzner, S., Gomig, T., Lugnani Gomes, C., De Felício, P. E., & Alberto Colnago, L. (2013). Fast determination of beef quality parameters with time-domain nuclear magnetic resonance spectroscopy and chemometrics. *Talanta*, 108, 88–91. <https://doi.org/10.1016/j.talanta.2013.02.070>

Pereira, F. M. V., Carvalho, A. de S., Cabeça, L. F., & Colnago, L. A. (2013). Classification of intact fresh plums according to sweetness using time-domain nuclear magnetic resonance and chemometrics. *Microchemical Journal*, 108, 14–17. <https://doi.org/10.1016/j.microc.2012.12.003>

R. W. Kennard & L. A. Stone. (1969). Computer Aided Design of Experiments. *Technometrics*, 11:1, 137–148.

Raddatz-Mota, D., Franco-Mora, O., Mendoza-Espinoza, J. A., Rodríguez-Verástegui, L. L., Díaz de León-Sánchez, F., & Rivera-Cabrera, F. (2019). Effect of different rootstocks on Persian lime (*Citrus latifolia* T.) postharvest quality. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108716>

Reguera-Alvarado, N., Blanco-Oliver, A., & Martín-Ruiz, D. (2016). Testing the predictive power of PLS through cross-validation in banking. *Journal of Business Research*. <https://doi.org/10.1016/j.jbusres.2016.04.016>

- Ren, W., Zhao, S., Lian, Y., Yang, Y., Tian, G., Zhao, C., ... Zheng, J. (2020). Effects of hydrosoluble calcium ions and organic acids on citrus oil emulsions stabilized with citrus pectin. *Food Hydrocolloids*. <https://doi.org/10.1016/j.foodhyd.2019.105413>
- Ribeiro, F. Z., Marconcini, L. V., de Toledo, I. B., de Vasconcellos Azeredo, R. B., Barbosa, L. L., & Colnago, L. A. (2010). Nuclear magnetic resonance water relaxation time changes in bananas during ripening: a new mechanism. *Journal of the Science of Food and Agriculture*, 90(12), 2052–2057. <https://doi.org/10.1002/jsfa.4051>
- Santos, A. (2019). A importância da citricultura nacional.
- Scott, W.C. and M.K.Veldhuis. (1966). Rapid estimation of recoverable oil in citrus juices by bromate titration. *J. AOAC*. 49:628, 633.
- Shaw, P. E. (1979). Review of qualitative analyses of citrus essential oils. *J. Agr. Food Chem.*, 27, 246–257.
- SIDRA, S. I. de R. A. (2019). Levantamento Sistemático da Produção Agrícola - julho 2019.
- Tsouvaltzis, P., Babellahi, F., Amodio, M. L., & Colelli, G. (2020). Early detection of eggplant fruit stored at chilling temperature using different non-destructive optical techniques and supervised classification algorithms. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2019.111001>
- Velázquez-núñez, M. J., Avila-sosa, R., Palou, E., & López-malo, A. (2013). Antifungal activity of orange (*Citrus sinensis* var . Valencia) peel essential oil applied by direct addition or vapor contact. *Food Control*, 31(1), 1–4. <https://doi.org/10.1016/j.foodcont.2012.09.029>
- Wu, W., Cronjé, P., Verboven, P., & Defraeye, T. (2019). Unveiling how ventilated packaging design and cold chain scenarios affect the cooling kinetics and fruit quality for each single citrus fruit in an entire pallet. *Food Packaging and Shelf Life*. <https://doi.org/10.1016/j.fpsl.2019.100369>
- Zhu, W., Wang, X., & Chen, L. (2017). Rapid detection of peanut oil adulteration using low-field nuclear magnetic resonance and chemometrics. *Food Chemistry*, 216, 268–274. <https://doi.org/10.1016/j.foodchem.2016.08.051>

4. CIÊNCIA DE DADOS: ANÁLISE AUTÔNOMA DE DADOS DE INFRAVERMELHO PRÓXIMO (NIR) PORTÁTIL PARA ANÁLISE DA QUALIDADE DE LARANJAS NO CAMPO

RESUMO

A informação presente em espectros de infravermelho próximo (NIR) é complexa, auto correlacionada e geralmente não está prontamente disponível para a aplicação analítica direta de quantificação. Os sensores NIR estão expostos a mais adversidades como variações desconhecidas do processo, da composição variável das amostras, presença de impurezas e temperatura variável. A detecção de amostras anômalas (*outliers*) que contenham algum erro de qualquer natureza, seja do sensor NIR ou do método primário, poder ser a tarefa mais difícil na etapa de calibração multivariada em um ambiente hostil como o industrial. Portanto, o objetivo deste trabalho foi desenvolver um código em Python utilizando diversos módulos estatísticos validados na comunidade científica para a detecção de outliers, separação do conjunto de dados em calibração (70%) e validação (30%) e construção de modelos de regressão por mínimos quadrados parciais (PLS) entre sinais de infravermelho próximo (NIR) e parâmetros de qualidade de laranjas de forma rápida, gratuita e autônoma. Os dados analisados foram compostos por 300 amostras de laranjas da variedade Pêra Rio inteiras com os respectivos espectros (X) do equipamento de infravermelho próximo portátil e mais 6 espectros das amostras 1 a 6 com adição de ruído aleatório para testar a função de detecção de outliers nos espectros. A rotina Python desenvolvida foi comparada ao software comercial The Unscrambler X e demonstrou alto potencial na análise autônoma de dados de NIR portátil para prever de forma não destrutiva parâmetros de qualidade de laranjas (Brix, pH) via calibração de curvas de PLS com exclusão automática de erros nos conjuntos de dados. Entretanto, os limites de confiança de 99% são indicados em relação aos limites de 95% para as distâncias de Mahalanobis devido a excluir um menor número de amostras e remover apenas problemas reais de outliers e não remover amostras que podem participar ativamente os modelos de calibração e tornar as curvas mais robustas.

Palavras-chave: Python, Análise de dados via PCA e PLS, Ciência de dados e distância de Mahalanobis, NIR, Análise não destrutiva de laranjas.

ABSTRACT

The information present in near infrared (NIR) spectra is complex, self-correlated and is generally not available for the direct analytical application of quantification. The NIR sensors are exposed to more adversities such as unknown variations in the process, the variable composition of the samples, the presence of impurities and variable temperature. The detection of anomalous samples (outliers) that contain any error of any kind, be it the NIR sensor or the primary method, may be the most difficult task in the multivariate calibration step in a hostile environment such as the industrial one. Therefore, the aim of this work was to develop a Python code using several validated statistical libraries in the scientific community for detecting outliers, separating the data set in calibration (70%) and validation (30%) and building regression models by partial least squares (PLS) between near infrared (NIR) signals and orange quality parameters quickly and autonomously. The analyzed data consisted of 300 samples of intact Pera Rio oranges with the respective spectra (X) of the portable near infrared equipment and another 6 spectra of samples 1 to 6 with the addition of random noise to test the outlier detection function in the spectra. The Python routine developed was compared to the commercial software The Unscrambler X and demonstrated high potential in the autonomous analysis of portable NIR data to non-destructively predict orange quality parameters (Brix, pH) by calibration of PLS curves with automatic exclusion of errors in the data sets. However, the 99% confidence limits are indicated in relation to the 95% limits for Mahalanobis distances due to excluding a lower number of samples and removing only real problems from outliers and not removing samples that can actively participate in the calibration models to make the curves more robust.

Keywords: Python, Data analysis by PCA and PLS, Data science and Mahalanobis distance, NIR, Non-destructive analysis of oranges.

4.1. Introdução

O desenvolvimento da instrumentação elevou a quantidade de dados gerados pelos sensores que produzem cada vez mais informação. A informação presente em espectros de infravermelho próximo (NIR) é muito complexa, auto correlacionada e geralmente não está prontamente disponível para a aplicação analítica de quantificação de analitos de interesse industrial. Isto se deve ao fato de que a matriz de dados da amostra que analisamos no NIR gera informação química sobreposta ao longo do espectro (O'Brien et al., 2012; Pérez-Marín et al., 2019; Phuphaphud, Saengprachatanarug, Posom, & Maraphum, 2019). Ainda, efeitos de espalhamento de luz devido a diferenças de tamanhos de partículas e outros interferentes elevam a complexidade de se analisar esses sinais e construir modelos multivariados de regressão robustos para quantificarmos os analitos de interesse (Celio Pasquini, 2018). No cenário industrial treinar/calibrar e validar/testar modelos de regressão multivariados utilizando espectros em condições totalmente controlada como é feito em laboratório nem sempre é uma alternativa viável em desenvolvimento de novas aplicações, em processos online ou em esteiras com amostradores (Butturi-Gomes & Petrere, 2020; Dambros, Moraes, Vasconcellos, & Franklin, 2020; Hao et al., 2019). Os sensores NIR estão expostos a mais adversidades como variações desconhecidas do processo, da composição variável das amostras, presença de impurezas e temperatura (Magwaza, Opara, et al., 2014; Ncama, Opara, Tesfay, Fawole, & Magwaza, 2017; Pérez-Marín et al., 2019). Além disso, os valores realizados pelas metodologias primárias devem ser de fontes validadas e confiáveis para que seja possível construir modelos de calibração eficientes. Entretanto, em ambientes industriais com diferentes turnos, amostras com diferenças sazonais, diferentes analistas e em alguns casos diferentes laboratórios a propagação de erro cresce significativamente (Bizzani et al., 2017; Tavakolian et al., 2013; Teye, Amuah, Mcgrath, & Elliott, 2019; Wu & Wang, 2014; Xu et al., 2012). Para conjuntos de dados desta natureza o uso de softwares comerciais mais utilizados como The Unscrambler X (Camo Analytics, 2019) e PLS_Toolbox (Eigenvector Research, 2019) necessita que o usuário seja avançado e que conheça todas as técnicas disponíveis para realizar um tratamento adequado de conjunto de dados tão complexos e extrair apenas os dados válidos para construção dos modelos. Mesmo neste cenário de usuário avançado o tempo demandado para identificar possíveis amostras fora dos intervalos

de confiança é demorado e pode estar sujeito a erro humano e por estas razões muitos grupos de pesquisa estudam alternativas automatizadas para esta tarefa (Luo, Bao, & Peng, 2019; Thennadil, Dewar, Herdsman, Nordon, & Becker, 2018) .

A detecção de amostras anômalas (*outliers*) que contenham algum erro de qualquer natureza, seja do sensor NIR ou do método primário, poder ser a tarefa mais difícil na etapa de calibração multivariada em um ambiente hostil como o industrial. Algumas ferramentas estatísticas e matemáticas avançadas podem auxiliar nesta tarefa onde o responsável por construir as calibrações desconhece ou não tem todas as informações necessárias para compreender 100% do seu conjunto de dados (C. Wang et al., 2019; J. Wang et al., 2020). A detecção de outliers já foi proposta por muitos métodos univariados. Estes métodos são usados para limpar erros baseados em relação à média ± 2 vezes o desvio padrão ou apenas distância Euclidiana. Entretanto, se as colunas no seu conjunto de dados estiverem correlacionadas, o que normalmente ocorre nos conjuntos de NIR, a distância euclidiana entre um ponto e o centro dos pontos pode fornecer informações enganosas sobre o quão perto um ponto realmente está dos demais e levar a falsa identificação de *outliers*. A base para detecção multivariada de outliers é o cálculo da distância de Mahalanobis (Tao, Wang, Wu, & Lin, 2020; Thennadil et al., 2018). O cálculo é baseado em transformar as colunas em variáveis não correlacionadas, escalar as colunas para variância igual a um e finalmente calcular a distância Euclidiana. Desta forma, o cálculo de Mahalanobis resolve o problema de escala e a alta correlação entre os dados, podendo encontrar *outliers* de forma mais assertiva dentre muitas outras aplicações (Filzmoser, 2004; García-Santillán & Pajares, 2018; Jiang, Ching, Yiu, & Qiu, 2018; Ruiz de la Hermosa González-Carrato, 2018; Tao et al., 2020).

Dentro desse contexto, necessitamos de ferramentas com flexibilidade e que possa ser automatizada para grandes volumes de informação. A ferramenta Python é uma linguagem de alto nível e código aberto. Proven diversos módulos para leitura e escrita de arquivos, computação científica, visualização de dados, entre outros (Himanen et al., 2019; Lee, Yoo, Kim, Lee, & Hong, 2019). Python está se estabelecendo como uma das linguagens mais populares para computação científica (Doupe et al., 2019; Mendes & Gontijo, 2017). Graças à sua natureza interativa de alto nível e ao seu ecossistema maduro de bibliotecas científicas é uma opção atraente para o desenvolvimento algorítmico e análise exploratória de dados (Moraes & Gradvohl, 2013). *Scikit-learn* é um modulo do Python que aproveita esse ambiente

rico para fornecer funções de ponta de muitos algoritmos de aprendizado de máquina conhecidos, mantendo uma interface fácil de usar totalmente integrada à linguagem Python. Isso responde à crescente necessidade de análise estatística de dados por não especialistas nas indústrias de software, bem como em áreas fora da ciência da computação, como biologia, física e ciência de alimentos (Pedregosa et al., 2011; Varoquaux et al., 2015).

Portanto, o objetivo deste trabalho foi desenvolver um código em Python utilizando diversos módulos estatísticos validados na comunidade científica para a detecção de outliers, separação do conjunto de dados em calibração (70%) e validação (30%) e construção de modelos de regressão por mínimos quadrados parciais (PLS) entre sinais de infravermelho próximo (NIR) e parâmetros de qualidade de laranjas de forma rápida e autônoma.

4.2. Metodologias

4.2.1. Análises dos parâmetros de qualidade das laranjas

As laranjas da variedade Pera Rio foram cedidas por uma empresa parceira de Araraquara, São Paulo, Brasil. Ao todo, 300 frutas foram selecionadas entre início, meio e fim de safra, higienizadas e após a leitura não invasiva por NIR foram cortadas na região equatorial e o suco foi extraído em extrator de bancada. Após a extração o suco de cada amostra foi reservado para as análises a seguir.

O teor de sólidos solúveis (BRIX) foi lido em refratômetro digital de bancada RX-5000 Atago, com ajuste de temperatura para 25°C. O resultado foi expresso em °Brix.

A análise de acidez total titulável (ATT) foi obtida por titulação de 10ml de suco homogeneizado em 50ml de água destilada com solução de NaOH 0.1N previamente padronizada com biftalato de potássio. Os resultados foram expressos em g de ácido cítrico por 100g de polpa (Instituto Adolfo Lutz, 2008) (Dahdouh et al., 2015).

A leitura do potencial hidrogeniônico (pH) foi realizada com mergulho completo do eletrodo do equipamento PHS-3B pH meter em 20ml de suco sob agitação.

O índice de maturação (RATIO) foi calculado pela divisão do teor de sólidos solúveis pela acidez total titulável.

Todas as análises foram realizadas em triplicata, sendo utilizados apenas os valores médios, totalizando 300 respostas médias para cada parâmetro.

4.2.2. Análise via infravermelho próximo portátil

Cada laranja foi lida três vezes pelo espectrofotômetro de infravermelho próximo (NIR) modelo MicroNIR Onsite W que opera via conexão Bluetooth com smartphone Android 7.0 ou superior e a bateria (Figura 1)(VIAVI, 2019). O aparelho trabalha com comprimento de onda de 900-1700 nm e é equipado com um filtro linear variável (LVF) diretamente conectado a um conjunto linear de 128 pixels do detector InGaAs com duas lâmpadas de tungstênio usadas como fonte de radiação. Um Spectralon foi utilizado como referência de branco (99% reflectância difusa) e o zero/escuro foi realizado com as lâmpadas apagadas. O tempo de integração utilizado foi de 10 ms e 150 *scans*. O software para aquisição de dados foi o SciR (Spectral Solutions, 2019) onde os espectros são exportados já processados com variação normal padrão (SNV), primeira derivada com polinômio de segundo grau e 5 pontos de suavização no formato (.xlsx). Apenas o sinal médio das 3 aquisições por laranja foi utilizado para os tratamentos estatísticos, totalizando 300 espectros médios.

Figura 1. MicroNIR onsite W sem fio e a bateria com conexão Bluetooth com smartphone Android em operação no app SciR.

4.2.3. Python e módulos

O código foi desenvolvido usando o PyCharm Community Edition 2019.2.4 x64 que é um ambiente de desenvolvimento integrado usado em programação de computadores, especificamente para a linguagem Python (JetBrains, 2019).

A versão do Python utilizada foi a 3.6 com os seguintes módulos importados diretamente pela seção de configuração do projeto no PyCharm: Xlrd 1.2.0; Openpyxl 3.0.0; Scikit-learn 0.21.2; Seaborn 0.9.0; matplotlib 3.1.0; Pandas 0.24.2; numpy 1.16.4; scipy 1.3.0; epip 19.1.1; O código fonte deste projeto pode ser clonado gratuitamente sob a licença MIT (gratuita pra usar, Anexo G) diretamente no domínio deste trabalho hospedado no GitHub (<https://github.com/floresdwm/Thesis>) (Anexo B).

4.2.4. Estrutura dos dados de entrada

O formato de dados deve ser em planilha Excel (.xls ou .xlsx). Quando executado o código Python irá permitir que você escolha a sua planilha com os dados a serem analisados. A planilha deve conter três abas. A primeira aba ('Labels') deve ser uma coluna, onde, a primeira célula é o cabeçalho da coluna. Esta primeira coluna é denominada 'Labels' que será os nomes das amostras. Cada linha representa uma

amostra nova. A segunda aba ('X') deve ser dos dados independentes (X), geralmente as variáveis do espectro ou conjunto de dados de alguma instrumentação. A primeira linha será considerada como cabeçalho e deve conter o nome de cada variável da coluna (comprimento de onda, tempo, etc.). Na sequência cada sinal do instrumento deve ser na mesma ordem em que a aba anterior, onde, cada linha é pertencente e alinhada à mesma amostra. A terceira aba ('Y') deve ser os dados dependentes (Y), geralmente as propriedades químicas ou físicas de interesse que desejamos modelar baseado nos dados da instrumentação. Da mesma forma que as demais, a primeira linha será o cabeçalho com os nomes das propriedades/parâmetros de qualidade e as linhas as amostras alinhadas na mesma ordem das abas 'Labels' e 'X' (Apêndice 1).

4.2.5. Fluxograma da análise automatizada em Python

Os dados importados serão validados quanto ao número de linhas corretas para multiplicação de matrizes. Em seguida são computadas as componentes principais dos dados X e Y individualmente e realizada o cálculo de Mahalanobis para os escores das componentes principais 1 e 2. As linhas/amostras que estiverem fora do limite de confiança serão excluídas da modelagem de forma automática para as duas matrizes (X e Y). Na sequência um modelo de PLS é realizado para cada parâmetro Y inserido no conjunto de calibração (70%) e validado no conjunto de validação (30%) escolhendo automaticamente o melhor fator baseado no menor RMSE obtido na validação cruzada da calibração (70%) e ao fim um relatório é gerado na tela com valores de outliers detectados/removidos, R^2 e RMSE (Apêndice 2-5) (Figura 2).

Figura 2. Fluxograma da rotina Python desenvolvida

4.2.6. Conjunto de dados

Os dados utilizados foram compostos por 300 amostras de laranjas da variedade Pera Rio inteiras com os respectivos espectros (X) do equipamento de infravermelho próximo portátil e mais 6 espectros das amostras 1 a 6 com adição de ruído aleatório para testar a função de detecção de outliers nos espectros. Os espectros foram processados previamente no software do equipamento com variação normal padrão (SNV), primeira derivada com polinômio de segundo grau e 5 pontos de suavização e as análises tecnológicas de teor de sólidos solúveis (BRIX), acidez total titulável (ATT), índice de maturação ($RATIO = BRIX/ATT$) e pH compondo os dados primários de referência (Y). Os dados foram coletados no estado de São Paulo, Brasil de diversas localidades e época da safra e climas diferentes. O objetivo deste conjunto de dados é construir uma regressão multivariada via PLS entre os espectros NIR (X) e as análises de qualidade de laranjas inteiras.

4.2.7. Avaliação dos resultados

Os resultados obtidos no código Python desenvolvido foram comparados aos resultados obtidos no software The Unscrambler X (Camo Analytics, 2019) com a finalidade de verificar se existiu queda nos valores de RMSE após a remoção autônoma via Mahalanobis. Um modelo de regressão por mínimos quadrados parciais (PLS) é avaliado quanto ao coeficiente de determinação (R^2), raiz do erro quadrático médio (RMSE) dos conjuntos de calibração e validação separados de forma aleatória, tanto no Python como no The Unscrambler X. Os dados foram modelados sem nenhuma exclusão de outliers no software The Unscrambler X e também no código Python. Então, os dados foram novamente modelados no código Python com exclusão de outliers com nível de significância de 95% e 99% para verificar os efeitos nos valores de R^2 e RMSE obtidos.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=0}^n (\hat{y}_i - y_i)^2} \quad (1)$$

4.3. Resultados

Na tabela 1 verificamos que os modelos de PLS calibrados de forma autônoma no código Python possuem valores de RMSECV próximos ou inferiores com os valores obtidos no software comercial The Unscrambler X (UNS) para o conjunto de dados sem exclusão de outliers. Isto se deve a otimização dos fatores pela rotina Python que busca o menor RMSECV para cada parâmetro em Y automaticamente. Ainda, uma pequena diferença entre UNS e a rotina Python pode ser atribuída ao algoritmo randômico de separação de dados de calibração e validação, tanto do UNS como do Python. Também podemos verificar que com base nos valores de RMSECV obtidos para BRIX e pH, estes parâmetros de qualidade possuem grande potencial de serem preditos apenas pela leitura do espectro em um NIR portátil diretamente na laranja sem destruí-la, assim como é observado na literatura (Cayuela & Weiland, 2010; Liu et al., 2010; Physics, 2017)(Farres et al., 2019; Magwaza, Opara, et al., 2014; A. Wang, Hu, & Xie, 2014).

Tabela 1. Conjunto de dados sem exclusão (Python n=306 com escolha automática de melhor fator, UNS = Unscrambler, n=306 com fator recomendado)

Software	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS
Parâmetro	BRIX	BRIX	ATT	ATT	RATIO	RATIO	pH	pH
RMSEC	0.8416	0.7989	0.354	0.3345	2.7508	2.5277	0.1910	0.1761
R2CAL	0.6542	0.6790	0.5143	0.5049	0.2479	0.3041	0.4890	0.5603
RMSECV	0.7897	0.8629	0.3357	0.3490	2.3807	2.6253	0.1711	0.1867
R2CV	0.6575	0.6286	0.4242	0.4653	0.2781	0.2525	0.5874	0.5094

A tabela 2 mostra que os modelos de calibração desenvolvidos de forma autônoma pelo Python com intervalo de confiança de 99% para exclusão de dados considerando as distancias de Mahalanobis para as componentes principais 1 e 2 tanto da matriz X como da matriz Y reduziu os valores de RMSECV dos modelos de PLS no conjunto de dados com exclusão de outliers detectados. Utilizando o limite de confiança de 99% foram removidas 12 amostras, que correspondem aos 6 espectros com ruído adicionados de forma proposital para teste da rotina e mais 6 amostras da matriz Y que são de origem do laboratório, totalizando assim, 3.92% de dados removidos. A redução dos valores de RMSECV e a baixa porcentagem de dados removidos pela rotina de Mahalanobis demonstraram alto potencial de avaliação autônoma em grande volume de dados com possíveis fontes de erros e outliers. Os dados encontrados pela rotina Python desenvolvida com a aplicação de Mahalanobis nas componentes principais batem com o que é reportado na literatura onde foram utilizados a distância de Mahalanobis modificada associada aos pesos do PLS para encontrar outliers em dados de espectroscopia Raman que é uma técnica fotônica de alta resolução que pode proporcionar, em poucos segundos, informação química e estrutural de quase qualquer material (Szostak & Mazurek, 2009; Thennadil et al., 2018; Zhao, Wu, Cheng, Shi, & Qiao, 2016).

Tabela 2. Conjunto de dados com exclusão limite de confiança 99% (Python n=294, outliers n=12, UNS = Unscrambler n=306)

Software	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS
Parâmetro	BRIX	BRIX	ATT	ATT	RATIO	RATIO	pH	pH
RMSEC	0.6991	0.7989	0.3190	0.3345	2.2795	2.5277	0.1812	0.1761
R2CAL	0.7611	0.6790	0.4770	0.5049	0.3233	0.3041	0.5613	0.5603
RMSECV	0.6794	0.8629	0.2717	0.3490	2.1952	2.6253	0.1457	0.1867
R2CV	0.7411	0.6286	0.5656	0.4653	0.3279	0.2525	0.6613	0.5094

A figura 3 mostra as amostras identificadas como outliers tanto na matriz X como na matriz Y através dos gráficos reconstruídos das componentes principais 1 e 2. Nas figuras podemos ver a alta distância das amostras e a evidência da rotina em detectar problemas nos espectros de NIR e no laboratório de forma autônoma dentro do limite de detecção configurado.

Figura 3. Análise de componentes principais e seleção das amostras consideradas outliers pela rotina Python via Mahalanobis 99% de confiança.

Na figura 4, é apresentado os espectros considerados outliers e os espectros selecionados para a calibração das curvas de PLS. Nela podemos verificar a média e o desvio padrão dos espectros, onde, fica nítido a presença de amostras com ruídos excessivos e em seguida os dados selecionados onde restam apenas os espectros selecionados com informação química relevante dentro dos limites de confiança de 99%.

Figura 4. Espectros de NIR selecionados e espectros de NIR indicados como outliers pela rotina Python a um limite de 99% de confiança.

A tabela 3 mostra que os modelos de calibração desenvolvidos de forma autônoma pelo Python com intervalo de confiança de 95% para exclusão de dados reduziram os valores de RMSECV dos modelos de PLS no conjunto de dados com exclusão de outliers detectados ainda mais. Entretanto, utilizando o limite de confiança de 95% foram removidas 54 amostras, que correspondem aos 6 espectros com ruído adicionados de forma proposital para teste da rotina e mais 11 espectros sem ruídos adicionados de forma proposital e mais 37 amostras da matriz Y que são de origem do laboratório, totalizando assim, 17.64% de dados removidos (Kubiak, Zhang, Ren, Yang, & Roskos, 2018). A redução dos valores de RMSECV e a alta porcentagem de dados removidos pela rotina de Mahalanobis para 95% de confiança demonstraram que se o limite de confiança for forçado, um grande número de amostras podem ser removidas e isto pode levar ao ‘*overfitting*’ ou super ajuste dos modelos de PLS o que não é indicado, pois, tira a robustez dos modelos de calibração (Gan & Ng, 2017; Kubiak et al., 2018).

Tabela 3. Conjunto de dados com exclusão limite de confiança 95% (Python n=252,outliers n=54, UNS = Unscrambler n=306)

Software	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS	PYTHON	UNS
Parâmetro	BRIX	BRIX	ATT	ATT	RATIO	RATIO	pH	pH
RMSEC	0.7785	0.718	0.2139	0.2952	1.7898	2.298	0.1325	0.1472
R2CAL	0.5897	0.7423	0.6565	0.6196	0.4770	0.4345	0.6607	0.6945
RMSECV	0.6473	0.7338	0.1974	0.3111	1.7989	2.4079	0.1566	0.1578
R2CV	0.7362	0.7328	0.6478	0.5776	0.3477	0.3798	0.5296	0.6491

A figura 5 mostra as amostras identificadas como outliers tanto na matriz X como na matriz Y através dos gráficos reconstruídos das componentes principais 1 e 2. Nas figuras podemos ver que muitas amostras foram removidas dentro do limite de confiança de 95% de forma autônoma, incluído até amostras que não são erros de leitura, mas sim amostras diferentes das demais ao nível de 95%. Amostras ‘*outliers*’ ou anômalas são amostras que podem conter erro, mas também podem ser só amostras diferentes das demais ao nível de significância estipulado, o que indica que durante a amostragem a variação total existente na qualidade de laranjas ainda não foi atingida, indicando que ainda existe a necessidade dos modelos de PLS serem incrementados com mais variabilidade. Na literatura diversos trabalhos com NIR em frutas mostram a importância de acompanhar as mudanças das frutas desde o início, meio e fim de safra (Milczarek et al., 2019). Isto é realizado para poder obter todas as variações químicas dos espectros NIR lidos diretamente na casca das frutas para que seja possível calibrar modelos de PLS com robustez (Malegori et al., 2017b; Maniwaru et al., 2019; Zhou, Zuo, Xu, & Wang, 2020).

Figura 5. Análise de componentes principais e seleção das amostras consideradas outliers pela rotina Python via Mahalanobis 95% de confiança.

Na figura 6, é apresentado os espectros considerados outliers e os espectros selecionados para a calibração das curvas de PLS. Nela podemos verificar a média e o desvio padrão dos espectros, onde, fica nítido a presença de amostras com ruídos excessivos, mas com a presença de sinais bons no conjunto considerado outlier. Mostrando que limites de confiança muito apertados como o 95% pode remover sinais de amostras diferentes sendo consideradas como outliers.

Figura 6. Espectros de NIR selecionados e espectros de NIR indicados como outliers pela rotina Python a um limite de 99% de confiança.

A rotina Python desenvolvida demonstrou alto potencial na análise autônoma de dados de NIR e de laboratório para calibração de curvas de PLS com exclusão automática de erros nos conjuntos de dados. Entretanto, os limites de confiança de 99% são indicados em relação aos limites de 95% para as distâncias de Mahalanobis devido a excluir menos amostras e remover apenas problemas reais de outliers e não remover amostras que podem participar ativamente os modelos de calibração e tornar as curvas mais robustas. Esta rotina desenvolvida pode ser aplicada a diversos conjuntos de dados e será bem aproveitada em grandes volumes de dados como gerados por indústrias onde muitos erros como de digitação de dados do laboratório e quando os espectros NIR vem com problemas na aquisição por alguma razão desconhecida na operação.

4.4. Conclusão

O código desenvolvido em Python se mostrou rápido e eficiente para lidar com conjuntos de dados de espectroscopia de infravermelho próximo e criar modelos de regressão multivariados via PLS. A exclusão automática de outliers utilizando a distância de Mahalanobis é eficiente para remover espectros de NIR ruidosos ou com problemas durante a aquisição de forma autônoma. Entretanto, utilizar limites de confiança de 99% é preferível ao limite de 95% devido ao grande número de amostras que podem ser removidas. A eficiência quando comparamos o número de amostras excluídas com 95% não compensa nos valores de R^2 e RMSE quando comparados aos valores obtidos com limite de confiança de 99% para o conjunto de dados avaliado. Ainda, os valores obtidos no software comercial The Unscrambler X é muito próximo aos valores obtidos no código Python variando apenas devido aos algoritmos randômicos de separação de dados para calibração e validação entre os dois softwares e a escolha dos fatores de cada modelo de PLS.

Portanto, o código desenvolvido em Python (gratuito) pode ser utilizado para avaliar de forma extremamente rápida e autônoma o potencial de espectros NIR para prever parâmetros de qualidade em alimentos em conjuntos de dados grandes com apenas poucos cliques.

REFERÊNCIAS

- Bizzani, M., Flores, D. W. M., Colnago, L. A., & Ferreira, M. D. (2017). Non-invasive spectroscopic methods to estimate orange firmness, peel thickness, and total pectin content. *Microchemical Journal*, 133, 168–174. <https://doi.org/10.1016/j.microc.2017.03.039>
- Butturi-Gomes, D., & Petrere, M. (2020). Edge influence and population aggregation: On point and interval statistical performances of Morisita patchiness index estimators in different sampling schemes. *Ecological Indicators*. <https://doi.org/10.1016/j.ecolind.2019.105736>
- Camo Analytics. (2019). *The Unscrambler X*. Oslo, Norway.
- Cayuela, J. A., & Weiland, C. (2010). Intact orange quality prediction with two portable NIR spectrometers. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2010.06.001>
- Dahdouh, L., Wisniewski, C., Ricci, J., Vachoud, L., Dornier, M., & Delalonde, M. (2015). Rheological study of orange juices for a better knowledge of their suspended solids interactions at low and high concentration. *Journal of Food Engineering*, 174, 15–20. <https://doi.org/10.1016/j.jfoodeng.2015.11.008>
- Dambros, C. S., Morais, J. W., Vasconcellos, A., & Franklin, E. (2020). Defining a termite sampling protocol for ecological studies: An effective method to increase statistical power. *European Journal of Soil Biology*. <https://doi.org/10.1016/j.ejsobi.2019.103145>
- Doupe, P., Faghmous, J., & Basu, S. (2019). Machine Learning for Health Services Researchers. *Value in Health*. <https://doi.org/10.1016/j.jval.2019.02.012>
- Eigenvector Research, I. (2019). *PLS Toolbox*. Retrieved from <https://eigenvector.com/software/pls-toolbox/>
- Farres, S., Srata, L., Fethi, F., & Kadaoui, A. (2019). Argan oil authentication using visible/near infrared spectroscopy combined to chemometrics tools. *Vibrational Spectroscopy*, 102(September 2018), 79–84. <https://doi.org/10.1016/j.vibspec.2019.04.003>
- Filzmoser, P. (2004). A multivariate outlier detection method. *Seventh International Conference on Computer Data Analysis and Modeling*, 1(1989), 18–22. Retrieved from <http://computerwranglers.com/com531/handouts/mahalanobis.pdf>

Gan, G., & Ng, M. K. P. (2017). k-means clustering with outlier removal. *Pattern Recognition Letters*. <https://doi.org/10.1016/j.patrec.2017.03.008>

García-Santillán, I. D., & Pajares, G. (2018). On-line crop/weed discrimination through the Mahalanobis distance from images in maize fields. *Biosystems Engineering*. <https://doi.org/10.1016/j.biosystemseng.2017.11.003>

Hao, C., Ma, J., Xu, N., Zhao, Q., Du, J., & Zhu, K. (2019). Uncertainty propagation analysis for control rod worth of PWR based on the statistical sampling method. *Annals of Nuclear Energy*. <https://doi.org/10.1016/j.anucene.2019.107054>

Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., ... Foster, A. S. (2019). DDescribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*. <https://doi.org/10.1016/j.cpc.2019.106949>

Instituto Adolfo Lutz. (2008). Análise sensorial. In *Métodos Físico-Químicos para Análise de Alimentos*.

JetBrains. (2019). *PyCharm*. Retrieved from <https://www.jetbrains.com/pycharm/>

Jiang, H., Ching, W. K., Yiu, K. F. C., & Qiu, Y. (2018). Stationary Mahalanobis kernel SVM for credit risk evaluation. *Applied Soft Computing Journal*. <https://doi.org/10.1016/j.asoc.2018.07.005>

Kubiak, R. J., Zhang, J., Ren, P., Yang, H., & Roskos, L. K. (2018). Excessive outlier removal may result in cut points that are not suitable for immunogenicity assessments. *Journal of Immunological Methods*. <https://doi.org/10.1016/j.jim.2018.10.001>

Lee, K. M., Yoo, J., Kim, S. W., Lee, J. H., & Hong, J. (2019). Autonomic machine learning platform. *International Journal of Information Management*. <https://doi.org/10.1016/j.ijinfomgt.2019.07.003>

Liu, Y., Sun, X., & Ouyang, A. (2010). Nondestructive measurement of soluble solid content of navel orange fruit by visible-NIR spectrometric technique with PLSR and PCA-BPNN. *LWT - Food Science and Technology*, 43(4), 602–607. <https://doi.org/10.1016/j.lwt.2009.10.008>

Luo, L., Bao, S., & Peng, X. (2019). Robust monitoring of industrial processes using process data with outliers and missing values. *Chemometrics and Intelligent Laboratory Systems*. <https://doi.org/10.1016/j.chemolab.2019.103827>

Magwaza, L. S., Opara, U. L., Cronje, P. J. R., Landahl, S., Nieuwoudt, H. H., Mouazen, A. M., ... Terry, L. a. (2014). Assessment of rind quality of “Nules Clementine” mandarin fruit during postharvest storage: 2. Robust Vis/NIRS PLS models for prediction of physico-chemical attributes. *Scientia Horticulturae*, 165, 421–432. <https://doi.org/10.1016/j.scienta.2013.09.050>

Malegori, C., Nascimento Marques, E. J., de Freitas, S. T., Pimentel, M. F., Pasquini, C., & Casiraghi, E. (2017). Comparing the analytical performances of Micro-NIR and FT-NIR spectrometers in the evaluation of acerola fruit quality, using PLS and SVM regression algorithms. *Talanta*. <https://doi.org/10.1016/j.talanta.2016.12.035>

Maniwara, P., Nakano, K., Ohashi, S., Boonyakiat, D., Seehanam, P., Theanjumol, P., & Poonlarp, P. (2019). Evaluation of NIRS as non-destructive test to evaluate quality traits of purple passion fruit. *Scientia Horticulturae*. <https://doi.org/10.1016/j.scienta.2019.108712>

Mendes, F. M., & Gontijo, C. C. (2017). Kpop: A Python package for population genetics analysis. *Forensic Science International: Genetics Supplement Series*. <https://doi.org/10.1016/j.fsigss.2017.09.159>

Milczarek, R. R., Liang, P. S., Wong, T., Augustine, M. P., Smith, J. L., Woods, R. D., ... Breksa, A. P. (2019). Nondestructive determination of the astringency of pollination-variant persimmons (*Diospyros kaki*) using near-infrared (NIR) spectroscopy and nuclear magnetic resonance (NMR) relaxometry. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2018.11.006>

Moraes, M. B. De, & Gradvohl, A. L. S. (2013). Viabilidade da Linguagem Python no Desenvolvimento de Aplicações de Alto Desempenho. *IV Escola Regional de Alto Desempenho de São Paulo*, (June 2014), 147–150. Retrieved from https://www.researchgate.net/profile/Andre_Gradvohl/publication/253357879_Viabilidade_da_Linguagem_Python_no_Desenvolvimento_de_Aplicacoes_de_Alto_Desempenho/links/0046351f90691297f9000000/Viabilidade-da-Linguagem-Python-no-Desenvolvimento-de-Aplicacoes-d

Ncama, K., Opara, U. L., Tesfay, S. Z., Fawole, O. A., & Magwaza, L. S. (2017). Application of Vis/NIR spectroscopy for predicting sweetness and flavour parameters of ‘Valencia’ orange (*Citrus sinensis*) and ‘Star Ruby’ grapefruit (*Citrus x paradisi* Macfad). *Journal of Food Engineering*. <https://doi.org/10.1016/j.jfoodeng.2016.08.015>

O'Brien, N. a., Hulse, C. a., Friedrich, D. M., Van Milligen, F. J., von Gunten, M. K., Pfeifer, F., & Siesler, H. W. (2012). *Miniature near-infrared (NIR) spectrometer engine for handheld applications*. 8374, 837404. <https://doi.org/10.1117/12.917983>

Pasquini, C. (2018). Analytica Chimica Acta Near infrared spectroscopy : A mature analytical technique with new perspectives e A review. *Analytica Chimica Acta*, 1026, 8–36. <https://doi.org/10.1016/j.aca.2018.04.004>

Pedregosa, F., Weiss, R., & Brucher, M. (2011). *Scikit-learn : Machine Learning in Python*. 12, 2825–2830.

Pérez-Marín, D., Calero, L., Fearn, T., Torres, I., Garrido-Varo, A., & Sánchez, M. T. (2019). A system using in situ NIRS sensors for the detection of product failing to meet quality standards and the prediction of optimal postharvest shelf-life in the case of oranges kept in cold storage. *Postharvest Biology and Technology*. <https://doi.org/10.1016/j.postharvbio.2018.09.009>

Phuphaphud, A., Saengprachatanarug, K., Posom, J., & Maraphum, K. (2019). *Vibrational Spectroscopy Prediction of the fibre content of sugarcane stalk by direct scanning using visible-shortwave near infrared spectroscopy*. 101(February), 71–80.

Physics, C. (2017). SCIO POCKET-SIZED MICRO SPECTROMETER. Retrieved from <https://www.consumerphysics.com/business/>

Ruiz de la Hermosa González-Carrato, R. (2018). Wind farm monitoring using Mahalanobis distance and fuzzy clustering. *Renewable Energy*. <https://doi.org/10.1016/j.renene.2018.02.097>

Spectral Solutions. (2019). SCIR. Retrieved from www.spectralsolutions.com.br

Szostak, R., & Mazurek, S. (2009). Simple transformation of spectra to effectively reduce quantification errors in FT-Raman multivariate analysis of complex systems. *Vibrational Spectroscopy*. <https://doi.org/10.1016/j.vibspec.2008.09.001>

Tao, F., Wang, T., Wu, J., & Lin, X. (2020). A novel KA-STAP method based on Mahalanobis distance metric learning. *Digital Signal Processing: A Review Journal*. <https://doi.org/10.1016/j.dsp.2019.102613>

Tavakolian, M. S. S., Silaghi, F. A., Fabbri, A., Molari, G., Giunchi, A., & Guarnieri, A. (2013). Differentiation of post harvest date fruit varieties non-destructively using FT-NIR spectroscopy. *International Journal of Food Science and Technology*, 48(6), 1282–1288. <https://doi.org/10.1111/ijfs.12088>

Teye, E., Amuah, C. L. Y., Mcgrath, T., & Elliott, C. (2019). *Spectrochimica Acta Part A : Molecular and Biomolecular Spectroscopy Innovative and rapid analysis for rice authenticity using hand-held NIR spectrometry and chemometrics*. 217, 147–154.

Thennadil, S. N., Dewar, M., Herdsman, C., Nordon, A., & Becker, E. (2018). Automated weighted outlier detection technique for multivariate data. *Control Engineering Practice*. <https://doi.org/10.1016/j.conengprac.2017.09.018>

Varoquaux, G., Buitinck, L., Louppe, G., Grisel, O., Pedregosa, F., & Mueller, A. (2015). Scikit-learn. *GetMobile: Mobile Computing and Communications*, 19(1), 29–33. <https://doi.org/10.1145/2786984.2786995>

VIAMI. (2019). MicroNIR onsite W. Retrieved December 2, 2019, from <https://www.viavisolutions.com/en-us/osp/products/micronir-onsite-w>

Wang, A., Hu, D., & Xie, L. (2014). Comparison of detection modes in terms of the necessity of visible region (VIS) and influence of the peel on soluble solids content (SSC) determination of navel orange using VIS-SWNIR spectroscopy. *Journal of Food Engineering*, 126, 126–132. <https://doi.org/10.1016/j.jfoodeng.2013.11.011>

Wang, C., Liu, Z., Gao, H., & Fu, Y. (2019). VOS: A new outlier detection model using virtual graph. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2019.104907>

Wang, J., Du, P., Hao, Y., Ma, X., Niu, T., & Yang, W. (2020). An innovative hybrid model based on outlier detection and correction algorithm and heuristic intelligent optimization algorithm for daily air quality index forecasting. *Journal of Environmental Management*. <https://doi.org/10.1016/j.jenvman.2019.109855>

Wu, G., & Wang, C. (2014). Investigating the effects of simulated transport vibration on tomato tissue damage based on vis/NIR spectroscopy. *Postharvest Biology and Technology*, 98, 41–47. <https://doi.org/10.1016/j.postharvbio.2014.06.016>

Xu, H., Qi, B., Sun, T., Fu, X., & Ying, Y. (2012). Variable selection in visible and near-infrared spectra: Application to on-line determination of sugar content in pears. *Journal of Food Engineering*, 109(1), 142–147. <https://doi.org/10.1016/j.jfoodeng.2011.09.022>

Zhao, N., Wu, Z., Cheng, Y., Shi, X., & Qiao, Y. (2016). MDL and RMSEP assessment of spectral pretreatments by adding different noises in calibration/validation datasets. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. <https://doi.org/10.1016/j.saa.2016.03.017>

Zhou, Y., Zuo, Z., Xu, F., & Wang, Y. (2020). Origin identification of *Panax notoginseng* by multi-sensor information fusion strategy of infrared spectra combined with random forest. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*. <https://doi.org/10.1016/j.saa.2019.117619>

ANEXOS

ANEXO A. Programa em Python para análise autônoma de dados NIR com aplicação em análise de alimentos

Estrutura dos dados

O formato de dados deve ser em planilha Excel (.xls ou .xlsx). Quando executado o programa irá permitir que você escolha a sua planilha com os dados a serem analisados. A figura abaixo mostra como a planilha Excel deve estar formatada. A planilha deve conter três abas.

A primeira aba ('Labels') deve ser uma coluna, onde, a primeira célula é o cabeçalho da coluna. Esta primeira coluna é denominada 'Labels' que será os nomes das amostras. Cada linha representa uma amostra nova.

A segunda aba ('X') deve ser dos dados independentes (X), geralmente as variáveis do espectro ou conjunto de dados de alguma instrumentação. A primeira linha será considerada como cabeçalho e deve conter o nome de cada variável da coluna (numero de onda, tempo, etc.). Na sequência cada sinal do instrumento deve ser na mesma ordem em que a aba anterior, onde, cada linha é pertencente e alinhada a mesma amostra.

A terceira aba ('Y') deve ser os dados dependentes (Y), geralmente as propriedades químicas ou físicas de interesse que desejamos modelar baseado nos dados da instrumentação. Da mesma forma que as demais, a primeira linha será o cabeçalho com os nomes das propriedades/parâmetros de qualidade e as linhas as amostras alinhadas na mesma ordem das abas 'Labels' e 'X'.

MicronirCitrusDataset - Microsoft Excel

Pronto

100%

MicronirCitrusDataset - Microsoft Excel

Pronto

100%

	A	B	C	D	E	F	G	H	I	J	K	L
1	Labels											
2	1											
3	2											
4	3											
5	4											
6	5											
7	6											
8	7											
9	8											
10	9											
11	10											
12	11											
13	12											
14	13											
15	14											
16	15											
17	16											
18	17											
19	18											
CP13	0											
1	1045.48	1049.65	1053.853	1058.09	1062.362	1066.668	1071.009	1075.386	1079.798	1084.247	1088.733	1093.256
2	-0.01138	-0.00931	-0.00722	-0.00509	-0.00279	-0.0004	0.002171	0.00476	0.007386	0.009636	0.011174	0.01273
3	-0.01315	-0.01075	-0.00816	-0.00545	-0.0028	0.000173	0.003244	0.00648	0.009781	0.012384	0.014073	0.015768
4	-0.01264	-0.01027	-0.00783	-0.00534	-0.00286	-3.68E-05	0.002831	0.00585	0.008966	0.011437	0.013161	0.014889
5	-0.01281	-0.01042	-0.00794	-0.00539	-0.00278	-0.00012	0.002767	0.005902	0.009205	0.01196	0.013626	0.015364
6	-0.0115	-0.00933	-0.00716	-0.00496	-0.00256	-7.43E-05	0.002557	0.005382	0.008199	0.010527	0.012123	0.013648
7	-0.00999	-0.00826	-0.00631	-0.00439	-0.00241	-0.00033	0.001919	0.004259	0.006656	0.008678	0.00997	0.011507
8	-0.01045	-0.00856	-0.00663	-0.00461	-0.00245	-0.00017	0.002148	0.004515	0.006943	0.009004	0.010402	0.011748
9	-0.01199	-0.00984	-0.00758	-0.00524	-0.00284	-0.00031	0.002379	0.005213	0.00821	0.010651	0.01225	0.014017
10	-0.0124	-0.0101	-0.0077	-0.00533	-0.00283	-0.00023	0.002521	0.005516	0.008527	0.011	0.012634	0.01437
11	-0.0113	-0.00925	-0.00703	-0.00474	-0.00251	-3.49E-05	0.002492	0.00519	0.008073	0.010372	0.011879	0.013459
12	-0.01441	-0.01162	-0.00873	-0.00571	-0.00267	0.000571	0.003883	0.007241	0.010766	0.013616	0.015501	0.017226
13	-0.01095	-0.0089	-0.00674	-0.00457	-0.00246	-0.00018	0.002197	0.00473	0.007439	0.009697	0.01123	0.012751
14	-0.01274	-0.01032	-0.00788	-0.00529	-0.00256	0.000388	0.003246	0.006229	0.009196	0.011764	0.013612	0.015262
15	-0.01248	-0.01027	-0.00773	-0.00525	-0.00274	-3.44E-05	0.002686	0.005726	0.008915	0.01145	0.013179	0.014933
16	-0.01038	-0.00851	-0.00659	-0.00455	-0.00249	-0.00029	0.00202	0.004443	0.006956	0.009018	0.010522	0.011915
17	-0.01079	-0.00878	-0.00668	-0.00465	-0.00252	-0.00028	0.002106	0.004688	0.00737	0.009605	0.011066	0.012626
18	-0.01058	-0.00866	-0.00669	-0.0047	-0.00259	-0.00033	0.001994	0.004485	0.007076	0.009241	0.010616	0.012099
19	-0.01305	-0.01055	-0.00803	-0.0054	-0.00277	2.85E-05	0.003023	0.006142	0.009371	0.012027	0.013808	0.015554

	A	B	C	D	E	F	G	H	I	J	K	L
1	SST	ATT	RATIO	Peso	pH	Altura	Largura	Formato				
2	10.83	0.84	12.87	201.06	3.74	73.4	73.46	1				
3	9.74	0.92	10.57	245.71	3.64	80.56	77.7	1.04				
4	9	0.7	12.9	203.08	3.73	76.15	72.86	1.05				
5	9.66	0.77	12.56	210.55	3.85	73.07	73.61	0.99				
6	10.38	0.94	11.03	206.26	3.57	73.83	73.47	1				
7	11.02	0.99	11.11	239.23	3.52	74.29	79.97	0.93				
8	11	0.79	13.86	189.68	3.81	74.66	69.32	1.08				
9	10.83	1.02	10.58	170.88	3.48	64.13	69.95	0.92				
10	9.97	0.77	12.87	262.89	3.79	77.71	81.41	0.95				
11	9.07	0.68	13.37	211.25	3.93	75.32	75.11	1				
12	9.1	1.02	8.94	306.76	3.68	85.36	82.77	1.03				
13	10.83	0.84	12.82	228.22	3.7	75.55	76.45	0.99				
14	8.64	0.92	9.38	278.56	3.75	75.48	84.51	0.89				
15	8.96	0.75	11.97	277.34	3.79	83.19	82.8	1				
16	10.95	0.77	14.21	206.82	3.73	76.16	72.3	1.05				
17	11.22	0.83	13.49	198.33	3.66	72.26	72.13	1				
18	9.99	0.64	15.5	216.14	3.66	78.3	74.03	1.06				
19	9.81	0.85	11.52	286.11	3.57	84.33	83.3	1.01				

Figura 1. Formato da planilha Excel com os dados de input.

Executando o programa

Com a planilha formatada, o arquivo foodscience-ml.py deve ser executado e o seguinte popup será inicializado (Figura 2). Escolha o local do seu arquivo e aperte ok para continuar (Figura 3).

Figura 2. Popup de inicialização.

Figura 5. Pasta gerada na sua área de trabalho.

Compreendendo os resultados

Dentro da pasta gerada irá conter os seguintes arquivos: PLS_Report, Outliers_Report, Mahalanobis_report e uma pasta 'Figures'. Dentro da pasta 'Figures' você irá encontrar todos os gráficos dos modelos de regressão criados, um mapa de correlação entre as variáveis Y e uma análise de detecção autônoma de outliers/amostras anômalas.

O arquivo PLS_Report irá trazer o resumo estatístico dos modelos de regressão multivariados treinados e testados no seu conjunto de dados.

O arquivo Outliers_Report irá trazer a % de amostras detectadas como anômalas as demais no seu conjunto de dados.

O arquivo Mahalanobis_report irá trazer todos os dados com os respectivos nomes das amostras selecionados para treinar e testar os modelos de regressão bem como em uma aba separada da planilha só os dados detectados como anômalos para avaliação.

Na pasta Figures, irá encontrar as figuras como o exemplo a seguir:

Figura 6. Gráfico dos sinais abertos pela rotina.

Figura 7. Gráfico dos dados analisados pela rotina.

Figura 8. Gráfico do PLS desenvolvido pela rotina.

Figura 9. Gráfico do PLS desenvolvido pela rotina.

Figura 10. Gráfico do PLS desenvolvido pela rotina.

Figura 11. Gráfico do PLS desenvolvido pela rotina.

ANEXO B. Código fonte da classe foodscience-ml.py

```

import Classes.FileIO as Fio
import Classes.MissingData as Miss
import Classes.ExploratoryAnalysis as Exp
import Classes.Helpers as help
import Classes.Plots as Plot
from datetime import datetime
import matplotlib.pyplot as plt
import Classes.RegressionAnalysis as Reg
import pandas as pd
import Classes.Configurations as cfg

# 1. import data
print('Starting at ' + str(datetime.now()))
outlier_confidence_level_x = cfg.outlier_confidence_level_x
outlier_confidence_level_y = cfg.outlier_confidence_level_y
confidence_pca = cfg.confidence_pca
labels, x, y, file_name = Fio.import_excel_data()
n_total = x.shape[0]

# 2. check if data has same n of rows
inputs_are_empty = Miss.check_matrix_empty(labels, x, y)
inputs_has_same_n = Miss.check_matrix_rows(labels, x, y)

if True == inputs_are_empty and inputs_has_same_n == False:
    print('Initializing routine failed to start at ' + str(datetime.now()))
    print('imported data is empty ? ' + str(inputs_are_empty))
    print('imported data has same size? ' + str(inputs_has_same_n))
    print('Routine was not executed and finished at ' + str(datetime.now()))
else:
    print('Starting machine learning routine at ' + str(datetime.now()))
    plt.style.use('seaborn')

# 3. plot imported data
if y.shape[1] >= 1:
    Plot.correlation_matrix(y, file_name)

# 4. exclude outliers based on Mahalanobis distance
x_data, y_data, df_cleaned_x, df_cleaned_xy, df_outliers_x, df_outliers_xy, n_outliers_x,
n_outliers_y, x_out, y_out = Exp.exclude_outliers(
    x, y, labels, outlier_confidence_level_x, outlier_confidence_level_y)
if n_outliers_y != 0 and n_outliers_x != 0:
    help.outliers_pcs_plot(df_cleaned_x, df_cleaned_xy, df_outliers_x, df_outliers_xy, labels,
n_outliers_x,
n_outliers_y, file_name)
Plot.x_data_outliers(x_data, x_out, file_name)

summary_outliers = ['N Total: ' + str(n_total),
'N Outliers X: ' + str(n_outliers_x) + ' - %: ' + str(((n_outliers_x * 100) / n_total)),
'N Outliers Y: ' + str(n_outliers_y) + ' - %: ' + str(
    ((n_outliers_y * 100) / (n_total - n_outliers_x)))]

# 5. build multivariate regression models
summary_models, models, x_train, y_train, x_test, y_test =
Reg.partial_least_square_regression(x_data, y_data,
cfg.train_split_percentage)
Plot.scatter_x_y(x_train, y_train, x_test, y_test, summary_models, models, file_name)

# 6. export full report
Fio.data_to_excel(file_name, df_cleaned_xy, pd.concat([df_outliers_x, df_outliers_xy],
axis=0, sort=True))
Fio.summary_data_to_excel(summary_models, file_name)
Fio.summary_outlier_to_excel(pd.DataFrame(summary_outliers), file_name)

```

```
print('Done at ' + str(datetime.now()))
print('Hello Github!')
plt.show()
```

ANEXO C. Código fonte da classe FileIO.py

```
import pandas as pd
import tkinter as tk
from tkinter import filedialog
from tkinter import messagebox
import os
import joblib

def import_excel_data():
    try:
        root = tk.Tk()
        root.withdraw()
        messagebox.showinfo("Import data (Excel file)", "Choose your matrix of inputs.")
        file_path = filedialog.askopenfilename()
        file_name = os.path.basename(file_path)
        labels = pd.read_excel(file_path, 0)
        label_df = pd.DataFrame(labels)
        xdata = pd.read_excel(file_path, 1)
        x_df = pd.DataFrame(xdata)
        ydata = pd.read_excel(file_path, 2)
        y_df = pd.DataFrame(ydata)
        return label_df, x_df, y_df, file_name
    except:
        print("An exception occurred while importing excel file.")

def import_excel_data_epo():
    try:
        root = tk.Tk()
        root.withdraw()
        messagebox.showinfo("Import data (Excel file)", "Choose your matrix of inputs.")
        file_path = filedialog.askopenfilename()
        file_name = os.path.basename(file_path)
        labels = pd.read_excel(file_path, 0)
        label_df = pd.DataFrame(labels)
        xdatau = pd.read_excel(file_path, 1)
        x_df_u = pd.DataFrame(xdatau)
        xdatas = pd.read_excel(file_path, 2)
        x_df_s = pd.DataFrame(xdatas)
        ydata = pd.read_excel(file_path, 3)
        y = pd.DataFrame(ydata)
        return label_df, x_df_u, x_df_s, y, file_name
    except:
        print("An exception occurred while importing excel file.")

def data_to_excel(file_name, all_df, outliers_df):
    path = os.path.expanduser("~/Desktop") + '/foodscienceml' + '/' + file_name.replace('.xlsx',
    '').replace('.xls', '')
    if not os.path.exists(path):
        os.makedirs(path)
    writer = pd.ExcelWriter(path + '\Mahalanobis_report.xlsx')
    all_data = pd.DataFrame(all_df)
    all_data.to_excel(writer, 'Selected data')
    outliers_data = pd.DataFrame(outliers_df)
    outliers_data.to_excel(writer, 'Outliers data')
    writer.save()

def summary_data_to_excel(df, file_name):
    path = os.path.expanduser("~/Desktop") + '/foodscienceml' + '/' + file_name.replace('.xlsx',
    '').replace('.xls', '')
    if not os.path.exists(path):
        os.makedirs(path)

    writer = pd.ExcelWriter(path + '\PLS_Report.xlsx')
    all_data = pd.DataFrame(df)
```



```

all_data.to_excel(writer, 'PLS SUMMARY')
writer.save()

def summary_outlier_to_excel(df, file_name):
    path = os.path.expanduser("~/Desktop") + '/foodscienceml' + '/' + file_name.replace('.xlsx',
    '').replace('.xls', '')
    if not os.path.exists(path):
        os.makedirs(path)

    writer = pd.ExcelWriter(path + '\Outliers_Report.xlsx')
    all_data = pd.DataFrame(df)
    all_data.to_excel(writer, 'OUTLIERS SUMMARY')
    writer.save()

def data_describe_to_excel(df, df_outliers, file_name):
    path = os.path.expanduser("~/Desktop") + '/foodscienceml' + '/' + file_name.replace('.xlsx',
    '').replace('.xls', '')
    if not os.path.exists(path):
        os.makedirs(path)

    writer = pd.ExcelWriter(path + '\PCA_stats_report.xlsx')
    df_describe = pd.DataFrame(df.describe())
    df_describe.to_excel(writer, 'Descriptive statistics inliers')
    df_describe_out = pd.DataFrame(df_outliers.describe())
    df_describe_out.to_excel(writer, 'Descriptive statistics outliers')
    writer.save()

def save_model_to_pkl(model, model_name, file_name):
    path = os.path.expanduser("~/Desktop") + '/foodscienceml' + '/' + file_name.replace('.xlsx',
    '').replace('.xls', '') + '/PKL'
    if not os.path.exists(path):
        os.makedirs(path)

    file_name_pls_final = os.path.join(path, str(model_name) + '.pkl')
    joblib.dump(model, file_name_pls_final)

```

ANEXO D. Código fonte da classe RegressionAnalysis.py

```

from math import sqrt
import numpy as np
import pandas as pd
from sklearn.cross_decomposition import PLSRegression
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split

def partial_least_squares_regression(x, y, train_split_percentage):
    # Define X and Y matrix after cleaned by PCA and Mahalanobis distance
    x_df = pd.DataFrame(x)
    y_df = pd.DataFrame(y)

    # split data to train and test
    x_test, x_train, y_test, y_train = train_test_split(x_df, y_df, test_size=train_split_percentage,
    random_state=1)

    # Train one PLS model for each Y parameter
    parameters = len(y_df.columns)
    models = []
    rmsec = []
    r2cal = []
    rmsecv = []
    r2cv = []
    for i in range(parameters):
        i_model, i_rmse, i_r2 = do_pls(x_train, y_train.iloc[:, i], train_split_percentage)
        models.append(i_model)
        rmsec.append(i_rmse)
        r2cal.append(i_r2)
        predicted_cv_y = i_model.predict(x_test)
        rmsecv.append(sqrt(mean_squared_error(y_test.iloc[:, i], predicted_cv_y)))
        r2cv.append(i_model.score(x_test, y_test.iloc[:, i]))

```

```

df_models_summary = pd.DataFrame(
    pd.concat([pd.DataFrame(list(y_df.columns)), pd.DataFrame(rmse), pd.DataFrame(r2cal),
pd.DataFrame(rmsecv), pd.DataFrame(r2cv)], axis=1))
s = pd.Series(['Parameter', 'RMSEC', 'R2CAL', 'RMSECV', 'R2CV'])
df_models_summary = df_models_summary.transpose().set_index(s)

return df_models_summary, models, x_train, y_train, x_test, y_test

def do_pls(data_x, data_y, train_split_percentage):
    latent_variables = []
    x_train, x_test, y_train, y_test = train_test_split(data_x, data_y,
test_size=train_split_percentage, random_state=0)
    for i in range(20):
        pls = PLSRegression(n_components=(i + 1), scale=True, max_iter=500, tol=1e-08)
        pls.fit(x_train, y_train)
        predicted_cv_y = pls.predict(x_test)
        mean_squared_error_cv = sqrt(mean_squared_error(y_test, predicted_cv_y))
        latent_variables.append(mean_squared_error_cv)

    best_factor = np.argmin(latent_variables)
    pls = PLSRegression(n_components=(best_factor + 1), scale=True, max_iter=500, tol=1e-08)
    pls.fit(x_train, y_train)
    predicted_cv_y = pls.predict(x_test)
    rmsev = sqrt(mean_squared_error(y_test, predicted_cv_y))
    r2v = pls.score(x_test, y_test)
    pls = PLSRegression(n_components=best_factor + 1, scale=True, max_iter=500, tol=1e-08)
    pls.fit(data_x, data_y)
    return pls, rmsev, r2v

def run_pls(x, model):
    y_hat = model.predict(x)
    return y_hat

```

ANEXO E. Código fonte da classe ExploratoryAnalysis.py

```

import seaborn as sns; sns.set()
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import numpy as np
import scipy as sp

def principal_component_analysis(data_to_analyze):
    scale = StandardScaler(copy=True, with_mean=True, with_std=True)
    scale.fit(data_to_analyze)
    df = scale.transform(data_to_analyze)
    n_of_pcs = data_to_analyze.shape[1]-1
    if n_of_pcs > 20:
        n_of_pcs = 20
    pca_model = PCA(n_components=n_of_pcs)
    pcs = pca_model.fit_transform(df)
    pcs_labels = []
    for i in range(n_of_pcs):
        pc_label = ('PC' + str(i+1) + ' ' +
"{0:.2f}".format(pca_model.explained_variance_ratio_[i]*100) + '%')
        pcs_labels.append(pc_label)
    pcs = pd.DataFrame(pcs, columns=[pcs_labels])
    return pca_model, pcs

def do_mahalanobis(data_to_analyze):
    df = pd.DataFrame(data_to_analyze)
    model, pcs = extract_pc1_pc2(df, 2)
    pca_df = pd.DataFrame(data=pcs, columns=['PC1', 'PC2'])
    covariance_matrix = pca_df.cov()
    inverse_covariance_matrix = sp.linalg.inv(covariance_matrix)
    xy_mean = pca_df['PC1'].mean(), pca_df['PC2'].mean()
    x_diff = np.array([x_i - xy_mean[0] for x_i in pca_df['PC1']])
    y_diff = np.array([y_i - xy_mean[1] for y_i in pca_df['PC2']])

```

```

diff_xy = np.transpose([x_diff, y_diff])
md = []
for i in range(len(diff_xy)):
    md.append(np.sqrt(np.dot(np.dot(np.transpose(diff_xy[i]), inverse_covariance_matrix),
diff_xy[i])))
mahalanobis_distance = pd.DataFrame(data=md, columns=['Mahalanobis'])
df = pd.concat([df.reset_index(), pca_df, mahalanobis_distance], axis=1)
df.index = pd.DataFrame(data_to_analyze).index
df.drop(['index'], axis=1, inplace=True)
return df

def drop_mahalanobis_outliers(data_to_analyze, sigma_confidence):
    df = pd.DataFrame(data_to_analyze)
    df_out = df.copy(deep=True)
    md = df['Mahalanobis']
    limit = sigma_confidence
    df_bool = df['Mahalanobis'] < limit
    df = df[df_bool]
    df = df.dropna()
    df_bool_out = df_out['Mahalanobis'] > limit
    df_out = df_out[df_bool_out]
    df_out = df_out.dropna()
    pcs_cleaned = pd.DataFrame(df)
    df.drop(['PC1'], axis=1, inplace=True)
    df.drop(['PC2'], axis=1, inplace=True)
    df.drop(['Mahalanobis'], axis=1, inplace=True)
    pcs_outliers = pd.DataFrame(df_out)
    df_out.drop(['PC1'], axis=1, inplace=True)
    df_out.drop(['PC2'], axis=1, inplace=True)
    df_out.drop(['Mahalanobis'], axis=1, inplace=True)
    return df_out, df, pcs_outliers, pcs_cleaned

def extract_pc1_pc2(data_to_analyze, n_components):
    scale = StandardScaler(copy=True, with_mean=True, with_std=True)
    scale.fit(data_to_analyze)
    df = scale.transform(data_to_analyze)
    pca_model = PCA(n_components=n_components)
    pcs = pca_model.fit_transform(df)
    pcs = pd.DataFrame(pcs, columns=['PC1', 'PC2'])
    return pca_model, pcs

def exclude_outliers(x, y, labels, outlier_confidence_level_x, outlier_confidence_level_y):
    main_df = pd.DataFrame(pd.concat([x, y], axis=1))
    main_df.index = labels
    x_shape = x.shape[1]
    y_start = x_shape

    n_total = x.shape[0]
    mahalanobis_x = pd.DataFrame(do_mahalanobis(main_df.iloc[:, 0:y_start]))
    df_cleaned_x = pd.DataFrame(pd.concat([mahalanobis_x, main_df.iloc[:, y_start:]], axis=1))
    df_outliers_x, df_cleaned_x, pcs_out_x, pcs_cleaned_x = drop_mahalanobis_outliers(df_cleaned_x,
outlier_confidence_level_x)
    n_outliers_x = df_outliers_x.shape[0]

    mahalanobis_y = pd.DataFrame(do_mahalanobis(df_cleaned_x.iloc[:, y_start:]))
    df_cleaned_xy = pd.DataFrame(
        pd.concat([df_cleaned_x.iloc[:, 0:y_start], mahalanobis_y], axis=1))
    df_outliers_xy, df_cleaned_xy, pcs_out_y, pcs_cleaned_y = drop_mahalanobis_outliers(df_cleaned_xy,
outlier_confidence_level_y)
    n_outliers_y = df_outliers_xy.shape[0]

    x_data = pd.DataFrame(df_cleaned_xy.iloc[:, 0:y_start])
    y_data = pd.DataFrame(df_cleaned_xy.iloc[:, y_start:])

    x_out = pd.DataFrame(df_outliers_x.iloc[:, 0:y_start])
    y_out = pd.DataFrame(df_outliers_xy.iloc[:, y_start:])

    return x_data, y_data, pcs_cleaned_x, pcs_cleaned_y, pcs_out_x, pcs_out_y, n_outliers_x,
n_outliers_y, x_out, y_out

```

ANEXO F. Licença MIT do software desenvolvido

MIT
License

Copyright (c) 2019 floresdwm

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.