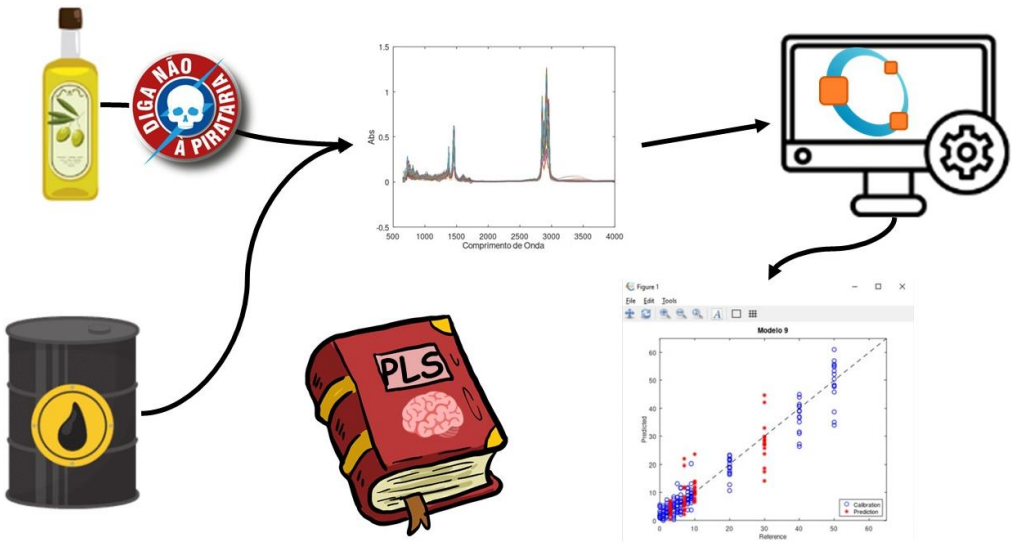
ISSN 2359-4799
Volume 10 / Número 4 / Ano 2024 – p. 01-16
DOI: 10.36524/ric.v10i4.2535

GRAPHICAL ABSTRACT



Tutorial para aplicação didática de quimiometria em software gratuito – Parte II: Regressão por Mínimos Quadrados Parciais (PLS) em dados de infravermelho médio e próximo para determinação de teor de adulterantes e propriedades físico-químicas.

Tutorial for didactic application of chemometrics in free software – Part II: Partial Least Squares (PLS) regression on mid- and near-infrared data to determine adulterant content and physicochemical properties.

Pedro Henrique Pereira da Cunha[1]*, Gabriely Silveira Folli[1], Sara Joaquina Inocencio Dionisio[1], Amanda Guedes Caldeira[1] e Paulo Roberto Filgueiras[1]

[1]Department of Chemistry, Center of Exact Sciences, Federal University of Espírito Santo – UFES, Goiabeiras, Vitória –ES, Brazil, Zip Code: 29075-910.

*pedrophenrique@hotmail.com

Artigo submetido em 15/12/2023, aceito em 27/03/2024 e publicado em 03/06/2024.

ORCID – Pedro H. P. da Cunha: https://orcid.org/0000-0003-1850-4664
ORCID – Gabriely S. Folli https://orcid.org/0000-0003-0665-7540
ORCID – Paulo R. Filgueiras: https://orcid.org/0000-0003-2617-1601

Resumo: O constante avanço tecnológico, instrumental e o aumento na capacidade de gerar dados resultou em um novo desafio para os químicos: como lidar com conjuntos de dados complexos, extensos e multidimensionais. Neste cenário, surgiu a quimiometria, um ramo da química especializado na aplicação de técnicas estatísticas e matemáticas para análise de dados analíticos de origem multivariada, resultando em conclusões científicas mais precisas e confiáveis. Por ser uma área relativamente nova, ainda há uma carência de recursos educacionais sobre o assunto e as abordagens quimiométricas estão, majoritariamente, confinadas a cursos de pós-graduação. Destarte, este artigo oferece um tutorial que pode ser usado em diversas instâncias acadêmicas – porém com o foco na graduação – para ensinar a regressão por mínimos quadrados parciais (PLS, do inglês *Partial Least Squares*). Considerado um dos métodos mais essenciais e comuns da quimiometria, destaca-se pela sua capacidade de lidar eficientemente com conjuntos de dados complexos e altamente correlacionados, cuja importância é evidente na análise de regressão quando há multicolinearidade significativa entre as variáveis independentes. Além disso, o PLS é conhecido por sua facilidade de uso, tornando-se uma ferramenta valiosa para profissionais que buscam uma abordagem eficaz na modelagem estatística complexa. Detalhamos todas as etapas para criar um modelo completo de PLS, desde a instalação do software gratuito GNU Octave até a produção das figuras finais. Também fornecemos todos os algoritmos desenvolvidos para a leitura e tratamento dos dados. Por fim, interpretamos os resultados obtidos, de forma que as discussões e conclusões sejam facilmente compreendidas por todos.

Palavras-chave: Quimiometria; PLS; tutorial; Matlab; Octave.

Abstract: The constant advance in technology, instrumentation, and the increase in data generation capacity has resulted in a new challenge for chemists: how to deal with complex, extensive, and multidimensional data sets. In this scenario, chemometrics emerged as a branch of chemistry specialized in the application of statistical and mathematical techniques for the analysis of multivariate analytical data, resulting in more precise and reliable scientific conclusions. As it is a relatively new area, there is still a lack of educational resources on the subject, and chemometric approaches are mostly confined to postgraduate courses. Therefore, this article provides a tutorial that can be used in various academic instances, with a focus on undergraduate education, to teach partial least squares regression (PLS). PLS is considered one of the most essential and common methods in chemometrics, and stands out for its ability to efficiently handle complex and highly correlated data sets, which is particularly important in regression analysis when there is significant multicollinearity among independent variables. In addition, PLS is known for its ease of use, making it a valuable tool for professionals seeking an effective approach to complex statistical modeling. We detail all the steps to create a complete PLS model, from installing the free GNU Octave software to producing the final figures. We also provide all the algorithms developed for data reading and processing. Finally, we interpret the obtained results so that the discussions and conclusions are easily understood by everyone.

Keywords: Chemometrics; PLS; tutorial; Matlab; Octave.

1 INTRODUÇÃO (títulos e subtítulos ficam a critério do(s) autor(es) – apenas quando houver)

Os avanços tecnológicos e o surgimento de computadores capazes de realizar cálculos matemáticos e estatísticos complexos e extensos simplificou a interpretação e compreensão de informações resultantes. Paralelamente, as técnicas de análise química foram modernizadas, resultando em uma vasta geração de dados. Esses processos resultaram em uma área da química moderna: a Quimiometria. Sua finalidade é processar conjuntos de dados complexos, transformando-os em informação útil através da estatística e matemática.[1]

A Quimiometria apresenta três subáreas: planejamento de experimentos, métodos não supervisionados e os métodos supervisionados. O planejamento de experimentos atua na otimização de um sistema experimental a partir da combinação reduzida de experimentos, buscando a eficiência na obtenção de dados. Os métodos não supervisionados abordam a análise exploratória de dados, identificando padrões e estruturas subjacentes sem depender de características prévias, enquanto os métodos supervisionados envolvem a construção de modelos preditivos baseados em vetor contendo alguma característica amostral, que pode ser qualitativa (classificação) ou quantitativa (regressão). Essas subáreas desempenham papéis cruciais na interpretação e extração de informações significativas de conjuntos de dados químicos complexos, contribuindo para avanços significativos em diversas áreas científicas e industriais.[2] Isso permite identificar analitos e simplificar resultados complexos e multivariados.[1,3–6]

1.1.DESENVOLVIMENTO NO BRASIL

A Quimiometria é uma área da química relativamente nova e com um começo incerto. Teve início formalmente na primeira metade da década de 70, quando a Análise por Componentes Principais (PCA do inglês "*Principal Component Analysis*") já proposta por Pearson[7] em 1901 e desenvolvida por Hotelling[8] 30 anos depois passou a ser usada como método de classificação. Ademais, também tem sido amplamente aplicada através do algoritmo NIPALS (do inglês, *Nonlinear Iterative Partial Least Squares*). Todavia, a Quimiometria se firmou definitivamente apenas com a chegada do computador ao laboratório químico.[39]

No final da década de 70, um pequeno grupo de quimiometristas começou a se formar na Universidade Estadual de Campinas (UNICAMP). O processamento de cálculos atualmente considerados simples, demoravam-se dias, a programação era feita em fitas magnéticas e o número de computadores era limitado. Uma melhora ocorreu quando os microcomputadores chegaram ao Brasil, no qual, a UNICAMP foi uma das pioneiras em obter estes aparelhos. Diante disso, o grupo de quimiometria com a posse do software ARTHUR – primeiro programa de quimiometria do mundo – adaptou as sub-rotinas para desenvolver trabalhos com a empresa Rio Doce Geologia e Mineração S.A. (Docegeo).[3] O programa ARTHUR não era facilmente adaptável em outros computadores. Isso gerou um obstáculo na divulgação da quimiometria em território nacional. A divulgação começou a se espalhar no ano de 1985 com a chegada dos computadores de 16 bits e o interesse da indústria em aplicar técnicas quimiométricas, trazendo grande incentivo e verba para o grupo da UNICAMP.

1.2.QUIMIOMETRIA NA ACADEMIA

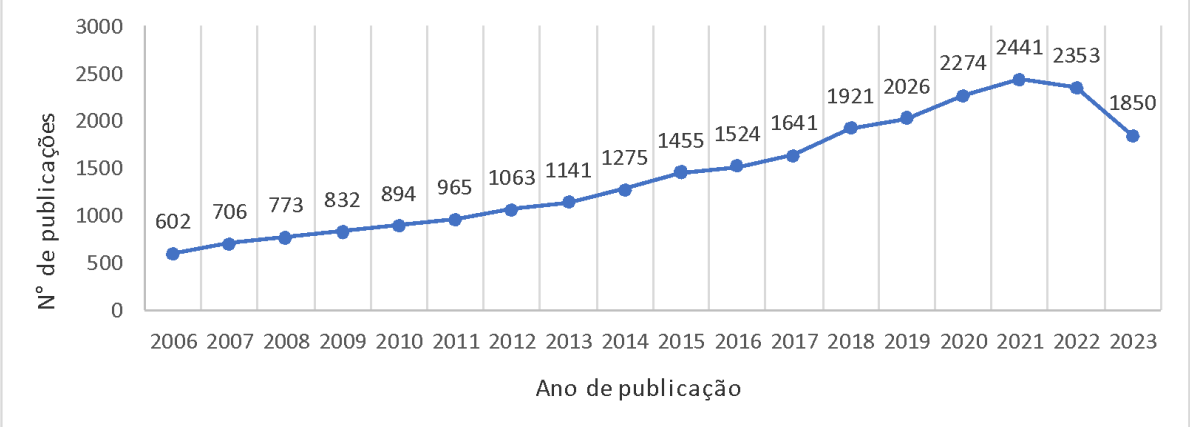
Uma pesquisa em 13 de dezembro de 2023, usando a palavra-chave "*chemometric**" na base de dados *Web of Science*, mapeou o progresso desse segmento da química entre 2006 e 2023. Foi evidenciado um crescimento significativo

de 422%, conforme ilustrado na **Figura 1**. Dito disso, a Quimiometria tem se destacado significativamente na academia, tanto nacional quanto globalmente. Este fenômeno é claramente evidenciado pelo **Figura 2**, que apresenta a análise da quantidade de artigos catalogados no *Web of Science* durante o mesmo período previamente abordado ao empregar a mesma palavra-chave na busca.

Figura 1 – Publicações relacionadas a quimiometria no Brasil por ano.

Fonte: Web of Science

Figura 2 – Publicações relacionadas a quimiometria no mundo por ano.



Fonte: Web of Science.

Se comparados os dois gráficos, observa-se que o progresso desse seguimento da química no país é evidente e significativo, apresentando um crescimento em torno de 422% de 2006 a 2023. Nota-se ainda que o crescimento nacional, em percentual, é bem superior ao crescimento mundial que foi igual a 192%, isso demonstra não só que a Quimiometria vem crescendo no Brasil, mas também sua contribuição em nível mundial.

Também foi analisado as principais faculdades brasileiras que publicam sobre Quimiometria, utilizando os mesmos parâmetros, no período de 2006 a atualmente. Na Tabela 1 estão os resultados obtidos.

Tabela 1. Principais instituições que publicam artigos na Quimiometria.

Posição	Instituição	N° de Publicações
1	UNICAMP	395
2	USP	243
3	EMBRAPA	192
4	UFSCar	149
5	UTFPR	145
6	Unesp	126
7	UFMG	124
8	UFRGs	121
9	UFES	96
10	UFBA	95
10	UFPB	95

Fonte: Web of science.

Percebe-se que apesar da predominância de faculdades públicas, nota-se o destaque da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), uma empresa pública voltada para pesquisa agrícola. Além disso, a UNICAMP se sobressai como precursora da Quimiometria no Brasil.

Para que ocorra a manutenção da taxa de crescimento apresentada pelo Brasil nesta área, bem como um impulsionamento da utilização desta ferramenta, é necessário que os químicos analíticos dominem e apliquem a Quimiometria. Para isso, é preciso que se rompa as barreiras impostas pelas limitações no ensino de Quimiometria seja por falta de conhecimento teórico ou por falta de habilidades no uso de programas específicos como o Octave e Matlab.

Um problema que pode estar associado a este descompasso é uma falta de materiais didáticos com linguagem e abordagem adequadas para cursos introdutórios, fazendo-se necessário o desenvolvimento de tutoriais e métodos com objetivo de encurtar o caminho do estudante de química à quimiometria, gerando químicos capacitados a aplicar essas técnicas no campo científico e industrial.

Destarte, este trabalho busca dar continuidade ao trabalho de Folli *et al*., 2023,[10] desenvolvendo um material simples, de fácil compreensão e completo, da teoria à prática. Fornece o material para aplicação passo a passo de PLS usando o Software Octave, também aplicável ao Matlab. O conhecimento teórico e prático adquirido do PLS, bem como as funções e rotinas necessárias, pode auxiliar em várias pesquisas e análises químicas, tanto acadêmicas quanto industriais. O principal intuito é contribuir com a divulgação da Quimiometria no Brasil.

2 REFERENCIAL TEÓRICO

Derivada da calibração univariada, a regressão multivariada é uma das grandes áreas da quimiometria e é assim denominada pois considera o comportamento de duas ou mais variáveis simultaneamente.[11] Difere de sua antecessora pois, a calibração univariada, é uma metodologia simples, fácil de aplicar e precisa somente de uma única variável.[12] Contudo, este método antigo depende que essa única variável seja completamente relacionada a variável dependente e que não sofra interferência de outras moléculas químicas.[13] Quando falamos de misturas complexas – como petróleo, azeite, café, cerveja, dentre outros[14–17] – isso é praticamente impossível de ocorrer, necessitando assim da utilização de uma calibração que utilize mais variáveis, chamada de regressão multivariada.[18,19]

2.1. PLS

A Regressão pelo método dos mínimos quadrados parciais (PLS, do inglês "*Partial Least Squares*") é um dos métodos mais utilizados da quimiometria, devido sua alta generalização e facilidade em resolver problemas de regressão. Diferente do PCR, o PLS decompõe simultaneamente a variável dependente (X) e independente (y) obtendo o que chamamos de variável latente (LV, do inglês *latent variable*), isso

proporciona uma vantagem para o método em relação ao método PCR. A matemática do PLS começa com o seguinte conjunto de equações:

$$\boldsymbol{X = T_A P_A^T + E} \quad \text{Equação 1}$$

$$\boldsymbol{y = U_A q_a^T + f} \quad \text{Equação 2}$$

Onde **T** e **U** são os escores do modelo, **P** e **q**, os *loadings* obtidos, **E** e **f** os resíduos e **A** o número de variáveis latentes. A matriz **T** é determinada com a combinação linear da matriz **X** com coeficientes ponderados por **W**, chamados de peso, assim como é demonstrado na equação 2.[20]

$$\boldsymbol{TA = X.WA} \quad \text{Equação 3}$$

Com o W, os coeficientes de regressão do PLS podem ser calculados utilizando a Equação 3.

$$\boldsymbol{b_{PLS} = W_A(P_A^T W_A)^{-1}\hat{q}\hat{q}} \quad \text{Equação 4}$$

Com o coeficiente de regressão do PLS podemos resumir a modelagem da seguinte maneira, Equação 4, onde as variáveis independentes são preditas com base na variável dependente.

$$\boldsymbol{y_{pred} = X.\, b_{PLS}} \quad \text{Equação 5}$$

Lembrando que a variável latente deve ser otimizada.[21]

3 MATERIAIS E MÉTODOS

A Parte II do tutorial para desenvolvimento de modelo PLS foi construída em sequência ao trabalho realizado por Folli e colaboradores. Com isso, é aconselhável ler a Parte I (rotina para desenvolvimento de modelos PCA) desse tutorial antes de desenvolver a Parte II para introdução dos conceitos iniciais de quimiometria.

Os dados para a Parte II foram divididos em duas etapas: dados para quantificação do °API em petróleo utilizando MIR e dados para quantificar diferentes óleos vegetais em azeite de oliva extra-virgem utilizando NIR portátil. Todas as rotinas, funções e os espectros e propriedades físico-química estão disponíveis no GitHub dos autores (https://github.com/Quimiometria-UFES/Tutorial-de-Quimiometria).

3.1. QUANTIFICAÇÃO DE API EM PETRÓLEO.

119 amostras de petróleo foram submetidas à aquisição espectral por espectroscopia de infravermelho na região do médio (MIR, do inglês *Midinfrared spectroscopy*). As condições do MIR foram dadas por uma célula horizontal de reflectância total atenuada (ATR) de seleneto de zinco (ZnSe) e ângulo de incidência de 45° (Pike Technologies, EUA). Cada amostra foi analisada em 32 varreduras com uma resolução óptica de 4 cm^{-1} na região de 4000 a 650 cm^{-1}. O número total de variáveis foi igual a 3350.

Essas amostras também foram submetidas a análises físico-químicas para determinação do grau API, variando entre petróleos extra-pesados (11,4) até leves (57,5), a partir da medida instrumental da densidade do óleo. A densidade do petróleo foi mensurada de acordo com a ASTM D-7042.[22]

3.2. QUANTIFICAÇÃO DE ÓLEO VEGETAL EM AZEITE DE OLIVA EXTRA-VIRGEM

Houve a criação de 229 amostras com diferentes concentrações (0,0, 1,0, 2,0, 3,0, 4,0, 5,0, 6,0, 7,0, 8,0, 9,0, 10,0, 20,0, 30,0, 40,0 e 50,0%) em massas para determinação da porcentagem m/m de óleo vegetal em amostras de azeite de oliva extra-virgem. Cinco óleos vegetais foram utilizados como adulterantes (óleos de soja, canola, algodão, milho e girassol). Os detalhes da construção dessas amostras podem ser encontrados no trabalho de Folli, 2022 e colaboradores.[23]

As amostras criadas foram submetidas à aquisição espectral por aparelho portátil de infravermelho na região do próximo (NIR, do inglês *near infrared spectroscopy*). O NIR portátil utilizado foi da marca MicroNir®, (Pro 1700 com *software* versão 3.0 da Viavi Solutions Inc.)[24,25]. As condições para a aquisição dos espectros foram: distribuição espectral de 1676 a 908 nm, tempo de integração de 8 ms, 100 varreduras por amostra, 125 pontos de aquisição por varredura e resolução de 6,15 nm entre pontos.

3.3. SOFTWARES E ALGORÍTMOS

Os algoritmos foram construídos no *software* GNU Octave (John W. Eaton, versão 7.1.0, 2022) e pode ser baixado gratuitamente (https://www.gnu.org/software/octave/). Paralelamente, utilizou-se o *software* Matlab (Versão 2013Ra) para verificar a assertividade dos resultados obtidos pelo Octave. A rotina para construção do modelo PLS foi nomeado “02 - Tutorial PLS”.

4 RESULTADOS E DISCUSSÃO

4.1. CONHECENDO O PLSMODEL – MODELOS PLA PARA PETRÓLEO

O ideal é que o primeiro passo no início de uma rotina seja sempre os seguintes comados:

```
>> clear       % Limpa o Workspace.
>> clc         % Limpa o Command Window.
>> close all   % Fecha qualquer imagem aberta.
```

Em seguida, exclusivamente para o uso do Octave, é necessário ativar os pacotes necessários.

```
>> pkg load statistics
>> pkg load io
```

É necessário ainda mudar o Diretório, o endereço do programa para local onde estão os dados que usaremos na rotina, através dos comandos:

```
>> cd('...\IFES Ciencia\PLS_Model')
```

Direciona o *software* para a pasta desejada.

```
>> load('Dados_API.mat')

% Note que os dados nos temos;

% Xmir      % Espectro MIR das nossas amostras.

% y         % São os dados quantitativos.

% objetos   % E a separação de calibração e teste das amostras separados por Kennard-Stone. [26,27]

% Nmir      % Comprimento de onda do espectro MIR.
```

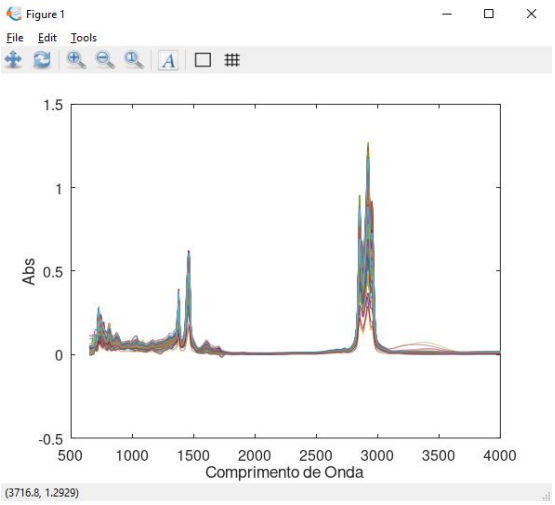
Exporta o arquivo que contém o conjunto amostral

Nos dados exportados tem-se quatro arquivos diferentes, são eles, ‘Xmir’, 119 amostras de espectro MIR, ‘y’, densidade API das 119 amostras, ‘Nmir’, comprimento de ondas do espectro MIR e ‘objetos’ um arquivo que contém a separação das amostras. Para visualizar o espectro de todas as amostras, pode-se utilizar os seguintes comandos:

```
>> plot(Nmir,Xmir)
>> xlabel("Comprimento de Onda")
>> ylabel("Abs")
>> set(gca,'FontSize',16)
```

Os comandos abrirão a seguinte subjanela, **Figura 3**:

Figura 3. Espectro das amostras de densidade API.



Fonte: Autoria Própria.

Para desenvolver um modelo de PLS, é preciso separar as amostras em conjunto calibração e teste, para tal, usa-se os comandos:

% Assim, teremos que separar os dados entre "calibração" e "teste" usando as seguintes linhas de comando.
>> Xcal = Xmir(objetos.cal,:); ycal = y(objetos.cal,:);

>> Xtest = Xmir(objetos.test,:); ytest = y(objetos.test,:);

Nesta função do PLS nós temos que criar um arquivo chamado "*options*" para mandar um conjunto de instruções de como o modelo deve ser criado.

>> options = [];

Garante que o arquivo esteja vazio.

>> options.Xpretreat = {'center'};

Define o pré-tratamento interno.

>> options.vene = 5;

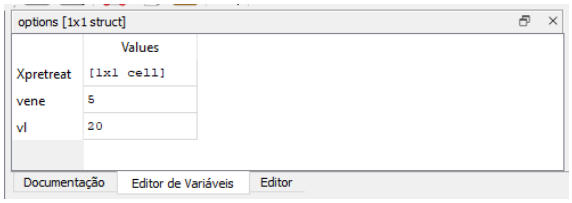
Tamanho da janela de calibração cruzada.

>> options.vl = 20;

Número de variáveis latentes.

Caso queira utilizar nenhum o pré-tratamento interno utilize "{'*none*'}". Para verificar se o "*options*" está programado da forma desejada, em "Ambiente de Trabalho", duplo clique no "*options*" transferirá a tela principal para a subjanela "Editor de Variáveis", '*Variables*' no Matlab, e poderá visualizar dentro do arquivo, como apresentado na **Figura 4**.

Figura 4. Editor de Variáveis do Octave.



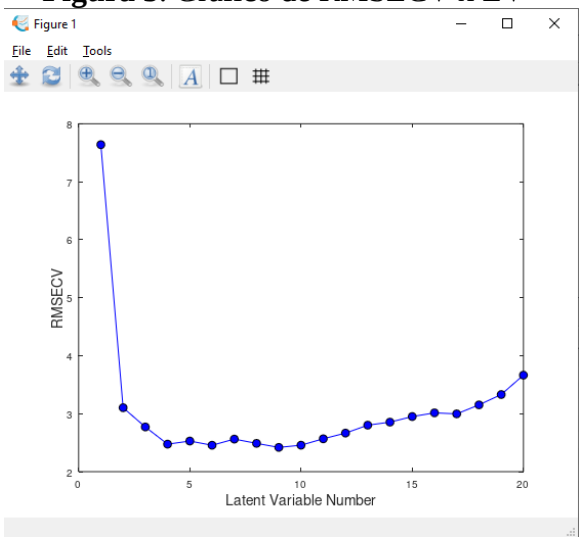
Fonte: Autoria Própria.

Pode-se rodar o PLS no modo otimização, através do comando abaixo:

>> modelo=plsmodel(Xcal,ycal,options);

Sob este comando o *software* fara os cálculos de otimização e abrirá a seguinte imagem, **Figura 5**.

Figura 5. Gráfico de RMSECV x LV



Fonte: Autoria Própria.

Neste gráfico podemos verificar como o RMSECV varia conforme o número de LV cresce. Neste caso, o número de LV

ideal parece ser 4, devido ao RMSECV variar pouco depois deste ponto.

Quando utilizamos uma função temos *inputs*, que são arquivos de entrada, e *outputs*, que são arquivos de saída, analogicamente a reações químicas onde os *inputs* seriam reagentes, os *outputs* produtos. Na última linha de comando utilizada, a que criou o gráfico acima, nós usamos a função "*plsmodel*" que tem três inputs, "**Xcal**", "**ycal**" e "*options*" e tem como output "modelo". Nesse modo, entregamos os *inputs* para a função e ela retorna os *outputs*.

Escolhendo o número 4 de LV, utilizamos os seguintes comandos:

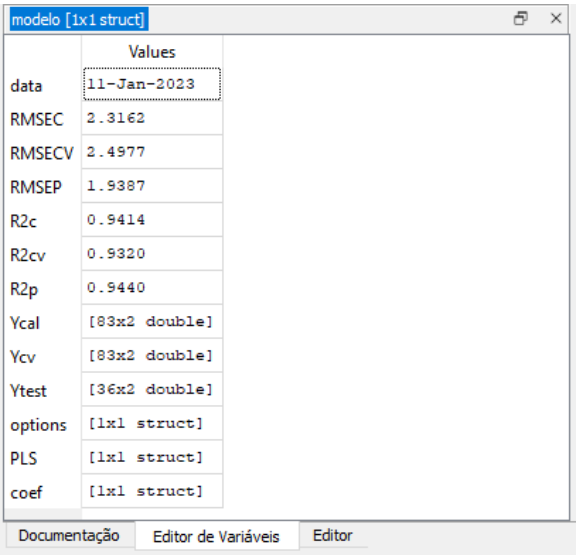
```
>> options.vl = 4;

>> modelo =
plsmodel(Xcal,ycal,Xtest,ytest,options);
```

Neste caso a função irá criar um modelo de PLS com ambos os conjuntos (Xcal e ycal). Após finalizar o cálculo, duplo clique no *input* "modelo".

A **Figura 6** mostra a estrutura do modelo que é composta pelos parâmetros de avaliação, dentre eles destaca-se:

- RMSEC (Erro quadrático médio de calibração):Erro das amostras de calibração.
- RMSEP (Erro quadrático médio de previsão): Erro das amostras de teste.
- R^2c (Coeficiente de determinação das amostras de calibração): quanto mais próximo a 1, melhor o modelo.
- R^2p (Coeficiente de determinação das amostras de teste).

Figura 6. Estrutura interna no modelo, no Octave.



Fonte: Autoria Própria.

Analisando o coeficiente de determinação, julgaríamos o modelo como bem-sucedido, entretanto, segundo a literatura, o RMSEP ideal é próximo ao 1.2 e o R^2c acima de 0.95. Testaremos o modelo com 5 LV:

```
>> options.vl  = 5;

>> modelo =
plsmodel(Xcal,ycal,Xtest,ytest,options);

>> modelo.R2c;   % 0.9537

>> modelo.R2p    % 0.9500
```

Também testaremos o modelo com 6LV:

```
>> options.vl          = 6;

>> modelo =
plsmodel(Xcal,ycal,Xtest,ytest,options);

>> modelo.RMSEC; % 1.9061

>> modelo.RMSEP; % 1.4480

>> modelo.R2c;   % 0.9614

>> modelo.R2p;   % 0.9582
```

Com base nos parâmetros de avalição sugere-se que o modelo o com seis variáveis latentes é melhor, com RMSE's menores e ambos R^2 mais próximos de 1, entretanto, é preciso comprovar com alguns testes. Primeiro, fazemos os dois modelos que serão comparados.

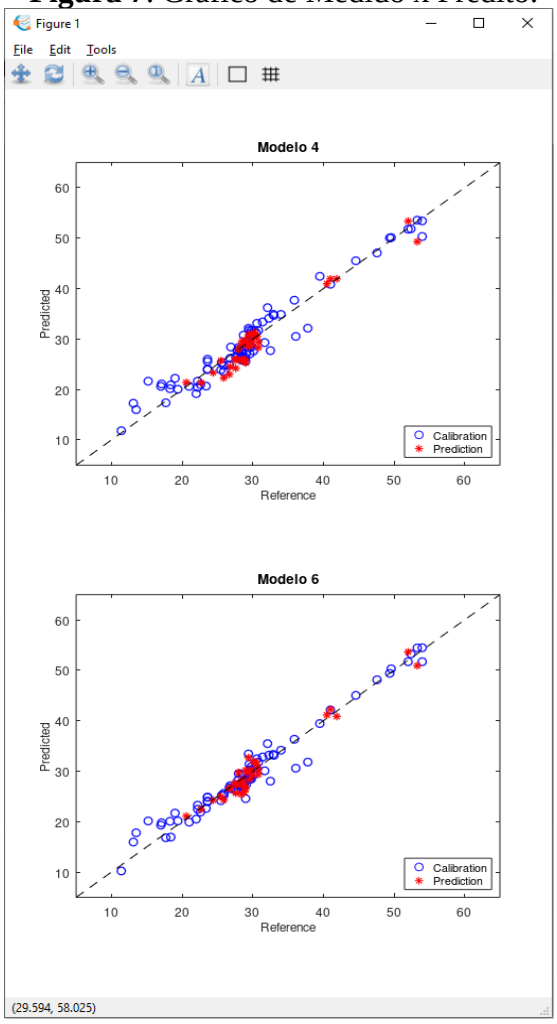
```
>> options.vl  = 4;

>> modelo4 =
plsmodel(Xcal,ycal,Xtest,ytest,options);

>> options.vl  = 6;

>> modelo6 =
plsmodel(Xcal,ycal,Xtest,ytest,options);
```

Em seguida vamos avaliar os modelos em conjunto utilizado o "Gráfico de Medido x Predito", para tal usaremos os seguintes comandos:

```
>> close all %Fechando imagens já criadas.

>> subplot(2,1,1)

>>plot(modelo4.Ycal(:,1),modelo4.Ycal(:,
2),'bo','LineWidth',1); hold on;

>>plot(modelo4.Ytest(:,1),modelo4.Ytest(:
,2),'r*','LineWidth',1); hold on;

>> ylim([5 65]); xlim([5 65]);

>>  plot(xlim,  ylim,  '--
k');legend('Calibration','Prediction','Locatio
n','southeast');

>> title('Modelo 4');

>>set(gca,'FontSize',12);xlabel('Reference'
,'fontsize',12);

>> ylabel('Predicted','fontsize',12);

>> subplot(2,1,2)

>>plot(modelo6.Ycal(:,1),modelo6.Ycal(:,
2),'bo','LineWidth',1); hold on;

>>plot(modelo6.Ytest(:,1),modelo6.Ytest(:
,2),'r*','LineWidth',1); hold on;

>> ylim([5 65]); xlim([5 65]);

>>  plot(xlim,  ylim,  '--
k');legend('Calibration','Prediction','Locatio
n','southeast');

>> title('Modelo 6');

>>set(gca,'FontSize',12);xlabel('Reference'
,'fontsize',12);

>> ylabel('Predicted','fontsize',12);
```

Como resultado obtém-se a **Figura 7**.

Figura 7. Gráfico de Medido x Predito.



Fonte: Autoria Própria.

Trata-se de uma análise estatística, o gráfico apresenta no eixo X os valores

medidos pela técnica e no eixo Y o que foi previsto pelo modelo, nesse modo podemos identificar tendências, grau de concordância e suspeita de *Outlier*. Na **Figura 7** podemos perceber que as amostras, em sua maioria, ficaram bem próximas da linha de tendência, o que é um bom indicador, repara-se ainda que ambos modelos tiveram o mesmo comportamento na faixa 5 a 37, e o modelo 6 obteve uma melhor performance após essa faixa. Mas ainda não é o suficiente, para confirmar que o modelo 6 é melhor, vamos utilizar uma comparação entre modelos com teste randômico de exatidão, "accuracy_test", utilizando os seguintes comandos:

```
>> tic

>> [pvalue,dist_tt,meandiff] =
accuracy_test(ytest,modelo4.Ytest(:,2),mo
delo6.Ytest(:,2),'randbi',500000,0.05);

>> toc
```

A função tic/toc informa o tempo, em segundos, entre o comando tic e o toc, interessante para saber o tempo de processamento de um teste. A função "*accuracy_test*" foi programada para dar a resposta assim que termina os cálculos, neste caso "Modelos com DIFERENÇAS na acurácia" os modelos são diferentes estatisticamente e podemos afirmar que o modelo 6 é melhor. Caso queira saber mais sobre a função que usamos, e outras, você pode utilizar a função "*help*" seguida do nome da função, como abaixo:

```
>> help accuracy_test
```

Recomenda-se que salve o "Ambiente de Trabalho" utilizando o seguinte comando;

```
>> save("Tutorial 02-1");
```

Esse comando salva todo o "Ambiente de Trabalho", o primeiro *input* que fica entre aspas é o nome do arquivo que será salvo, no caso do nosso exemplo "Parte 01".

4.2. EXEMPLO REAL – AZEITE DE OLIVA.

Esta parte detalhará o dia-a-dia de um Quimiometrista, mostrando análise de amostras reais, separação de conjuntos, melhor pré-tratamento, verificação outlier e determinação do melhor modelo. Iniciaremos com a extração de dados numa planilha.

```
>> clear all;
>> clc;
>> close all;
>>cd('C:\Users\Exemplo\...\Tutorial\PLS_
Model');

>> % Como extrair dados de planilha excel.

>>[y,~,~]=xlsread('Oleos_Adulterados.xls
x','Plan1','B2:B230'); % Vetor de regressão
```

A função xlsread, trata-se de um leitor de planilhas e a podemos simplificar no seguinte esquema:

```
% [A,B,C]=xlsread('XXX','YYY','ZZZ')
% XXX = Nome da planilha, incluindo o
formato.
% YYY = Aba da planilha.
% ZZZ = Faixa que deixar extrai.
% A = Quanto se captura números.
% B = Quando se captura caracteres.
% C = Quando deseja capturar número e
letras.
```

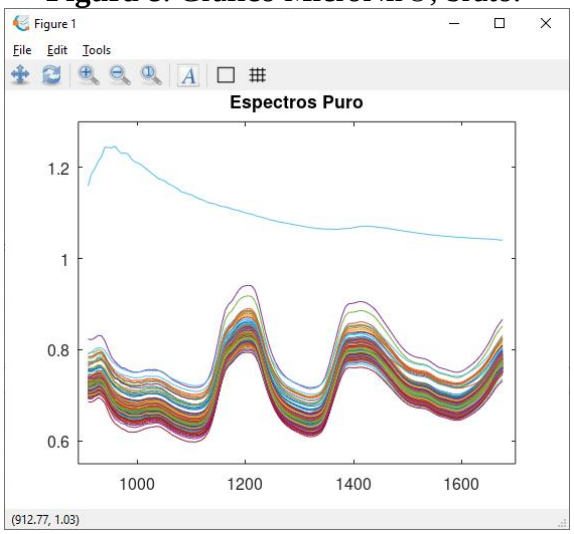
Assim, continuamos a extração dos dados.

```
>>[num,~,~]=xlsread('Oleos_Adulterados.
xlsx','Plan1','C1:DW1');

>>[X,~,~]=xlsread('Oleos_Adulterados.xls
x','Plan1','C2:DW230');

>>[~,Sample,~]=xlsread('Oleos_Adulterad
os.xlsx','Plan1','A2:A230');
```

Esses espectros são de infravermelho próximo, adquiridos por meio do MicroNir®, de azeite adulterado com óleo comercial, utilizadas em um artigo publicado e pegas com autorização dos autores,[16] além disso, por medida de segurança, as amostras comerciais foram renomeadas por códigos. A primeira coisa que devemos fazer é analisar o perfil espectral, verificar se temos alguma amostra inconsistentes, se os espectros são semelhantes aos encontrados na literatura. Faremos isso através do seguinte comando:

```
>> plot(num,X);
>> title('Espectros Puro')
>> set(gca,'FontSize',16);
>> xlim([890 1700]);ylim([0.55 1.3]);
```

Figura 8. Gráfico MicroNir®, bruto.



Fonte: Autoria Própria.

Nota-se pelo gráfico das amostras, **Figura 8**, que se tem uma amostra com um perfil anômalo, só isso já é o suficiente para podermos removê-la como *outlier* pela função *find*;

```
>> AAA = find(X(:,1) > 1);
>> X(AAA,:) = []; y(AAA,:) = [];
```

A função '*find*' tem o objetivo de encontrar uma amostra em uma condição especifica, no caso estamos procurando uma amostras no 'X(:,1)', e a condição foi definida ao observar o espectro.

Como se trata de dados de um artigo iremos utilizar a mesma separação cal/test dele, com amostras de 3, 7, 10 e 30% pertencentes ao grupo teste. Para isso iremos combinar for e if. Para melhor compreensão do for, é recomendado que leia o "00 - Conhecendo a função for".

Primeiro criamos a variável "objetos" para armazenar onde cada amostra estará.

```
>> objetos.cal  = [];
>> objetos.test = [];
```

Agora iremos usar um for, uma função cíclica, que o "qi" irá variar de 1 até a quantidade de linhas de X, ou seja, número de amostras.

```
>> for qi=1:1:size(X,1);
```

Se y tem valor 3 7 10 e 30;

```
>> if y(qi) == 3 || y(qi) == 7 || y(qi) == 10
||y(qi) == 30;
```

Caso o y(qi) obedeca a condição.

```
>> objetos.test = [objetos.test;qi];
```

else

Caso o y(qi) não obedeca a condicao.

```
>> objetos.cal = [objetos.cal;qi];
>> end
```

Separando as amostras em conjunto calibração e teste.

```
>> Xcal = X(objetos.cal,:); Xtest =
X(objetos.test,:);
>> ycal = y(objetos.cal,:); ytest =
y(objetos.test,:);
```

```
>> close all
>> plot(1:1:size(ycal,1),ycal,'bo'); hold on;
>> plot(1:1:size(ytest,1),ytest,'r*'); hold on;
```

Note que esta separação segue todos pré-requisitos citados. Agora vamos para a modelagem.

```
>> close all
>> options=[];
>> options.vene          = 5;
>> options.vl            = 20;
```

```
>> modelo=plsmodel(Xcal,ycal,options)

>> options.vl          = 9;
>>                modelo9          =
plsmodel(Xcal,ycal,Xtest,ytest,options);
>> options.vl          = 15;
>>                modelo15         =
plsmodel(Xcal,ycal,Xtest,ytest,options);
```

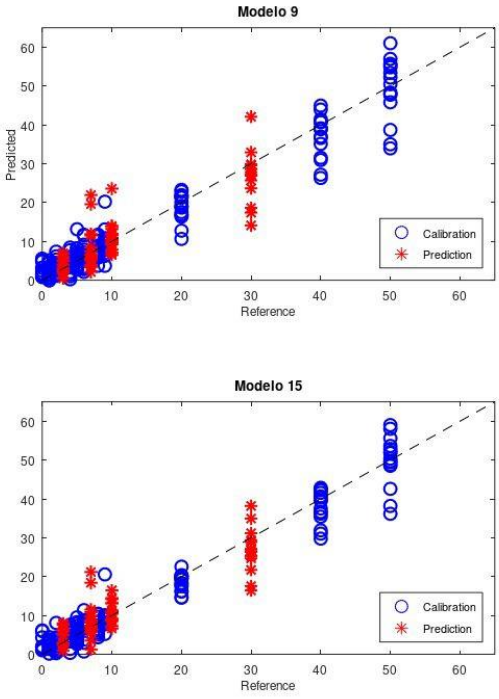
%VL	9	15
%RMSEC	4.1114	3.4256
%RMSEP	5.0965	4.6783
%R2c	0.9377	0.9584
%R2p	0.7597	0.7937

O VL melhor foi o 15 como podemos analisar com base nos parâmetros de avaliação, todavia, vamos avaliar o gráfico de medido e previsto.

Agora, vamos analisar o gráfico de medido e predito das amostras;

```
>> subplot(2,1,1);

subplot(2,1,1);

plot(modelo9.Ycal(:,1),modelo9.Ycal(:,2),'
bo','LineWidth',1); hold on;

plot(modelo9.Ytest(:,1),modelo9.Ytest(:,2)
,'r*','LineWidth',1); hold on;

ylim([0 65]); xlim([0 65]);

plot(xlim,          ylim,          '--
k');legend('Calibration','Prediction','Locatio
n','southeast');

title('Modelo 9');

set(gca,'FontSize',12);xlabel('Reference','fo
ntsize',12);

ylabel('Predicted','fontsize',12);



>> subplot(2,1,2);

subplot(2,1,2);

plot(modelo15.Ycal(:,1),modelo15.Ycal(:,
2),'bo','LineWidth',1); hold on;

plot(modelo15.Ytest(:,1),modelo15.Ytest(:,
2),'r*','LineWidth',1); hold on;

ylim([0 65]); xlim([0 65]);

plot(xlim,          ylim,          '--
k');legend('Calibration','Prediction','Locatio
n','southeast');

title('Modelo 15');

set(gca,'FontSize',12);xlabel('Reference','fo
ntsize',12);
```

Figura 9. Gráfico de medido e predito modelo azeite.



Fonte: Autoria Própria.

Quando analisamos o gráfico de medidos e preditos, **Figura 9**, percebemos que algumas amostras estão consideravelmente afastadas da linha de referência, cinco amostras de teste, na faixa 10, 7 e 30, além de algumas amostras de calibração.

Ao observar o gráfico, nota-se pouca diferença entre os modelos a partir do 20%.

Contudo na faixa de 10% o modelo de VL 15 se mostrou superior.

Agora vamos para o teste de acurácia.

```
>> tic
>> [pvalue,dist_tt,meandiff] = accuracy_test(ytest,modelo9.Ytest(:,2),modelo15.Ytest(:,2),'randbi',500000,0.05);
>> toc  %{1068 sec}
```

O teste considerou os modelos iguais, utilizando o princípio da parcimônia, onde menos é mais. Escolhemos o modelo com VL 9 como o melhor modelo.

4.3. PRÁTICA

Finalizamos este tutorial deixando uma sugestão de exercício para praticar. O objetivo é utilizar o terceiro conjunto de amostras, chamado “Nitrogênio Total”, e desenvolver uma regressão utilizando todos os conhecimentos aplicados no tutorial. Detalhes;

- *Indet*: Identificação das amostras.
- num: Numero de onda do espectro de infravermelho.
- X: Espectro de Infravermelho Médio das amostras. [Fonte Analítica]
- y: Vetor de concentração de Nitrogênio total.

Além disso, os modelos utilizados como exemplo aqui, nos tutoriais, não são os melhores resultados obtidos pela equipe do laboratório, então, se sinta desafiado a tentar encontrar os modelos. O Gabarito se encontra no final do tutorial.

5 CONCLUSÃO

Este tutorial didático para a aplicação do PLS no software gratuito GNU Octave é uma alternativa educacional economicamente acessível, que visa facilitar a ampla disseminação da quimiometria nos mais diversos níveis acadêmicos. As rotinas e funções desenvolvidas no ambiente do GNU Octave possuem grande semelhança com as do software MATLAB (programa mais comumente empregado em trabalhos utilizando a quimiometria). Assim, essa abordagem favorece uma interação mais significativa, robusta e eficaz entre os alunos e professores de química, facilitando o processo de aprendizagem. Ademais, o fato de os dados experimentais estarem disponíveis de forma gratuita e online enriquece ainda mais todo o conteúdo explorado neste trabalho, promovendo uma compreensão integral dele, bem como uma maior divulgação da Quimiometria no Brasil.

AGRADECIMENTOS

Os autores agradecem às empresas de fomento Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES) [88887.487966/2020-00], Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) [032/2023; 356/18;83552723,442/2021, 3530.503.20537.12092017 e 76459934/16], e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [422515/2016-7, 445987/2014-6, 310349/2021-4e 465450/2014-8] por todo suporte financeiro, à Petróleo Brasileiro S.A. (PETROBRAS), Núcleo de Competências em Química do Petróleo (NCQP) e Instituto Federal do Espírito Santo (IFES) pelo fornecimento das amostras, aquisição espectrais e espaço físico para desenvolvimento da pesquisa.

REFERÊNCIAS

1. Bruns, R. E. & Faigle, J. F. G. Quimiometria. *Quimica Nova* vol. 8 84–99 at (1985).
2. Ferreira, M. M. C. *Quimiometria: conceitos, métodos e aplicações*. (Editora da Unicamp, 2015). doi:10.7476/9788526814714.
3. Barros Neto, B. de, Scarminio, I. S. & Bruns, R. E. 25 anos de quimiometria no Brasil. *Quim. Nova*

29, 1401–1406 (2006).
4. Pereira, F. & Pereira-Filho, E. APLICAÇÃO DE PROGRAMA COMPUTACIONAL LIVRE EM PLANEJAMENTO DE EXPERIMENTOS: UM TUTORIAL. *Quim. Nova* **2018**, 1061–1071 (2018).
5. Hebling e Tavares, J. P., da Silva Medeiros, M. L. & Barbin, D. F. Near-infrared techniques for fraud detection in dairy products: A review. *J. Food Sci.* **87**, 1943–1960 (2022).
6. Correia, R. M. *et al.* PORTABLE NEAR INFRARED SPECTROSCOPY APPLIED TO THE QUALITY CONTROL OF COFFEE ADULTERED BY GROUNDS. *Quim. Nova* **45**, 392–402 (2022).
7. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
8. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
9. Agnoletti, B. Z. *et al.* Multivariate calibration applied to study of volatile predictors of arabica coffee quality. *Food Chem.* **367**, 130679 (2022).
10. Silveira Folli, G., Pereira da Cunha, P. H., Kuster Moro, M. & Roberto Filgueiras, P. Tutorial para aplicação didática de quimiometria em software gratuito – Parte I: Análise de Componentes Principais em dados de infravermelho médio e propriedades físico-químicas de amostras de petróleo. *Rev. Ifes Ciência* **9**, 01–14 (2023).
11. de Paulo, E. H. *et al.* Study of coffee sensory attributes by ordered predictors selection applied to 1H NMR spectroscopy. *Microchem. J.* **190**, 108739 (2023).
12. Oliveira, B. G. *et al.* Controlling the quality of grape juice adulterated by apple juice using ESI(-)FT-ICR mass spectrometry. *Microchem. J.* **149**, 104033 (2019).
13. Barros Neto, B. de, Pimentel, M. F. & Araújo, M. C. U. Recomendações para calibração em química analítica: parte I. Fundamentos e calibração com um componente (calibração univariada). *Quim. Nova* **25**, 856–865 (2002).
14. de Paulo, E. H. *et al.* Particle swarm optimization and ordered predictors selection applied in NMR to predict crude oil properties. *Fuel* **279**, 118462 (2020).
15. Coelho Neto, D. M. *et al.* Study of the Chemical Profile of Brazilian Beers: an Assessment Between Artisanal and Industrial Drinks. *Quim. Nova* **45**, 518–530 (2022).
16. Belchior, V., Botelho, B. G. & Franca, A. S. Comparison of Spectroscopy-Based Methods and Chemometrics to Confirm Classification of Specialty Coffees. *Foods* **11**, 1655 (2022).
17. Santos, P. *et al.* DETERMINAÇÃO DA AUTENTICIDADE DE AMOSTRAS DE AZEITE COMERCIAIS APREENDIDAS NO ESTADO DO ESPÍRITO SANTO USANDO UM ESPECTROFOTÔMETRO PORTÁTIL NA REGIÃO DO NIR. *Quim. Nova* **43**, 891–900 (2020).
18. Bagheri Garmarudi, A., Khanmohammadi, M., Ghafoori Fard, H. & de la Guardia, M. Origin based classification of crude oils by infrared spectrometry and chemometrics. *Fuel* **236**, 1093–1099 (2019).
19. Long, J., Wang, K., Yang, M. & Zhong, W. Rapid crude oil analysis using near-infrared reflectance spectroscopy. *Pet. Sci. Technol.* **37**, 354–360 (2019).
20. Salehi, M., Zare, A. & Taheri, A.

Artificial Neural Networks (ANNs) and Partial Least Squares (PLS) Regression in the Quantitative Analysis of Respirable Crystalline Silica by Fourier-Transform Infrared Spectroscopy (FTIR). *Ann. Work Expo. Heal.* **65**, 346–357 (2021).

21. Amsaraj, R., Ambade, N. D. & Mutturi, S. Variable selection coupled to PLS2, ANN and SVM for simultaneous detection of multiple adulterants in milk using spectral data. *Int. Dairy J.* **123**, 105172 (2021).
22. ASTM. Standard D7042: Test Method for Dynamic Viscosity and Density of Liquids by Stabinger Viscometer (and the Calculation of Kinematic Viscosity). *Am. Natl. Stand. Inst.* **12a**, 1–11 (2013).
23. Folli, G. S. *et al.* Food analysis by portable NIR spectrometer. *Food Chem. Adv.* **1**, 100074 (2022).
24. Santos, F. D. *et al.* Discrimination of oils and fuels using a portable NIR spectrometer. *Fuel* **283**, 118854 (2021).
25. Beć, K. B., Grabska, J. & Huck, C. W. Principles and Applications of Miniaturized Near-Infrared (NIR) Spectrometers. *Chem. – A Eur. J.* **27**, 1514–1532 (2021).
26. Kennard, R. W. & Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **11**, 137 (1969).
27. da Cunha, P. H. P. *et al.* Variable selection by permutation applied in support vector regression models. *J. Chemom.* **36**, 1–14 (2022).