

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO
CENTRO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE QUÍMICA

***Tutorial de quimiometria, voltado para o ensino de PLS na
graduação***

Pedro Henrique Pereira da Cunha

Monografia de Conclusão de Curso

Vitória-ES
2023

Pedro Henrique Pereira da Cunha

***Tutorial de quimiometria, voltado para o ensino de PLS na
graduação***

Monografia apresentada ao Departamento de Química-CCE, Universidade Federal do Espírito Santo, como parte dos requisitos para obtenção do título de Licenciatura em Química.

Orientador: Prof. Dr. Paulo Roberto Filgueiras

Vitória-ES
2023

Pedro Henrique Pereira da Cunha

***Tutorial de quimiometria, voltado para o ensino de PLS na
graduação***

Monografia apresentada ao Departamento de Química-CCE, Universidade Federal do Espírito Santo, como parte dos requisitos para obtenção do título de Licenciatura em Química.

Vitória, 03 de Fevereiro de 2023

BANCA EXAMINADORA

Prof. Dr. Paulo Roberto Filgueiras - Orientador

Prof^a. Dr^a. Maria de Fátima Fontes Lelis

Prof^a. Dr^a. Rosângela Cristina Barthus

Vitória-ES
2023

Demorei mais tempo pensando a quem dedicar esse trabalho do que realmente escrevendo a dedicatória. Peço que ninguém se sinta esquecido ou desconsiderado por isso, apenas existem tantas pessoas que me fizeram ser o que sou hoje que fico imaginando que seria um desrespeito com cada uma delas. Então deixo aqui um MEGA OBRIGADO a todas as pessoas que entraram no meu caminho e me fizeram crescer nem que seja um pouco nessa vida, obrigado pais, família, amigos, professores, orientadores e todas as pessoas que contribuíram para meu crescimento. MUITO OBRIGADO.

Agradecimentos

Primeiramente a deus, que me deu fé para acreditar nos meus passos, manter meu foco, mexer meu corpo e acalmar a minha mente.

A minha noiva, Luiza Machado Fischer, que nos momentos de frustração, me acalentou, nos momentos de tristeza, me amou e quando eu não acreditei em mim, acreditou por nós dois, espero, sinceramente retribuir o triplo para você.

A minha família, Alvaro Roque Tosta da Cunha, Leila Vaz Pereira e Isabela Pereira da Cunha, que cuidaram da minha sanidade mental, mesmo sem saber, pois, me nutriam com amor, carinho e orgulho. E alguns puxões de orelha também.

Aos meus amigos, Amalia, Kunu, Cururu, Byeano, Cauani, Yohann, Gian e outros, que acabaram com o meu stress do dia a dia me ouvindo, conversando besteira, bebendo uma gelada e me tirando da inercia social.

Aos meus amigos de laboratório, Marcia Helena, Gabriely Folli, Barbara Zani, Madson Zanoni, Livia e outros, por que rimos nos momentos de desespero e pelos vários ombros amigos que recebi ao longo dos anos.

A família da minha sogra Rita de Cassia, Renato Fischer, Gustavo, Julia, Kamila, Danilo e Juliana, que me propuseram diversas risadas e momentos descontraídos.

Ao meu cachorro, Ralph, que apesar nem ter tanta consciência do mundo complexo em que vivemos, acalmava-me com seu olhar e fofura.

A Fátima Fonte Lelis, por aceitar fazer parte dessa banca e por ter me feito um Químico.

A Rosângela Cristina Barthus, por aceitar fazer parte dessa banca e pela paciência e carinho que sempre teve ao dar aula para mim.

Ao Paulo Roberto Filgueiras, por me aceitar como aluno pesquisador na graduação e pós graduação por mais de cinco anos, por me ensinar a pesquisar e divulgar ciência e pelos puxões de orelhas.

A UFES, CCE e a todos os professores pelos anos de estudos e aprendizados que me foram ofertados de bom grado.

Ao LabPetro por disponibilizar o recurso e espaço necessário para a realização desse projeto.

“Na vida, não existe nada a se temer, apenas a ser compreendido.” - Marie Curie

SUMÁRIO

1. INTRODUÇÃO	12
2. REVISÃO BIBLIOGRÁFICA	14
2.1. Quimiometria.....	14
2.1.1. Desenvolvimento no Brasil	16
2.1.2. Ensino no Brasil	18
2.1.3. Academia no Brasil.....	17
2.2. Quimiometria Aplicação.....	21
2.2.1. Obtenção dos dados	21
2.2.2. Processamento de Dados	22
2.2.3. Avaliação	27
2.3. Regressão Multivariada	29
2.3.1. PLS	30
2.3.2. PLS e suas aplicações	31
2.4. Programação	33
3. OBJETIVOS	35
3.1. Objetivo Geral	35
3.2. Objetivos Específicos	35
4. PROCEDIMENTO EXPERIMENTAL.....	36
4.1. Amostragem.....	36
4.1.1. MIR.....	36
4.1.2. NIR portátil.....	36
4.1.3. Propriedades Físico-químicas.....	36
4.1.4. Azeite	37
4.2. Softwares e Algoritmos	37
5. RESULTADOS E DISCUSSÃO	38
5.1. UFES	38
5.2. Parte 00 - Instalação de pacotes.....	39
5.3. Parte 01 - Conhecendo o plsmodel	41
5.4. Parte 02 - Aprimorando o modelo.	51
5.5. Parte 03 - Exemplo real.	64
5.6. Parte 04 - Exercício para praticar.	72
6. CONCLUSÃO.....	73
7. REFERÊNCIAS	74

LISTA DE FIGURAS

Figura 1. Conhecimentos necessários para ser um Quimiometrista.....	15
Figura 2. Computador PDP-10 utilizado pela Unicamp na década 70.....	16
Figura 3. Gráfico de artigos publicados com a palavra “chemometric*”, pesquisado no Web of Science dia 09 de Janeiro de 2023.	18
Figura 4. Gráfico de artigos publicados, com a palavra “chemometric*”, por ano, mundo (a) e Brasil (b). Pesquisado no Web of Science dia 09 de Janeiro de 2023.....	18
Figura 5. Gráfico de artigos publicados, com a palavra “chemometric*”, por ano e país. Pesquisado no Web of Science dia 09 de Janeiro de 2023.	19
Figura 6. Subjanela Editor no Octave.....	39
Figura 7. Janela de Comando do Octave.....	40
Figura 8. Ambiente de Trabalho do Octave.....	42
Figura 9. Espectro das amostras de densidade API.	43
Figura 10. Editor de Variáveis do Octave.....	44
Figura 11. Gráfico de RMSECV x LV	45
Figura 12. Estrutura interna no modelo, no Octave.....	46
Figura 13. Gráfico de Medido x Predito.....	49
Figura 14. Gráfico bruto e tratado.	52
Figura 15. Gráfico bruto e tratado com derivada.	53
Figura 16. Gráfico de Medido x Predito do modelo 6 e 9.....	57
Figura 17. Gráfico de Leverage.....	59
Figura 18. Gráfico de Resíduos.....	60
Figura 19. Gráfico Lev_res.	61
Figura 20. Gráfico de Lev_res do modelo 6.....	62
Figura 21. Gráfico microNIR bruto.....	66
Figura 22. Separação caltest.	68
Figura 23. Gráfico de medido e predito modelo azeite.	71

LISTA DE TABELAS

Tabela 1. Analise da presença de cursos importantes para a Quimiometria no currículo.	17
Tabela 2. Principais instituições que publicam artigos na quimiometria.....	20
Tabela 3. Principais instituições que publicam artigos na quimiometria por faixa de 5 anos.....	21

LISTA DE ABREVIATURAS E SIGLAS

CARS - Amostragem reponderada adaptativa competitiva (do inglês “*Competitive Adaptive Reweighted Sampling*”)

CENPES - Centro de Pesquisas Leopoldo Américo Miguel de Mello.

Docegeo - Rio Doce Geologia e Mineração S.A. (Docegeo)

EMBRAPA - Empresa Brasileira de Pesquisa Agropecuária.

GA - Algoritmo genético (do inglês “*Genetic Algorithm*”)

HCA - Análise hierárquica de agrupamentos (do inglês “*Hierarchical Cluster Analysis*”)

KS - Kennard-Stone.

LV - Variável latente. (do inglês *Latent Variable*)

MLR - Regressão linear múltipla. (do inglês *Multiple Linear Regression*)

MSC - Correção multiplicativa de sinal (do inglês *Multiplicative Signal Correction*)

NCQP/UFES - Núcleo de Competências em Química do Petróleo da UFES.

NIPALS - (do inglês, ‘*Nonlinear Iterative Partial Least Squares*’)

NIR - Infravermelho próximo. (do inglês *Near Infrared*)

NISPA - Ruído incorporado a análise de janela permutada. (do inglês “*Noise Incorporated Subwindow Permutation Analysis*”)

NMR C¹³ - Ressonância magnética nuclear de carbono. (do inglês “*Nuclear Magnetic Resonance Carbon-13*”)

OM - Octave/Matlab.

PCA - Análise por componentes principais (do inglês “*Principal Component Analysis*”)

PCR - Regressão por componentes principais. (do inglês *Principal Component Regression*)

PLS - Regressão pelo Método dos Quadrados Mínimos Parciais. (do inglês “*Partial Least Squares*”)

PLS-DA - Análise discriminante por mínimos quadrados parciais (do inglês “*Partial Least-Squares Discriminant Analysis*”)

R²c - Coeficiente de determinação do conjunto de calibração.

R²cv - Coeficiente de determinação de validação cruzada.

RMSEC - Raiz quadrada do erro médio da calibração. (do inglês *Root Mean Square Error Of Calibration*)

RMSEP - Raiz quadrada do erro médio da previsão. (do inglês *Root Mean Square Error of Prediction*)

RMSECV - Raiz quadrada do erro médio da validação cruzada. (do inglês *Root Mean Square Error Of Cross-Validation*)

R²p - Coeficiente de determinação da predição.

SPA - Análise de janela permutada. (do inglês “*Subwindow Permutation Analysis*”)

SNV - Variação padrão Normal. (do inglês “*Standard Normal Variate*”)

SVC - Classificação por vetores de suporte, (do inglês “*Support Vector Classification*”)

SVR - Regressão por vetores de suporte. (do inglês *Support Vector Regression*)

THP - Resíduos de hidrocarbonetos totais de petróleo. (do inglês *Total Petroleum Hydrocarbon*)

UV-VIS - Ultravioleta/visível.

UNICAMP - Universidade Estadual de Campinas. (UNICAMP)

RESUMO

Com o avanço da tecnologia, os computadores chegaram aos laboratórios de química e, com isso, o volume e a complexidade dos dados cresceram, criando a necessidade de novos métodos para sua interpretação. Nesse contexto, surgiu a Quimiometria, cuja proposta era e ainda é possibilitar a extração do máximo de informação útil do alto volume de dados criados. A aplicação da Quimiometria exige conhecimento básico de programação, matemática e estatística que, em geral, não são ofertados em grande parte dos cursos de Química no Brasil, o que dificulta sua divulgação, uso dos métodos e familiaridade dos químicos leigos nesta área. Pesando nesta problemática, este estudo visa desenvolver um tutorial didático de Quimiometria, usando exemplos rotineiros e de fácil compreensão, além de incluir funções simples para iniciar aplicação de métodos mais comuns. O tutorial foi desenvolvido para o Matlab, software mais utilizado para este fim, e também adaptado ao software gratuito Octave. O tutorial foi desenvolvido para a metodologia mais utilizada para calibração multivariada: a Regressão pelo Método dos Quadrados Mínimos Parciais (PLS, do inglês “*Partial Least Squares*”), instruindo como otimizar seus parâmetros, avaliar, comparar os modelos, identificar *outliers*, aplicar pré-tratamentos, entre outras práticas relevantes. Dessa forma, o principal objetivo desse estudo foi facilitar o acesso às funções mais utilizadas e divulgar o conhecimento quimiométrico.

Palavras-chave: Tutorial; PLS; Octave; MatLab; Regressão Multivariada;

Abstract

With the advancement of technology, computers came to chemistry labs, and thus, the volume and complexity of data grew, creating the need for new methods for their interpretation. In this context, came chemometrics, whose purpose was and still is to make it possible to extract the maximum amount of useful information from the high volume of data created. The application of Chemometrics requires basic knowledge of programming, mathematics, and statistics that, in general, are not offered in most Chemistry courses in Brazil which hinders its dissemination, use, and familiarity of lay chemists in this area. Bearing this in mind, this study aims to develop a didactic tutorial on Chemometrics, using routine and easy-to-understand examples, in addition to including simple functions to start applying more common methods. The tutorial was developed for Matlab, the most used software for this purpose, and also adapted to the free Octave software. The tutorial was developed for the most used methodology for multivariate calibration: Partial Least Squares, instructing how to optimize its parameters, evaluate, compare models, identify outliers, apply pre-treatments, among other relevant practices. Thus, the main objective of this study was to facilitate access to the most used functions and disseminate chemometric knowledge.

Keywords: Tutorial; PLS; Octave; MatLab; Multivariate Regression.

1. INTRODUÇÃO

A Quimiometria surgiu por consequência do aumento exponencial do volume de dados a serem trabalhados no laboratório de química.¹ Conseguindo ser aplicada de formas diferentes, seja no planejamento de experimento,² para reduzir ensaios ou otimizar um sistema, seja na identificação de amostras, na classificação de câncer de pele,³ seja em problemas de regressão multivariados,⁴ para determinar propriedades físico químicas de petróleo, ou até possibilitando a aplicação de análises químicas rápidas em campo,⁵ onde antes, só era possível dentro do laboratório.

Embora tenha vários avanços, a utilização de quimiometria, como uma importante ferramenta do laboratório de química, têm encontrado barreiras no ensino na graduação,⁶ muitas vezes inexistente. Mesmo sendo aplicada com recorrência na Química Analítica, tanto no contexto acadêmico, como no industrial, é de se esperar que químicos dessa área dominem e apliquem quimiometria, todavia, não é o comum. Isso pode ser um reflexo tanto da falta de conhecimento teórico e prático tanto quanto na falta de acesso das ferramentas necessárias para sua aplicação, como funções envolvendo a modelagem desejada, como falta de compreensão em softwares de cálculo, voltados para computador, como Octave, um software gratuito, e Matlab, o mais recomendado, porém pago.

A Quimiometria pode ser separada em três grandes áreas, planejamento de experimentos, reconhecimento de padrões e calibração multivariada.⁷ Nesta última área é possível utilizar um conjunto de variáveis dependentes (Matriz **X**), ou como será utilizado aqui, Fonte Analítica, para estimar variáveis independentes quantitativa (Vetor **y**). Um exemplo desta aplicação é a utilização de infravermelho ou ressonância magnética nuclear para estimar a quantidade de enxofre e nitrogênio total, duas importantes propriedades para conhecer a aplicação e valor do barril de petróleo, utilizando Regressão pelo Método dos Quadrados Mínimos Parciais (PLS, do inglês “*Partial Least Squares*”).⁸ Este método de calibração multivariada foi introduzido na década de 1970 como uma abordagem na área econômica e posteriormente aplicado na área da química e nos dias atuais se trata do método de regressão mais aplicado na quimiometria.⁹ Além disso, o PLS é um método que é capaz de resolver problemas com rapidez e de maneira computacionalmente eficiente. Com isso, o conhecimento teórico e prático do PLS, além das funções e

rotinas necessárias, pode auxiliar em diversas pesquisas e análises químicas, tanto na academia quanto na área industrial. Assim, este estudo visa desenvolver um Tutorial completo, indo da teoria a prática e oferecendo o material necessário para a aplicação, além de dados para a prática.

2. REVISÃO BIBLIOGRÁFICA

2.1. Quimiometria

A Quimiometria é uma área da química relativamente nova, com um começo incerto, dependendo da definição e referencia, esta área chegou por consequência do grande avanço tecnológico criado com a evolução dos hardwares e softwares. Segundo a IUPAC¹⁰ podemos definir a quimiometria pela seguinte frase:

“Chemometrics is the application of statistics to the analysis of chemical data (from organic, analytical or medicinal chemistry) and design of chemical experiments and simulations”

Todavia, eu creio que a melhor forma de definir um quimiometrista é:

"Um químico com o conhecimento estatístico, matemático e de programação necessários para extrair e combinar informações químicas multivariáveis de diversas fontes analíticas."

No meu pensamento eu tento diferenciar o Químico Analítico do Quimiometrista com base num refinamento maior da matemática e estatística aplicadas em dados multivariados e complexos e a necessidade de utilizar programação (Matlab, Octave, Python e etc.) como ferramenta de aplicação do conhecimento em conjunto, em outras palavras, o Quimiometrista é um químico em Y. Onde nas beiradas temos um conhecimento específico de Matemática, Estatística e Propagação, unidas pelo conhecimento Químico (**Figura 1**). A compreensão da Matemática e Estatística, para dominar o passo-a-passo das ferramentas, tem que dominar alguma linguagem de programação para poder aplicar, corrigir e criar as ferramentas e isso tudo para obter informação química de maneira relevante. Vale ressaltar que não estou dizendo que uma área é melhor que a outra, longe disso, só estou enfatizando que elas têm focos diferentes e específicos.

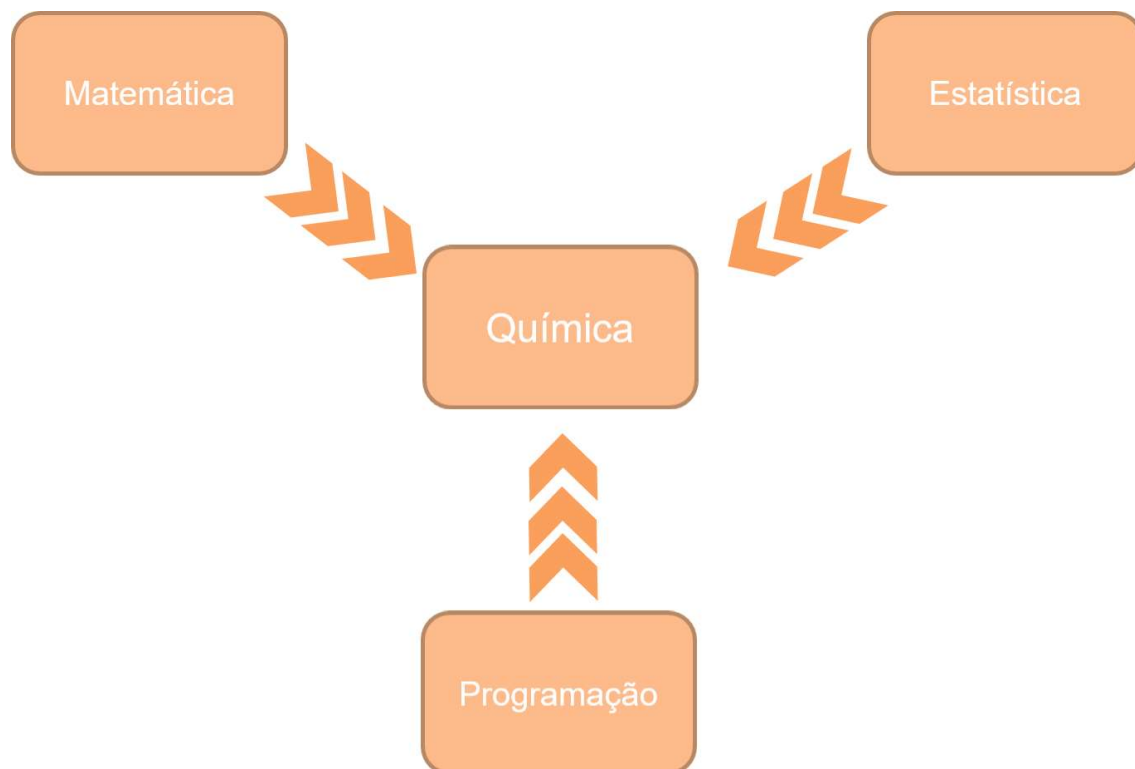


Figura 1. Conhecimentos necessários para ser um Quimiometrista.

Com base na definição da IUPAC poderíamos estimar que a Quimiometria começou com a Lei de Lambert-Beer, anterior ao século 19, nesta lei é admitido que há uma relação entre a absorvância de uma solução e a concentração de uma substância,¹¹ todavia, esta lei é basicamente aplicada em métodos univariados, sem interferentes químicos e necessidade de um cálculo computacional, assim, muitos não consideram como o começo da quimiometria.

O maior consenso sobre o começo da quimiometria está na proposta da análise por componentes principais (PCA do inglês “*Principal Component Analysis*”) feita por Pearson, em 1901, e desenvolvida por Hotelling 30 anos depois, mas só popularmente aplicada na década de 70,¹ devido ao avanço do algoritmo NIPALS (do inglês, ‘*Nonlinear Iterative Partial Least Squares*’), e ainda continua sendo largamente aplicado nos dias de hoje.¹²

A Programação é essencial para a aplicação da quimiometria, devido aos inúmeros cálculos necessários para a aplicação das técnicas mais simples. O uso de computadores permitiu aos químicos aplicar uma calibração mais elaborada, como Moraes, C. L. M. *et al*, em 2017,¹³ demonstrou ao aplicar infravermelho para conseguir distinguir espécies de fungos, ou Santos, F. D., *et al*, em 2021,¹⁴ que utilizou espectros de infravermelho portátil

para conseguir identificar óleo adulterado em trabalhos em campo. Caso um pesquisador quisesse aplicar os cálculos necessários para estes estudos usando somente calculadora científica, papel e lápis, levaria meses ou talvez anos só para terminar os cálculos, enquanto hoje em dia um computador de laboratório de química demoraria segundos. Além disso, quando analisamos os computadores utilizados no primeiro artigo de Quimiometria e comparamos com a capacidade processional dos computadores de hoje, percebemos um salto gigantesco que tivemos e podemos ir mais além.

2.1.1. Desenvolvimento no Brasil

No Brasil a quimiometria começou a passos lento, como Neto B. B. et al, em 2006, narrou no artigo intitulado “25 ANOS DE QUIMIOMETRIA NO BRASIL”.¹ No final da década de 70, um pequeno grupo de quimiometristas começou a se formar na Universidade Estadual de Campinas (UNICAMP), devido às limitações das época, eles usavam um computador comunitário de grande porte do tipo PDP-10, **Figura 2**, que ostenta uma grande memória ram de 10 Megabytes, para fins de comparação um Iphone 11, lançado em setembro de 2019, tem, no mínimo, 4 Gigabytes de ram, o que equivale a 400 vezes mais memória que o PDP-10.



Figura 2. Computador PDP-10 utilizado pela Unicamp na década 70. Fonte: <https://www.ccuec.unicamp.br/ccuec/site-historico/anos-60> (09/01/2023)

A tecnologia da época era bem defasada se comparada as atuais, o processamento de cálculos considerados simples hoje em dia, demorava dias, a programação era feita em fitas magnéticas e o número de computadores era limitado. Nesse cenário, Ieda S. Scarminio conseguiu defender a primeira dissertação de quimiometria do Brasil, em 1981, e publicar o primeiro artigo em 1982, um estudo que analisou a concentração de 18 em amostra de água minerais de São Paulo, todavia, até o dia de hoje, não conseguimos acesso a essa dissertação e artigo para podermos detalhar melhor.

Nessa época, o maior empecilho para se aplicar quimiometria estava na necessidade de computadores e na dificuldade de ter estes no laboratório. Isso melhorou quando os microcomputadores chegaram ao Brasil. Na UNICAMP existia um microcomputador que tinha duas entradas de disquetes, uma para adicionar o programa e outra para salvar os dados, ambos com capacidade de 32 kbytes. O grupo de quimiometria com a posse do software ARTHUR, o primeiro programa de quimiometria do mundo, adaptou as sub-rotinas deste para os dois disquetes e conseguir assim desenvolver trabalhos com a empresa Rio Doce Geologia e Mineração S.A. (Docegeo), contudo, esse programa não era facilmente adaptável em outros computadores, o que dificultou a divulgação da quimiometria em território nacional.

A grande mudança veio no ano de 1985, uma combinação de fatores possibilitou o grande crescimento da quimiometria no Brasil. Entre eles, a chegada dos computadores de 16 bits, computadores menores, mais baratos e de fácil uso com diversos programas derivados do ARTHUR. O interesse da indústria em aplicar técnicas quimiométricas, o que trouxe grande incentivo e verba para o grupo da UNICAMP. Com cursos intensivos de extensão em quimiometria aplicados por Roy E. Bruns em outras faculdades, o primeiro curso de pós graduação de Quimiometria, com 60 h semanais, no IQ-Unicamp, com dois alunos. Além de um artigo de revisão de Quimiometria publicado na Química Nova.¹⁵ Na **Figura 3**, podemos ver o resultado ao pesquisar “*chemometrics**” na plataforma *Web Of Science*, sem intervalo de tempo e só artigos publicados por instituições brasileiras.

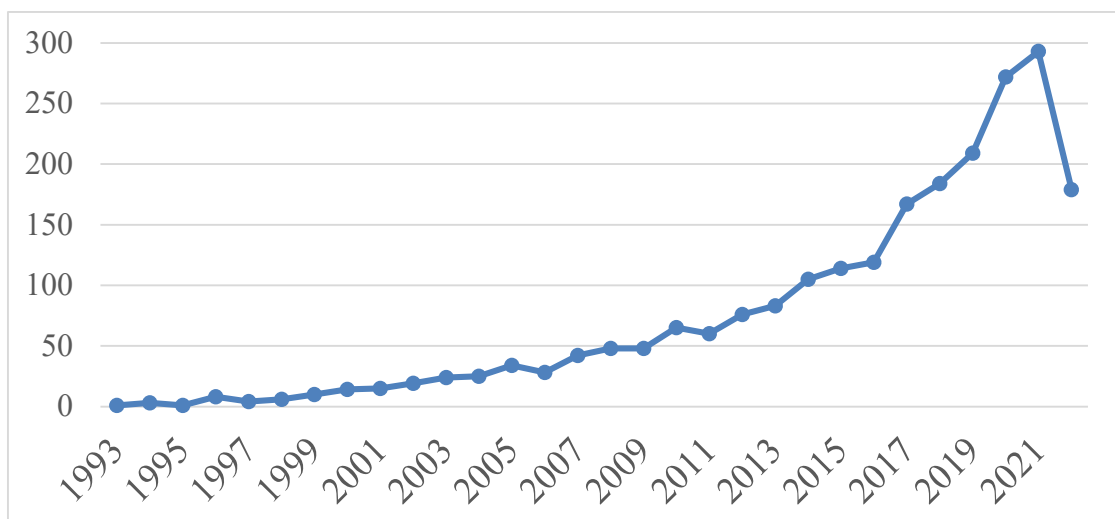


Figura 3. Gráfico de artigos publicados com a palavra “chemometric*”, pesquisado no Web of Science dia 09 de Janeiro de 2023.

Apesar de nitidamente a pesquisa incluir os artigos citados anteriormente, os pioneiros, fica evidente que a quimiometria avançou consideravelmente ao longo das últimas décadas, indo de 1 artigo em 1993 a 293 artigos em 2021. Isso provavelmente se deve a um conjunto de fatores, entre eles, o avanço da tecnologia da computação, que possibilitou computadores menores e com maior potência de processamento, a divulgação da quimiometria, com ensino na pós-graduação e cursos intensivos, e divulgação de funções e rotinas para a aplicação de quimiometria.

2.1.2. Ensino no Brasil

Uma análise dos principais cursos de química do país, de acordo com a classificação do MEC, como pode ser visto na **Tabela 1**, aponta ainda um pobre ensino de quimiometria em nível de graduação. Entre os cursos mais bem colocados, em sua maioria, possuíam de modo obrigatório ou optativo a disponibilidade de disciplinas que tange áreas relacionadas, como estatística e programação. Neste diagnóstico essas disciplinas foram encontradas como Estatística básica ou Introdução à estatística, estatística em/voltada à química, programação, química computacional e quimiometria, sendo estatística básica a mais frequente, presente na grade curricular de dezoito dos trinta cursos analisados.

Tabela 1. Análise da presença de cursos importantes para a Quimiometria no currículo.

Posição Mec	Sigla	Pública / Privada	Avaliação de Mercado	Avaliação de Ensino	CPC	ENADE	Ensino				
							Estatística Básica	Estatística em Química	Programação	Química Computacional	Quimiometria
1	USP	Pub	36	63.38	-	-	Obrigatória	-	-	Optativa	Obrigatória
2	UNICAMP	Pub	35.6	61.44	4	5	-	-	-	-	Optativa
3	UFRJ	Pub	35.6	58.95	4	3	-	Optativa	Obrigatória	Optativa	-
4	UFMG	Pub	34.4	57.7	4	5	Obrigatória		Optativa		
5	UFSCAR	Pub	33.2	57.48	4	4	-	-	-	Obrigatória	-
6	UFRGS	Pub	30.8	57.17	5	5	Optativa	-	-	Optativa	Optativa
7	UNESP	Pub	35.6	52.14	5	5	Obrigatória	-	-	-	-
8	UFPR	Pub	30.8	48.53	4	4	Obrigatória	-	-	-	Optativa
9	UFC	Pub	34.4	43.5	4	4	Obrigatória	-	-	Obrigatória	-
10	UFSC	Pub	30.8	46.67	3	3	-	Obrigatória	-	Optativa	-
11	UFF	Pub	33.2	43.2	4	3	-	-	-	Optativa	-
12	UFPE	Pub	33.2	40.12	4	4				Eletiva	Obrigatória
13	UNB	Pub	30.8	42.2	4	3	Optativa	-	Optativa	Optativa	Optativa
14	UFSM	Pub	22.4	47.66	4	3	Obrigatória	-	-	-	Optativa
15	MACKENZIE	Privada	33.2	34.95	3	3					
16	UFRRJ	Pub	22.4	43.36	4	4	Obrigatória	-	-	Obrigatória	-
17	UFV	Pub	22.4	43.1	4	3	Obrigatória	Optativa	Obrigatória		Optativa
18	UFU	Pub	22.4	42.44	3	2	Obrigatória	-	-	Optativa	-
19	PUC Rio	Privada	22.4	39.26	5	5	-	-	Optativa	-	Obrigatória
20	PUC RS	Privada	30.8	30.6	3	2	-	-	-	-	-
21	UFABC	Pub	9.2	44.81	4	4	Obrigatório				

22	UFJF	Pub	30.8	22.9	5	5	Obrigatória	-	Obrigatória	-	-
23	UFBA	Pub	33.2	20.18	3	3	Obrigatória	Optativa	-	-	-
24	UERJ	Pub	34.4	18.13	4	4	-	-	-	-	-
25	UNEB	Pub	30.8	21.29	3	4	Obrigatória	-	-	-	-
26	UEM	Pub	30.8	20.21	4	3	-	Obrigatória	-	-	-
27	UEL	Pub	30.8	20.14	3	3	Obrigatória	-	-	-	-
28	UFRN	Pub	30.8	19.61	3	2	-	Optativa	Optativa	-	Obrigatória
29	UFS	Pub	30.8	19.6	4	3	Obrigatória	-	-	Optativa	Obrigatória
30	UFG	Pub	30.8	18.61	3	3	Obrigatória	-	-	-	Optativa
39	UFES	Pub	22,4	21,32	4	4	Optativa	-	-	-	-

Estatística básica, ou Introdução à estatística, resume sua ementa em capacitar o aluno a organizar e descrever conjuntos de dados de forma estatística, dominar os fundamentos básicos de probabilidade, inferência estatística, algarismo significativos e outros. **Estatística voltada para química** se trata do ensino de métodos estatístico que auxiliem no cotidiano do químico, tendo maior ligação com a Química Analítica, podendo ser usado para ensinar algarismos significativos, identificar erros de tendência, erros de método, Distribuição t de student e até na aplicação de regressão linear, como a Lei de Lambert-Beer.

Programação visa habilitar o aluno em organização sistemas de computação, algoritmos, tipos de dados e programas, introdução a uma linguagem de programação orientada e suas aplicações. **Química computacional** trata de aproximações numéricas, aritmética de ponto flutuante, erros, características básicas de organização de um computador, algoritmos de programas, programação básica, representação de números e solução de problemas numéricos e não numéricos por computadores.

Quimiometria, varia conforme o curso, podendo se basear somente em métodos não supervisionados em inteligência artificial, análise de componentes principais (PCA) e análise hierárquica de agrupamentos (HCA, do inglês “*Hierarchical Cluster Analysis*”), além de poder baseado em métodos simples supervisionados, como no caso de reconhecimento de padrões, análise discriminante por mínimos quadrados parciais (PLS-DA, do inglês “*Partial Least-Squares Discriminant Analysis*”) e na calibração de multivariada com PLS.

Relacionando a atuação destas disciplinas na classificação dos cursos se observa que dos trinta primeiros colocados dezenove deles possuíam duas ou mais destas disciplinas em sua grade, oito deles possuíam apenas uma e somente três não tinham nenhuma delas entre seus componentes curriculares e estas não estavam entre as dez primeiras colocações.

2.1.3. Academia no Brasil

A Quimiometria vem ganhando destaque no meio acadêmico como a **Figura 3a** demonstra, nessa figura podemos analisar a quantidade de artigos encontrados no *Web of Science* no intervalo entre os anos de 2006 a 2022, utilizando a palavra “*chemometric**”, de modo que fosse possível mapear o crescimento de publicações que abordam quimiometria no Mundo. Já na **Figura 3b** podemos ver o mesmo gráfico contabilizando

os artigos publicados no Brasil.

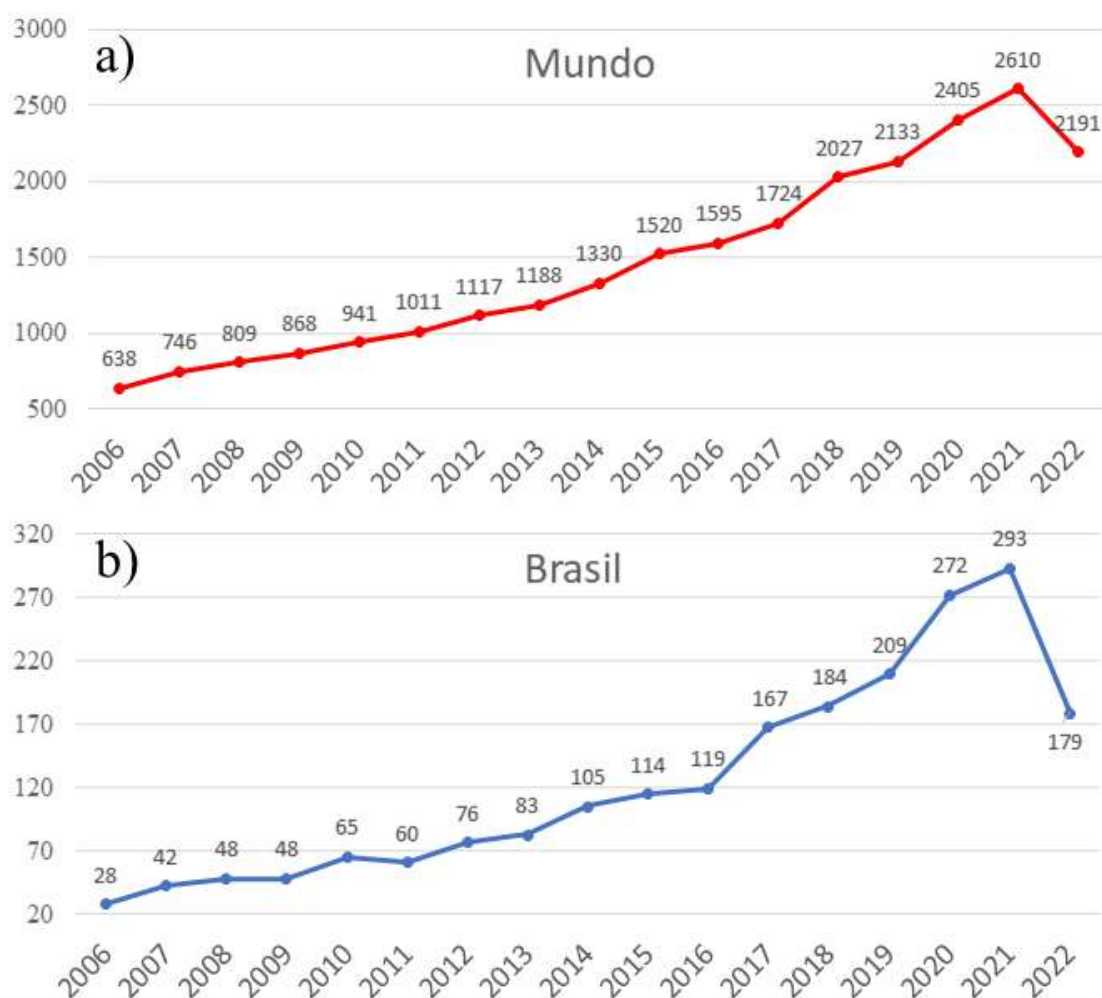


Figura 4. Gráfico de artigos publicados, com a palavra “chemometric*”, por ano, mundo (a) e Brasil (b). Pesquisado no Web of Science dia 09 de Janeiro de 2023.

O progresso desse seguimento da química no país é evidente e significativo apresentando um crescimento em torno de 539% de 2006 a 2022. Nota-se ainda que o crescimento nacional, em percentual, é bem superior ao crescimento mundial que foi igual a 243%, isso demonstra não só que a Quimiometria vem crescendo no Brasil como também que a participação mundial do Brasil vem crescendo também.

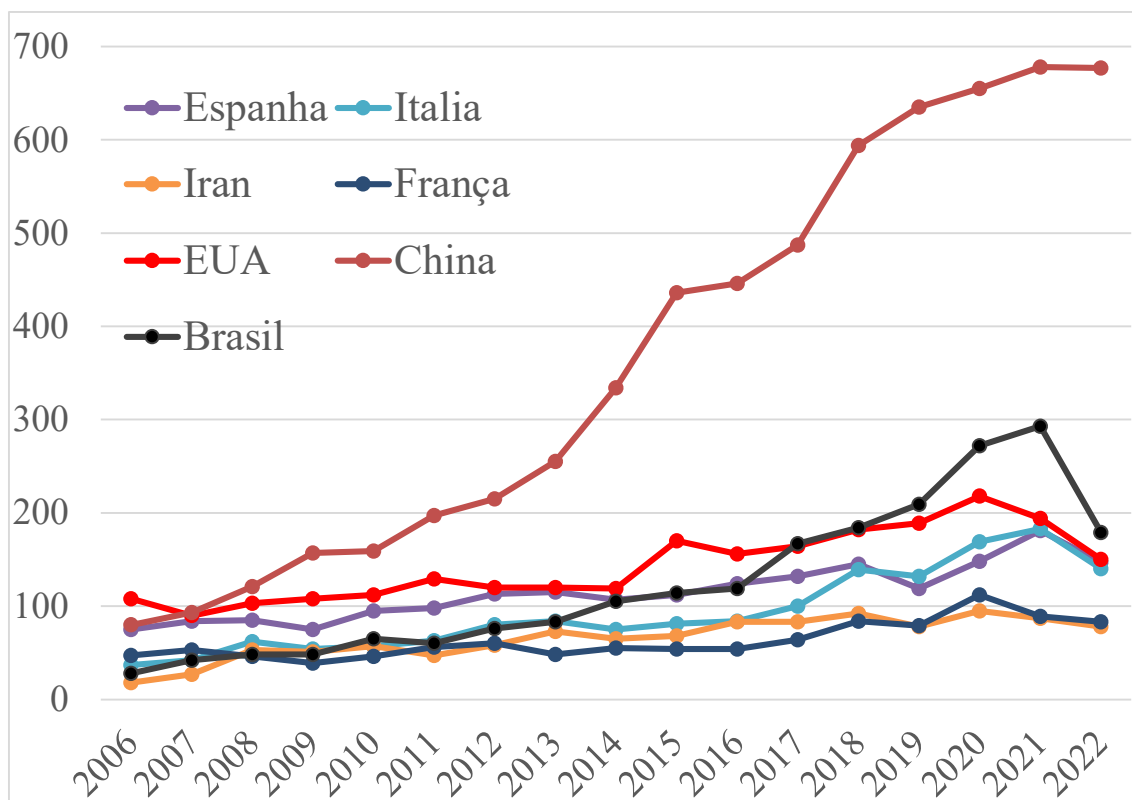


Figura 5. Gráfico de artigos publicados, com a palavra “chemometric*”, por ano e país. Pesquisado no Web of Science dia 09 de Janeiro de 2023.

Na Erro! Fonte de referência não encontrada. podemos ver a publicação de artigos com base no ano e no país de origem, inicialmente, em 2006 os EUAs liderava com 108 artigos totais, o que equivaleria a 16,9% da produção mundial, enquanto que o Brasil tinha 28 artigos, 4,4%, nos dias atuais a China lidera sendo responsável por 677 publicações em 2022, o que equivale a 30,9% da publicação mundial, e o Brasil teve 179 publicações, 8,2% da mundial. Este gráfico reforça a ideia de o Brasil estar crescendo a sua participação na publicação de quimiometria no mundo, ficando em segundo lugar nos últimos anos.

Também foi analisado as principais faculdades que publicam sobre quimiometria, utilizando os mesmos parâmetros, no período de 2005 a atualmente. Na **Tabela 2** estão os resultados obtidos, percebe-se que apesar da predominância de faculdades públicas, temos o destaque da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) uma empresa pública que desenvolve pesquisa voltadas para pesquisa agrícola, além disso, temos o destaque da UNICAMP, a precursora da quimiometria no Brasil.

Tabela 2. Principais instituições que publicam artigos na quimiometria.

PUBLICAÇÕES DE 2005 A 2023 (BRASIL)		
POSICÃO	INSTITUIÇÃO	Nº PUBLICAÇÕES
1	UNICAMP	417
2	USP	247
3	EMBRAPA	172
4	UFSCar	149
5	UTFPR	134
6	Unesp	130
7	UFMG	120
8	UFRGs	118
9	UFBA	98
10	UEPB	96
11	UFPE	93
12	UFES	89

Ademais, se analisarmos esse crescimento em intervalos menores, a cada 5 anos, percebe-se que os dois primeiros lugares ficam sempre UNICAMP e a universidade de são Paulo (USP), respectivamente, nas demais posições ocorrer uma variação. Além disso, todas faculdades da **Tabela 3**, também estão presentes na **Tabela 2**.

Tabela 3. Principais instituições que publicam artigos na quimiometria por faixa de 5 anos.

PUBLICAÇÕES DE 2005 A 2009 (BRASIL)			PUBLICAÇÕES DE 2010 A 2014 (BRASIL)		
POSICÃO	INSTITUIÇÃO	Nº PUBLICAÇÕES	POSICÃO	INSTITUIÇÃO	Nº PUBLICAÇÕES
1	UNICAMP	89	1	UNICAMP	101
2	USP	35	2	USP	48
3	UFPE	14	3	UFSCar	37
4	UFScar	14	4	UFPB	28
5	Unesp	12	5	EMBRAPA	24
21	UFES	4	12	UFES	16
PUBLICAÇÕES DE 2015 A 2019 (BRASIL)			PUBLICAÇÕES DE 2020 A 2023 (BRASIL)		
POSICÃO	INSTITUIÇÃO	Nº PUBLICAÇÕES	POSICÃO	INSTITUIÇÃO	Nº PUBLICAÇÕES
1	UNICAMP	120	1	UNICAMP	107
2	USP	81	2	USP	83
3	EMBRAPA	65	3	EMBRAPA	74
4	UFRGS	60	4	UTFPR	63
5	UFSCar	51	5	UFC	48
18	UFES	26	8	UFES	43

2.2. Quimiometria Aplicação

A construção de um modelo em Quimiometria pode ser dividida em três etapas distintas; Obtenção dos dados, onde as amostras são coletadas, definidas e organizadas; Processamento dos Dados, onde os dados são aperfeiçoados e modelados e a última Etapa; Avaliação, onde os modelos obtidos são avaliados e validados.

2.2.1. Obtenção dos dados

A Obtenção dos Dados consiste na determinação de qual será o objeto de análise e suas variáveis independentes (vetor y) e dependentes (matriz X), este também chamado de fonte analítica. Essas são organizadas em matrizes X , onde o vetor linha representa as amostras e o vetor coluna representa as variáveis, neste caso todas as amostras devem ter a mesma faixa de resolução, ou quantidade de colunas. Para n amostras com m variáveis espectrais, a matriz de dados terá dimensão $X_{(n,m)}$ (n linhas por m colunas), assim como

é mostrado na **Equação 1**.

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \quad \text{Equação 1}$$

O vetor \mathbf{y} , caso seja mais de uma variável independente damos o nome de matriz \mathbf{Y} , é organizada de forma semelhante a matriz \mathbf{X} , com o vetor linha representando a amostra. Nesta etapa é de suma importância escolher variáveis dependente com alguma relação com a variável independente. Como por exemplo:

O trabalho de Folli G. S. *et al*, em 2022,⁵ que aplicou infravermelho próximo (NIR, do inglês near infrared) portátil, ou MicroNir, para determinar adulterantes em alimentos, como óleo em azeite, açúcar em mel e outros. Neste caso a fonte analítica é o infravermelho próximo portátil e a variável independente é a quantidade de adulterantes.

Outro caso interessante, foi publicado por Paulo E. H. *et al*, em 2022,¹⁶ que utilizou como Matriz \mathbf{X} a ressonância magnética nuclear de carbono (NMR C^{13} do inglês “Nuclear magnetic resonance carbon-13”) e como vetor \mathbf{y} o poder calorífico das amostras de petróleo, neste caso utilizou a fonte analítica para prever indiretamente um propriedade físico química das amostras.

2.2.2. Processamento de Dados

O Processamento de Dados é a parte onde aplicamos aperfeiçoamentos, para melhorar a extração de informações relevantes, e a parte da modelagem, onde criamos modelos de predição. Eu gosto de explicar essa parte utilizando a “Filosofia do Mineiro”, que podemos ver logo abaixo:

“O mineiro tem o dado bruto em suas mãos, mas ele não é maleável, difícil de extrair e complicado para lidar, então, o mineiro, em sua persistência, minera o dado até ele se tornar maleável, de fácil extração e simples de lidar.”

A parte de tornar maleável e fácil extração refere-se a parte de aperfeiçoamento que antecede a parte de modelagem, nessa parte visamos remover ruídos e variáveis pouco informativas, da variável dependente, enquanto destacamos as variáveis

informativas e mais relacionadas a nossa variável independente. Esta etapa será melhor explicada adiante. A última parte e não menos importante, a modelagem é onde construímos os modelos de previsão da nossa variável independente, podendo ir de uma simples aplicação de PCA a uma complexa classificação por vetores de suporte, (SVC, do inglês “*support vector classification*”).

2.2.2.1 Aperfeiçoamento

Nesta etapa temos como objetivo aplicar métodos nos dados brutos de forma a aprimorar os modelos. Podemos aperfeiçoar um dado de formas diferentes, como; separação em conjunto calibração e teste, onde os dados brutos são separados em conjunto de calibração, que cria o modelo, e conjunto de teste, que valida o modelo, pré-tratamento,⁹ onde a fonte analítica é tratada de forma a aperfeiçoar a interpretação dos dados, seleção de variáveis,¹⁷ onde se remove variáveis que não contém informações desejada e seleciona aquelas que tem, e a fusão de dados,¹⁸ onde combina diferentes fontes analíticas para obter um melhor resultado.

Separação em conjunto calibração e teste, nesta etapa do processo temos o objetivo de criar dois conjuntos de amostras, o conjunto calibração, também chamado de treinamento no reconhecimento de padrões, que tem como objetivo construir e otimizar o modelo e o conjunto teste, que contém amostras que não fazem parte do conjunto calibração e tem como objetivo medir a capacidade preditiva. Além desses dois conjuntos podemos ter o conjunto desconhecidos, onde só se tem a fonte analítica das amostras e as variáveis dependentes são desconhecidas, até mesmo para o Quimiometrista que criou o modelo, este conjunto é utilizado comumente na indústria para conferir com maior certeza se o modelo é adequado.

A Separação pode ser feita de diversas formas, desde a simples manual, onde o Quimiometrista decide quais amostras ficam no conjunto calibração e quais ficam no conjunto teste, como utilizar o método de Kennard-Stone (KS).¹⁹ Neste método o objetivo principal é selecionar um conjunto de calibração representativo, de máxima variabilidade, utilizando o centro do conjunto de dados e a distância euclidiana desses. Contudo, não existe uma regra sobre qual a melhor forma de separar os conjuntos amostrais, dependendo de o Quimiometrista conhecer seus dados e objetivo e decidir a melhor forma.

Os **pré-tratamentos** modificam a fonte analítica para facilitar a extração de

informação, são altamente recomendados, todavia tem que se tomar cuidado com a sua aplicação, pois uma aplicação incorreta pode remover informações relevantes em vez de destacá-la. Entre os pré-tratamentos mais utilizados podemos citar; Variação padrão Normal (SNV, do inglês *Standard normal variate*), Correção multiplicativa de sinal (MSC, do inglês *Multiplicative Signal Correction*), derivada (Deriv) Centragem na média (Center) e Auto-escalamento (Auto).

Existem dois tipos de pré-tratamento, os que são aplicados em relação a amostra, trabalham com o vetor linha da matriz \mathbf{X} , e os que são aplicados em relação a variável, trabalham com a coluna da matriz. No primeiro caso temos a derivada, este método é constantemente utilizado em fontes analíticas que tem problemas de deslocamento e inclinação de linha de base,⁹ recomendado ser usado somente em variáveis contínuas, como infravermelho,¹⁴ e em alta resolução, além disso, a derivada pode ser aplicada uma vez, primeira derivada, ou duas vezes, segunda derivada, no mesmo espectro. A primeira derivada corrige o deslocamento da linha de base enquanto a segunda a inclinação da linha, em ambos os casos o perfil espectral muda consideravelmente, como veremos na 5.4. Parte 02 - Aprimorando o modelo.

Outro método que trabalha em relação a amostra é o MSC, este método é recomendado utilizar em fontes analíticas que sofrem efeitos aditivos e multiplicativos, efeito que são causados por fenômenos físicos, como mudança de caminho ótico, temperatura e pressão. Neste método, MSC, utiliza-se o espectro médio como referência, uma constante de absorbância para corrigir o efeito aditivo, e o coeficiente angular para corrigir efeitos multiplicativos.⁹ Este método é recomendado para fontes analíticas como infravermelho,²⁰ ultravioleta/visível (UV-Vis) e Raman, diferente da derivada no MSC o perfil espectral não é modificado e trata o efeito multiplicativo, contudo não corrige a inclinação da linha de base. Neste método tem que se tomar cuidado para a constante de absorbância e coeficiente angular não estarem relacionados com a variável independente (vetor \mathbf{y}), pois caso seja, poderá perder informações importantes para o modelo.

Entre os pré-tratamentos que trabalham com o vetor coluna da matriz \mathbf{X} , temos o auto-escalamento, nesse método é aplicado um ajuste matemático onde cada coluna é centralizada na média e dividida pelo desvio-padrão, com isso a matriz \mathbf{X} fica adimensional, dando o mesmo peso para cada variável, este tipo de tratamento é recomendado para fontes analíticas de natureza distintas, como concentrações e propriedades físico químicas de diferentes dimensões,⁹ apesar de funcionar com outras fontes como o infravermelho¹⁴ e NMR C¹³.¹⁶ Além desse método, temos também o Center,

este é o pré-tratamento mais utilizado em dados espectrais, neste método se calcula a média de cada coluna da matriz X e em seguida subtrai este valor de cada uma das variáveis da coluna, o resultado é uma translação de eixos para o valor médio de cada um deles.⁹

2.2.2.2 Modelagem

A Modelagem é o coração da quimiometria, nesta etapa é utilizada métodos estatísticos e matemáticos para conseguir modelos com a capacidade preditiva desejada. Podemos separar esses modelos em dois grupos principais, reconhecimento de padrões não supervisionado e reconhecimento de padrões supervisionado, no primeiro podemos citar o PCA,²¹ e no segundo podemos separar em dois subgrupos classificação multivariada, que trabalha com predição qualitativa de variáveis dependentes,²² onde podemos utilizar modelagem como PLS-DA²³ e SVC,³ e no segundo subgrupo regressão multivariada,²⁴ que trabalha com predição quantitativas, que podemos utilizarmos métodos como regressão por componentes principais (PCR, do inglês *principal component regression*), PLS²⁵ e regressão por vetores de suporte (SVR, do inglês *support vector regression*).²⁶

No **não supervisionado** o seu principal objetivo é encontrar tendências de agrupamento e semelhança entre as amostras. Um exemplo clássico é a PCA, o precursor da quimiometria, neste método é extraído uma nova variável da matriz X, preservando a maior quantidade de variância possível, esta variável é denominada componente principal, o que sobra da extração é chamado de matriz resíduo, dessa matriz é extraído uma nova componente principal de forma a ser ortogonal a anterior, dessa forma as variáveis não são correlacionadas, e este processo é repetido até chegar ao número de componentes principais solicitadas.²⁷ Além disso, é obtido o que chamamos de Scores e Loadings, o primeiro informa a posição de uma amostra em determinada componente principal e o segundo quais variáveis são importantes para aquela componente principal. As informações adquiridas pela aplicação desta técnica podem ser aplicadas de diversas formas, contudo diferente das técnicas mais modernas, sua avaliação é manual dependendo da experiência e conhecimento do quimiometrista.

Wu B. et al, em 2022,¹² demonstraram uma aplicação da PCA ao analisar a variação espacial e risco ambiental em campos petrolíferos causados por resíduos de hidrocarbonetos totais de petróleo (THP, do inglês Total petroleum hydrocarbon) em

campos Chineses utilizando como informação base teor de aromático e saturados e outras propriedades importantes do solo. A PCA foi utilizada para verificar quais variáveis tem maior relação com o índice de aromáticos e saturados do petróleo. Utilizando os loadings, o artigo concluiu que Aromáticos e Saturados são diretamente relacionados, além disso, também se percebe uma forte relação entre a presença de Argila e a concentração de Material Orgânico, o que pode ser justificado por que ambos tem a capacidade de absorver moléculas aromáticas e saturadas, ainda, o artigo relacionada negativamente a Temperatura Acumulada, o que deve ocorrer devido ao potencial de degradar as moléculas orgânicas foco do estudo.

Na **classificação multivariada**, do grupo dos métodos supervisionados, cada amostra tem sua classe, valor qualitativo, pré-estabelecida e essa informação é utilizada na construção do modelo. Assim, nesse modo, o modelo é treinando já sabendo em qual classe as amostras pertencem e quais são as semelhantes dentro do mesmo conjunto e quais as diferentes entre conjuntos diferentes. Um exemplo desse método é o PLS-DA, este método se trata de uma adaptação do PLS aplicado na classificação multivariada, a adaptação é realizada na forma que é utilizado a matriz Y, em vez de um vetor y quantitativo. A matriz é formada por n colunas, sendo “n” o número de classes existentes, onde os valores presentes são 1, ou 0, sendo 1 representando que a amostra pertencente aquela classe e 0 que não pertence. Para definir em qual classe uma amostra será colocada é aplicado uma técnica conhecida como estatística Bayesiana, onde um limite é traçado entre as classes “pertence” e “não pertence”, quem estiver abaixo do limite é classificado como “não pertence” e acima do limite como “pertence”.²³

O Potencial do PLS-DA foi demonstrado por Mohammadi M. *et al*, em 2022,²⁸ neste estudo tentou classificar amostras de petróleo com base no seu teor de enxofre, entre amostras ácidas, com maior concentração de enxofre, e doce, com menor concentração de doce. Onde observou o destaque de cinco bandas importantes para o modelo. Em 727 cm^{-1} temos a parte que chamamos de impressão digital, onde temos informação de vibração de flexão, em 1300 cm^{-1} temos estiramento e vibração de grupos funcionais de N e S. Os picos em 2860 e 2950 cm^{-1} são picos característicos de petróleo, o primeiro tem relação com bandas de CH e CH₃ de vibração de flexão e o segundo com funções aromáticas. O PLS-DA conseguiu errar somente uma única amostra em sua modelagem, o artigo ainda compara o método PLS-DA com SVC, um método não linear, entretanto, ambos obtiveram a mesma acurácia, o que sugere que esta propriedade seja linear.

Como o foco principal deste trabalho é a **Regressão Multivariada**, veremos sobre

o assunto no próprio tópico com maiores detalhes.

2.2.3. Avaliação

A Avaliação é a última das etapas, consiste em validar o modelo utilizando métricas já estabelecidas na literatura. Podemos utilizar **parâmetros de avaliação**, antigamente chamados de figura de mérito, **deteção de outlier**, onde identificamos e removemos amostras anômalas, e **testes de comparação**, onde analisamos se dois modelos são estatisticamente iguais, ou não. Quando falamos de avaliação, precisamos pensar no objetivo que temos, dependendo do objetivo teremos uma avaliação diferente.

Quando avaliamos uma classificação multivariada, por exemplo, podemos utilizar; Tabela de contingência, onde é apresentado as amostras de cada classe que foram corretamente e incorretamente classificadas, Sensibilidade, que é a habilidade do modelo em classificar corretamente amostras de uma determinada classe, Especificidade, a capacidade do modelo de classificar corretamente amostras que não pertencem a uma classe especificar e por último Acurácia, um parâmetro único que avalia todo o desempenho do modelo, analisando todas amostras corretamente classificadas.⁶

Nos parâmetros de avaliação voltados para regressão multivariada podemos destacar; Raiz quadrada do erro médio (RMSE, do inglês root mean square error), que pode ser tanto de calibração (RMSEC), quando referente ao conjunto de calibração, tanto para validação cruzada (RMSECV) e predição (RMSEP). Este parâmetro mede a diferença entre o valor medido e o previsto pelo modelo, ou seja, quanto mais próximo de zero, mais preciso é o modelo. Podemos ver seu cálculo na **Equação 2**;

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

Onde y_i é o valor medido, \hat{y}_i é o valor predito, n é o número de amostras ou de calibração, para o caso do RMSEC e RMSECV, ou de predição, para o RMSEP. O v tem o valor do número de amostras de calibração (n_{cal}) quando aplicado no RMSECV e o RMSEP utiliza o número de amostras de teste (n_{pred}) e no RMSEC, para o PLS, $n_{cal} - VL - 1$, isso se deve a perda do grau de liberdade conforme a quantidade de VL.

O Coeficiente de determinação (R^2), que assim como o RMSE pode ser para calibração (R^2_c), validação cruzada (R^2_{cv}) e predição (R^2_p). Neste parâmetro é avaliado

o grau de concordância entre os valores medidos pelos saios experimentais e os valores preditos pelos modelos. O cálculo podemos ver na **Equação 3**.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad \text{Equação 3}$$

Onde \bar{y} é a média das observações. O valor do R^2 varia de 0 a 1, sendo quanto mais próximo de 1 maior é o grau de concordância do que foi previsto pelo modelo e o que foi realmente medido. O R^2 tem a vantagem, em comparação com o RMSE, de não depender a faixa do conjunto amostral utilizado no modelo, assim, é interessante utilizá-lo para comparar modelos de diferentes faixas amostrais.

Outros parâmetros de avaliação podem ser necessários, dependendo do objetivo da modelagem, como é o caso do limite de quantificação (LoD, do inglês *limit of detection*) e limite de quantificação (LoQ, do inglês *limit of quantification*), o LoD define o menor valor da variável dependente que pode ser detectado pelo modelo e o LoQ o menor valor que pode ser quantificado.²⁹ Estes dois são de grande importância quando temos o objetivo de detectar adulterantes.⁵

Uma forma gráfica de analisar os dados da regressão multivariada é o uso do gráfico de medido x predito é construído ao colocar y_i , o valor medido, e o \hat{y}_i , valor predito pelo modelo, num gráfico com uma linha de tendência. O objetivo da construção é analisar como as amostras se comportam ao longo da faixa amostral, conseguindo assim identificar tendências e variância do modelo. Amostras muito afastada da linha de tendência podem ser indicativas da presença de outlier.

A Detecção de Outlier, como o nome sugere, é analisar as amostras do modelo de forma a identificar amostras que tem um comportamento anômalo, isso pode ocorrer de diversas formas, erro experimental, erro de digitação e até ser uma amostra que não pertencente aquele conjunto amostral.⁹ Assim, a presença desse tipo de amostra pode prejudicar os parâmetros de avaliação do modelo, sendo ainda mais grave quando se encontra no conjunto de calibração, ou treinamento. Nesse modo, sua detecção e remoção do conjunto amostral é de vital importância, neste estudo optamos por ensinar a aplicação da análise de Resíduo e Leverage, sem tradução.

O Leverage é uma forma de avaliar a influência de uma amostra na construção do modelo de regressão. Um valor pequeno indica que a amostra tem pouca influência e um valor alto, grande influência, caso uma amostra seja uma amostra anômala ela tem a

tendência de ter uma alta influencia no modelo,³⁰ contudo, precisa passar por um limite determinado pela ASTM 1655-05.³¹ Sozinho, este parâmetro só oferece uma tendência da amostra se tratar de um outlier, por isso, também analisamos o Resíduo, que é o valor absoluto da diferença entre o valor medido e predito pelo modelo, caso este também passe do limite, o dobro do RMSEP do modelo, temos um forte indicio que a amostra se trata de um outlier. Caso a amostra esteja no conjunto calibração, precisamos refazer todo o modelo, caso só no conjunto teste, somente a parte final.

Por último, a comparação entre modelos, quando um quimiometrista lida com uma análise real, normalmente, é desenvolvido diversos modelos diferentes, alguns com parâmetros de avaliação bem distinto, outros com parâmetros de avaliação próximo, neste último caso, as vezes, não é possível determinar se um modelo é igual, ou diferente, do outro apenas com as ferramentas aqui apresentadas, então aplicamos um teste de comparação entre modelos, entre elas temos o teste randômico de exatidão,³² trata-se da aplicação do teste t randomizado para verificar se dois modelos são iguais preditivamente, utilizado justamente para confirmar estatisticamente se dois modelos tem, ou não, a mesma capacidade preditiva. Este teste é relativamente rápido e consegue resultados precisos independente se é um modelo quantitativo ou qualitativo.^{33,34}

2.3. Regressão Multivariada

A Regressão Multivariada é uma das grandes áreas da Quimiometria e tem como principal objetivo prever variáveis contínuas, quantitativas, utilizando variáveis dependentes. Esta área deriva da antiga da calibração univariada, uma metodologia simples, fácil de aplicar e precisa somente de uma única variável, contudo, este método antigo depende que essa única variável seja completamente relacionada a variável dependente e que não sofra interferência de outras moléculas químicas. Quando falamos de misturas complexas como petróleo,³⁵ Azeite,³⁶ café³⁷ e entre outros,^{38,39} isso é praticamente impossível de ocorrer, necessitando assim da utilização de uma calibração que utilize mais variáveis, Regressão Multivariada.

Na regressão multivariada existe diversas metodologias que são possíveis aplicar, entre elas; O **PCR**,⁴⁰ que transforma a fonte analítica em componentes principais e depois aplica regressão, removendo assim a alta correlação entre algumas variáveis dependentes, contudo, apesar de simples este método tem a desvantagem de desconsiderar a variável

que deseja prever no cálculo. A **regressão linear múltipla** (MLR, do inglês *Multiple Linear Regression*),⁴¹ este método é a ampliação da regressão linear simples para uma versão com maior número de variáveis, com as modificações matemáticas e estatísticas necessárias. Este método tem a desvantagem de depender que haja maior número de amostras do que de variáveis utilizadas, problema de colinearidade, pré-requisito difícil de se obter com as fontes analíticas tradicionais como; MIR,⁴² NIR,⁴³ RMN ¹H⁴⁴ e ¹³C⁴⁵ que ultrapassam, normalmente, três mil variáveis.

2.3.1. PLS

A Regressão pelo método dos quadrados mínimos parciais (PLS)²⁵ é o método mais utilizado da quimiometria, devido a sua alta generalização e facilidade em resolver problemas de regressão, apesar de conseguir ser adaptado para modelo não lineares através de métodos estatísticos e matemáticos. 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 outro. A matemática do PLS começa com o seguinte conjunto de equações;

$$X = T_A P_A^T + E \quad \text{Equação 4}$$

$$y = U_A q_a^T + f \quad \text{Equação 5}$$

Onde **T** e **U** são os scores do modelo, **P** e **q** os loadings obtidos, **E** e **f** os resíduos e **A** o número de variáveis latentes.²⁵ 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 5**.⁴⁶

$$T_A = X \cdot W_A \quad \text{Equação 6}$$

Com o **W**, os coeficientes de regressão do PLS podem ser calculados utilizando a **Equação 6**.

$$b_{PLS} = W_A (P_A^T W_A)^{-1} \hat{q} \quad \text{Equação 7}$$

Com o coeficiente de regressão do PLS podemos resumir a modelagem da seguinte maneira, **Equação 7**, onde as variáveis independentes são preditas com base na variável dependente.

$$y_{pred} = X.b_{PLS} \quad \text{Equação 8}$$

Lembrando que a variável latente deve ser otimizada.⁴⁷

2.3.2. PLS e suas aplicações

O PLS é conhecido como o método mais aplicado na quimiometria, o seu uso não é de hoje, como demonstra o primeiro artigo utilizando Petróleo e PLS, publicado em 1991,⁴⁸ que aplicou o método aliado a radiação ultravioleta para conseguir prever a quantidade de asfalto em petróleo pesado, conseguindo bons resultados. No caso da matriz petróleo o conhecimento da concentração de asfalto é de suma importância, este grupo de compostos são intrinsicamente relacionados a viscosidade do óleo fazendo com que a indústria precise modificar o processo de refino dependendo da sua concentração.⁴⁹

Atualmente, na mesma matriz, Long J. *et al*, em 2019,⁵⁰ analisou a viabilidade de um novo modelo de aparelho de coleta de infravermelho próximo e para confirmar sua capacidade espectroscopia analisou amostras de petróleo com o intuito de prever densidade API e teor de enxofre, os resultados se mostraram promissores para ambas as propriedades físico químicas, o suficiente para serem aplicados na indústria. A Densidade API é de suma importância na indústria do petróleo, com essa propriedade é possível determinar qual é o produto final do refino que melhor se adequa ao tipo de Petróleo. O Teor de Enxofre também tem importância, afinal, altos índices de enxofre são venenosos para catalisadores de refino de petróleo e para motores de carro.

Uma das formas de aperfeiçoar um modelo é aplicando seleção de variáveis e com

o PLS isso não é diferente, como Paulo E. H. et al, em 2022, demonstrou ao aplicar oito tipo de seleção de variáveis em RMN C^{13} para predizer poder calorífico em petróleo, a necessidade dessa aplicação é justificada ao analisar o modelo PLS puro, R^2c de 0,9326 e R^2p de 0,4809, um resultado muito longe do ideal. Ao utilizar otimização por partícula de exames para selecionar 701 variáveis de 61606 o PLS conseguiu obter um R^2c de 0,9953 e um R^2p de 0,9698 uma melhora considerável em comparação ao modelo sem seleção de variáveis.

O PLS tem alta aplicabilidade na matriz Petróleo, como Moro K. M. et al, em 2021,⁵¹ demonstrou no seu artigo de revisão, que analisou determinação de propriedades físico químicas do petróleo utilizando infravermelho e ressonância magnética nuclear aliado a ferramentas quimiométricas. Contudo, o PLS não se aplica bem somente nessa matriz, também podemos utilizar espectroscopia raman em leite, como Du Y. et al, em 2021,⁵² demonstraram ao utilizar PLS e PLS com seleção de variáveis para conseguir determinar enterotoxina B provinda de *Staphylococcus aureus*, essa bactéria é responsável por mais de 200 mil casos de envenenamento alimentar por ano nos EUA, então é de suma importância que sua detecção seja rápida e precisa. Neste trabalho o autor conseguiu um R^2p de 0,9438 e um RMSEP de 0,0994 $\mu\text{g/ml} \cdot 10^{-6}$, o parâmetro foram considerados bons pelos autores.

Ainda no ramo alimentício Folli G. S. et al, em 2022,⁵ utilizou infravermelho portátil, um novo tipo de modelo de espectroscópico que permite análises em campo, para conseguir identificar adulterante em azeite e mel. No caso dos azeites foram testados cinco adulterantes sendo o que obteve o pior resultado foi o óleo de canola com RMSEP 4,4 wt% e R^2p 0,90 enquanto o melhor resultado foi óleo de milho com RMSEP 2,1 wt% e R^2p 0,97, dois resultados bons. Já no mel foram testados três adulterantes, o pior resultado foi com o adulterante glucose com RMSEP de 1,5 wt% e R^2p 0,80 e o melhor

com néctar RMSEP 0,57 wt% e R^2_p 0,98, resultados promissores.

Como dito anteriormente, PLS é um método de regressão linear, o que faz com que tenha baixo desempenho quando lidamos com matrizes não lineares, contudo, isso pode ser contornado com pequenas modificações como Wang Y. et al, em 2015,⁵³ demonstrou ao aplicar diversas modificações de linearidade para aperfeiçoar o modelo. Neste artigo os autores utilizaram amostras de carne em NIR para predizer três propriedades físico químicas e análise de MIR para medir a concentração de gases em gás produzido em refinaria. No estudo das amostras de carne a aplicação de aperfeiçoamento de matrizes para não linear conseguiu o melhorar o modelo em ambas três propriedades, vale destacar que no caso da Umidade, o R^2_p do PLS puro ficou em 0,7554 enquanto o melhor modelo não linear ficou em R^2_p de 0,9783, uma melhora considerável. Na análise das concentrações de gases, os modelos não lineares se destacaram novamente como os melhores, vale ressaltar a determinação de CO, monóxido de carbono, cujo o modelo tradicional conseguiu R^2_p de 0,9341 e o melhor não linear conseguiu 0,9660. Neste artigo ficou demonstrado que o PLS tem alta capacidade de conseguir resolver problemas não lineares ao ser adaptado corretamente.

2.4. Programação

A Programação é um dos conhecimentos necessários para um químico conseguir aplicar a metodologia quimiométrica, como já foi enfatizado neste trabalho. O conhecimento de programação, mesmo o básico, é de suma importância para o químico não só apertar botões e sim saber e compreender o que está sendo feito no computador.

Como os métodos quimiométricos são metodologias que envolvem estatística e matemática, ou seja, diversos cálculos numéricos, é interessante utilizar o software adequado, como o Matlab. Este programa visa justamente o que é necessário para a aplicação da quimiometria, de forma fácil de usar e com funções já embutidas, tornando-se um dos programas mais tradicionais.

O Matlab, foi desenvolvido na década de 70 para fazer cálculos lineares simples e chegou aos tempos atuais podendo ser utilizado para controle de sistemas, aprendizagem de máquina, robótica, tratamento de sinal e outras aplicações. Sua linguagem de programação foi desenvolvida misturando C, C++, Fortran, Java e Python, de uma forma para que leigos possam facilmente aplicar, além de disponibilizar o serviço de nuvem e tutorias para aprendizado.⁵⁴ Apesar de todas essas vantagens o Matlab apresenta uma grande desvantagem, é um programa pago, pensando nisso surgiu algumas versões gratuitas.

O Octave é um software livre que surgiu no fim da década de oitenta com intuito de fazer computação matemática, espelhando-se no Matlab, o software segue uma linguagem semelhante e é quase completamente compatível com as mesmas funções. A principal desvantagem do Octava está em sua velocidade de cálculo em comparação ao software pago, contudo, ainda é uma boa alternativa para os pesquisados.

A programação pode ser aplicada em dois níveis diferentes, os químicos que tem o conhecimento básico para aplicar as ferramentas e funções necessárias e os químicos que tem um conhecimento mais avançado e conseguem desenvolver rotinas, ferramentas e funções para facilitar o próprio trabalho e auxiliar outros químicos. Como o conhecimento de programação não é um pré-requisito para se tornar químico, ter rotinas, ferramentas e funções de fácil compreensão, acesso e aplicação são de suma importância para divulgação da quimiometria. Com este problema em mente, surgiu este estudo.

3. OBJETIVOS

3.1. Objetivo Geral

Desenvolver um tutorial simples e de fácil compreensão para a aplicação passo-a-passo de PLS com a utilização dos Software Octave, com aplicabilidade para o Matlab, visando dados químicos e voltado para graduação ou superior. Com intuito de facilitar e divulgar a quimiometria no Brasil.

3.2. Objetivos Específicos

- Selecionar o melhor conjunto de dados para o tutorial.
- Desenvolver as funções de forma a ser viável em Octave e Matlab.
- Criar uma rotina computacional que possa ser usada no Octave e no MatLab.
- Desenvolver o Tutorial.
- Facilitar o aprendizado de Quimiometria.
- Divulgar Quimiometria.

4. PROCEDIMENTO EXPERIMENTAL

4.1. Amostragem

Neste trabalho foram utilizados três conjuntos de dados, dois relacionados a petróleo e um relacionado a adulterantes em alimentos. O primeiro conjunto é para ser usado na Parte 01 e 02, trata-se de um infravermelho médio como matriz X e um vetor y de densidade API de amostras de petróleo. Na Parte 03 foi escolhido um conjunto de infravermelho próximo obtido em um portátil como matriz para prever a quantidade de adulterantes em amostras de azeite. Na Parte 04 um conjunto de infravermelho médio para prever nitrogênio total.

Todas rotinas, funções e conjunto de dados podem ser encontrados no GitHub do autor: <https://github.com/PHPCunha/Monografia.git>

4.1.1. MIR

Os espectros de infravermelho médio foram obtidos utilizando um equipamento Spectrum 400 da PerkinElmer. As análises foram feitas no laboratório de instrumentação em triplicata e com 32 scans, com resolução de 1 cm^{-1} e intervalo de leitura de $4000\text{--}650\text{ cm}^{-1}$. A análise demorou aproximadamente 10 min e foi necessário 1 mL de amostra.

4.1.2. NIR portátil

Os espectros de infravermelho próximo foram adquiridos no equipamento portátil de modelo MicroNIR TM Pro 1700 with software version 3.0 from Viavi Solutions Inc. Com o comprimento de onda de $908\text{--}1676\text{ nm}$, com 100 scans de varredura e uma distância de 6.15 entre pontos. Um vial foi utilizado como acessório para a análise que demorou menos de 5 minutos.

4.1.3. Propriedades Físico-químicas.

Os ensaios das propriedades físico-químicas do petróleo foram obtidos pelo centro de pesquisa Centro de Pesquisas Leopoldo Américo Miguel de Mello (CENPES) e os resultados foram entregues ao Núcleo de Competências em Química do Petróleo da UFES (NCQP/UFES).

A determinação da densidade API foi feita com base na norma ISO 12185,⁵⁵ foi utilizado um viscosímetro digital Anton Paar (modelo Stabinger SVM 3000) com limite de detecção de $0,0002 \text{ g,cm}^{-3}$ a $20 \text{ }^{\circ}\text{C}$. Foi usado um volume de amostra de 5 mL e para determinar a densidade API, os resultados de viscosidade foram convertidos para a densidade equivalente a $20 \text{ }^{\circ}\text{C}$. O Nitrogênio Total foi determinado utilizando a ASTM D4629-17, este ensaio consiste na combustão das amostras, seguida de determinação de nitrogênio utilizando quimiluminescência e curva de calibração.

4.1.4. Azeite

Este conjunto amostral deriva de um artigo já aceito⁵ e foi liberado pelos autores para participar deste estudo. Foram utilizadas 10 marcas diferentes e cinco tipos diferentes de óleo para adulterar o azeite, são eles: Soja, canola, algodão, milho e girassol. As concentrações de adulterante colocada em cada amostras foram 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 m/m. adulterante/azeite. Na separação do conjunto calibração e teste, foi escolhido para conjunto calibração: 0.0, 1.0, 2.0, 4.0, 5.0, 6.0, 8.0, 9.0, 20.0, 40.0 e 50.0 m/m e para teste as concentrações 3.0, 7.0, 10.0 e 30.0 m/m.

4.2. Softwares e Algoritmos

As funções e rotinas foram construídas utilizando o Matlab versão 2013a e posteriormente adaptadas ao GNU Octave (John W. Eaton, versão 7.3.0, 2022), que pode ser gratuitamente obtido no link: <https://octave.org/download>. As rotinas, funções, espectros, pacotes e demais itens estão disponíveis no GitHub do autor (<https://github.com/PHPCunha/Monografia.git>). Recomenda-se abrir a rotina “00 - Instalação de pacotes” caso seja a primeira vez que utilize o Octave, caso não, recomendamos abrir “02 - Tutorial PLS”.

5. RESULTADOS E DISCUSSÃO

5.1. UFES

Na análise dos principais cursos de química do país, 2.1.2. Ensino no Brasil, com base na classificação do MEC, a Universidade Federal do Espírito Santo - UFES aparece na trigésima nona posição, **Tabela 1**, enquanto o nosso curso tem somente a disciplina de Noções de Estatística, ou Introdução à estatística, como optativa para o curso de química bacharelado. Pode até relacionar o posicionamento no ranking com o empobrecimento do ensino de quimiometria, todavia, isso provavelmente é uma correlação sem casualidade, ou seja, uma mera coincidência.

Provavelmente o ranking tenha uma relação com currículo mais encorpado, com muito mais variedades de matérias eletivas e optativas, do que com a ausência da matéria de quimiometria. Neste cenário, torna-se relevante apontar que o currículo do curso precisa de uma atualização, adicionar uma matéria obrigatória chamada de **Estatística aplicada na Química**, uma matéria onde é ensinado estatística aplicadas e necessárias na química, desde algarismos significativos a aplicação de lei de Beer para duas variáveis, como explicado na Revisão Bibliográfica, além de adicionar **Quimiometria** como matéria eletiva, ensinado o conteúdo não supervisionado até o PLS. Com essas duas adições, o curso de Química-Bacharelado de UFES teria maior capacidade de preparar os Químicos, principalmente analíticos, para aplicar e compreender quimiometria.

Analisando o cenário acadêmico, 2.1.3. Academia no Brasil, observou-se um aumento de publicações na área da quimiometria por parte da nossa faculdade, com um crescimento da primeira faixa, 2005 a 2009, para a atual, 2020 a 2023, de 1075% aproximadamente. Além disso, ao analisar a **Tabela 3**, percebe-se que a UFES vem crescendo no número de publicações e crescendo no ranking das posições, o que demonstra um grande esforço dos laboratórios para o desenvolvimento de novos estudos.

Ao analisarmos a **Figura 4**. Gráfico de artigos publicados, com a palavra “chemometric*”, por ano, mundo (a) e Brasil (b). Pesquisado no Web of Science dia 09 de Janeiro de 2023., podemos ver que conforme os anos avançaram houve um crescimento em ambos os gráficos, **a** Mundial e **b** Brasil, nota-se que o crescimento percentual do gráfico nacional (539%) é maior que o internacional (243%), o que aponta um aumento da relevância do Brasil nesta área, saindo de 4,39% dos artigos em 2006

serem brasileiros indo para 8,17% em 2022. Esse cenário é favorável para preparar futuros químicos não só para aplicar os novos estudos, como também desenvolver novos estudos.

5.2. Parte 00 - Instalação de pacotes

O Octave é um software livre, que pode ser gratuitamente baixado (<https://octave.org/download>), desse modo, o primeiro passo é fazer o download do programa. Após isso iremos instalar alguns pacotes de ferramentas para podermos utilizar as rotinas de quimiometria.

O Octave não tem todas ferramentas matemáticas e estatísticas necessárias para a aplicação de um PLS, então temos que baixar e instalar. O Editor é uma subjanela do Octave, como é apresentada na **Figura 6**, nesta subjanela ficam as rotinas que utilizamos, na figura podemos ver as rotinas “00 – Instalação de pacotes” e “02 – Tutorial PLS” que serão utilizadas neste tutorial.

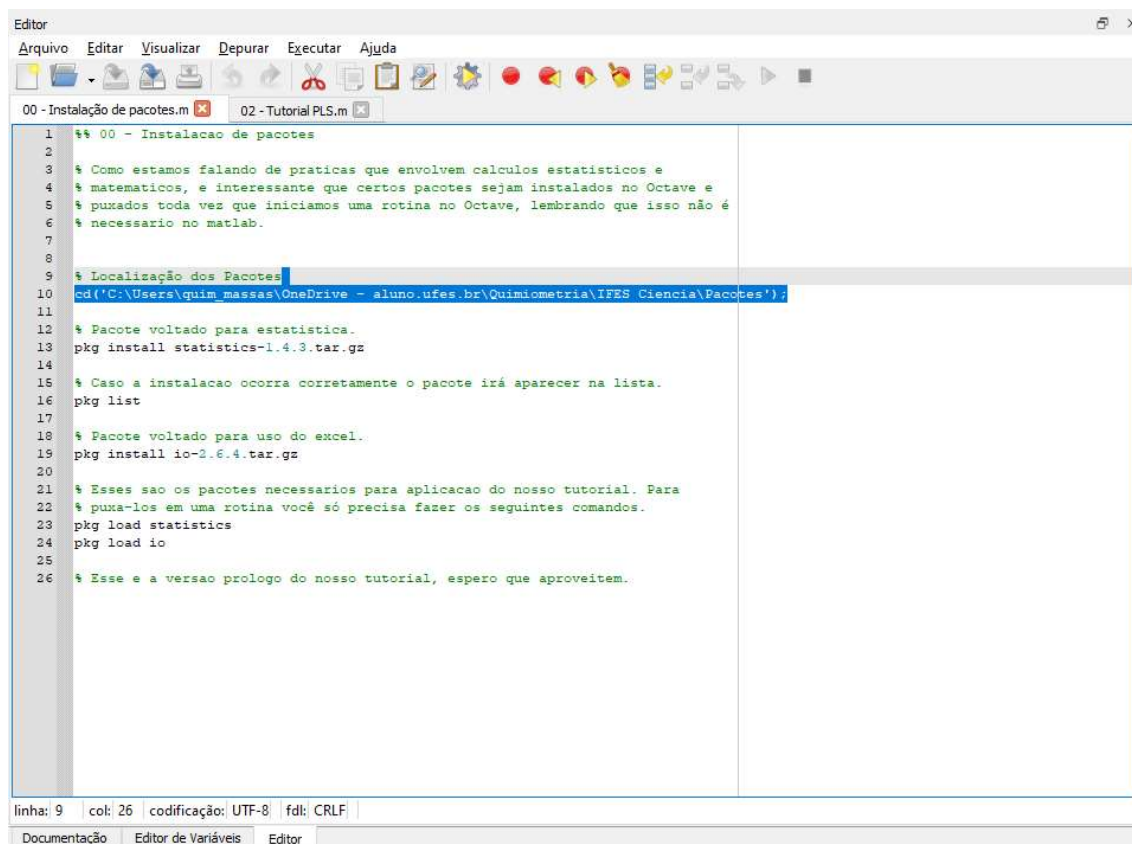


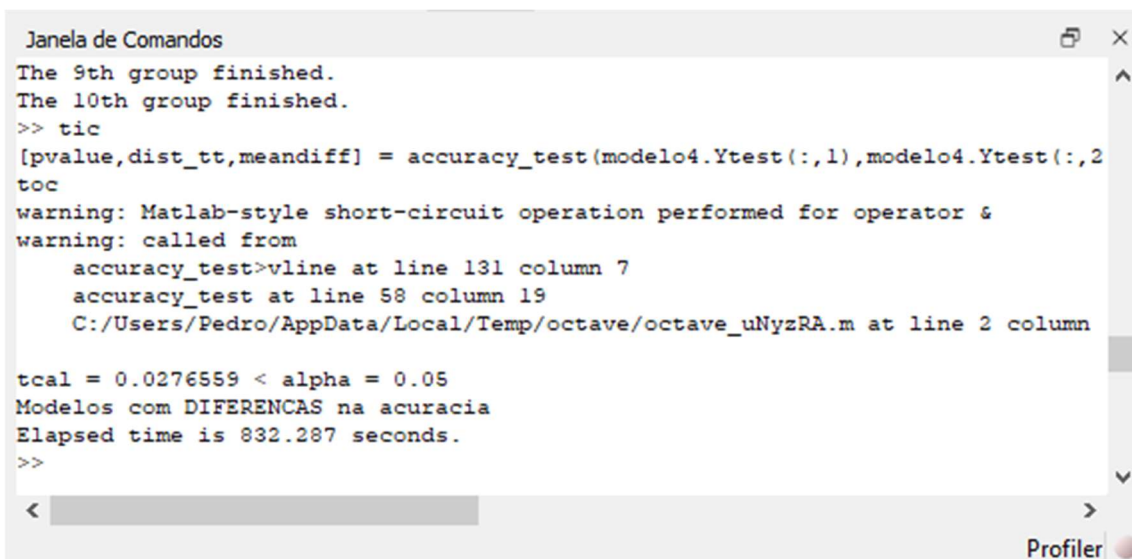
Figura 6. Subjanela Editor no Octave.

Para ativar um comando da rotina basta seleciona-la e apertar o botão F9, assim

como está na **Figura 6**, o primeiro comando é muito simples, o “cd” é utilizado para mudar o diretório, ou pasta, que o Octave/Matlab (OM) está utilizando como referência. (Quando utilizamos uma função, ou carregamos algum dado para dentro do programa, precisamos deixar o OM na pasta que o arquivo se encontra.)

```
>>% Localização dos Pacotes
>> cd('C:\Users\Exemplo\...\Monografia\Pacotes');
```

Ao utilizar o comando acima (todos comandos nesse tutorial serão simbolizados com “>>”), provavelmente não ocorrerá nada, isso porque o diretório não existe, o que causa um erro de localização, que será apresentado na Janela de Comandos, **Figura 7**. Para corrigir isso, substituir o 'C:\Users\Exemplo\...\Monografia\Pacotes' pelo diretório da pasta Pacotes que foi baixada do Gitrib, neste tutorial sempre usaremos a entrada “C:\Users\Exemplo\...” como exemplo para todos diretórios, para fim de referência.



```
Janela de Comandos
The 9th group finished.
The 10th group finished.
>> tic
[pvalue,dist_tt,meandiff] = accuracy_test(modelo4.Ytest(:,1),modelo4.Ytest(:,2)
toc
warning: Matlab-style short-circuit operation performed for operator &
warning: called from
    accuracy_test>vline at line 131 column 7
    accuracy_test at line 58 column 19
    C:/Users/Pedro/AppData/Local/Temp/octave/octave_uNyzRA.m at line 2 column

tcal = 0.0276559 < alpha = 0.05
Modelos com DIFERENCAS na acuracia
Elapsed time is 832.287 seconds.
>>
```

Figura 7. Janela de Comando do Octave.

Detalhe extra, o “%” anula qualquer comando posterior a ele, ou seja, enquanto, “>> 2+2”, resultará em 4, “>> 2%+2”, resultará em 2. É interessante utilizar o “%” para adicionar suas anotações na rotina, pois assim tu não interferes na parte que o OM processa. A seguir, vamos instalar o primeiro pacote, este é voltado para estatística.

```
>> pkg install statistics-1.4.3.tar.gz
```

Para verificar se um pacote foi corretamente instalado, basta utilizar o seguinte comando e verificar sua presença na lista.

```
>> pkg list
```

Em seguida, vamos instalar o pacote voltado para planilhas.

```
>> pkg install io-2.6.4.tar.gz
```

Temos todos pacotes instalados, só tem mais um detalhe (valido somente para quem usa Octave) caso precise utilizar um desses pacotes, precisará ativa-los toda vez que abrir um novo Octave com o seguinte comando:

```
>> pkg load statistics
```

```
>> pkg load io
```

5.3. Parte 01 - Conhecendo o plsmode

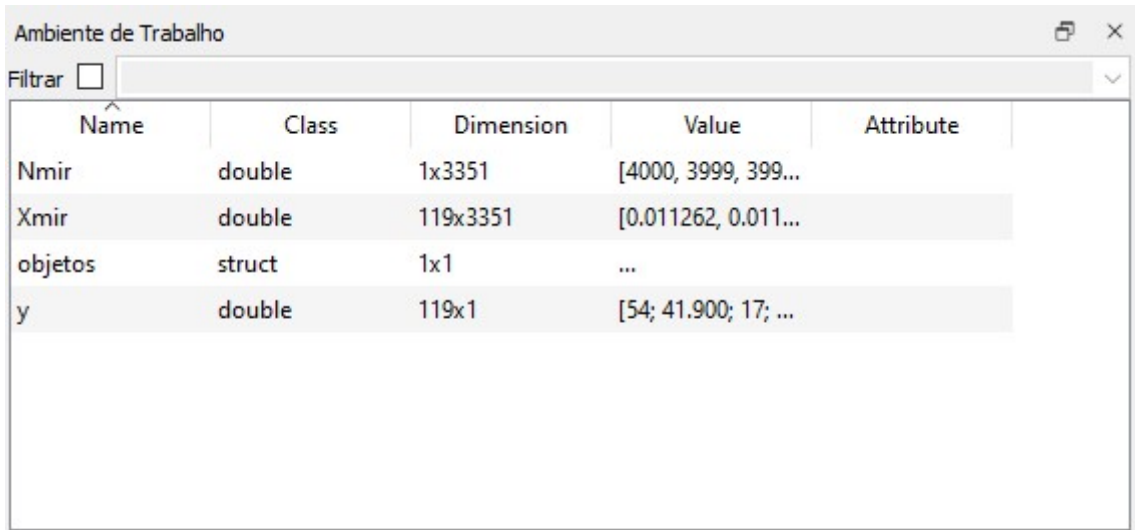
Quando iniciamos uma rotina é recomendado utilizar três comandos:

```
>> clear    % Limpa o Ambiente de Trabalho/Workspace
```

```
>> clc      % Limpa o Editor/Command Window
```

```
>> close all % Fecha qualquer imagem aberta.
```

Como detalhado, esses comandos irão limpar a subjanela Editor, Command Windows no Matlab, e o Ambiente de Trabalho, Workspace no Matlab. O Ambiente de Trabalho é o local onde fica informações carregadas pelo OM, **Figura 8**.



The screenshot shows the 'Ambiente de Trabalho' (Workspace) window in Octave. It contains a table with the following data:

Name	Class	Dimension	Value	Attribute
Nmir	double	1x3351	[4000, 3999, 399...	
Xmir	double	119x3351	[0.011262, 0.011...	
objetos	struct	1x1	...	
y	double	119x1	[54; 41.900; 17; ...	

Figura 8. Ambiente de Trabalho do Octave.

Em seguida, caso esteja utilizando o Octave, caso seja Matlab ignore, vamos ativar os pacotes extras;

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

A primeira coisa que temos que fazer é mudar o endereço do programa e exportar os dados que usaremos na primeira lição;

```
>> cd('C:\Users\Exemplo\...\PLS_Model');
>> load('Dados_API.mat')
```

Com o primeiro comando estamos direcionando o software para a pasta desejada, onde esteja as nossas funções e amostras, no segundo comando usamos a função “load” para exportar o arquivo que contém o nosso conjunto amostral.

No “Dados_API.mat” temos 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 amostras, pode-se utilizar os seguintes comandos;

```
>> plot(Nmir,Xmir);
>> xlabel("Comprimento de Onda")
```



```
>> ylabel("Abs");
>> set(gca,'FontSize',16);
```

Com estes comandos a seguinte subjanela deve aparecer automaticamente, **Figura 9**. Podemos ver que não existe uma amostra aparentemente anômala, todas tem o mesmo perfil espectral.

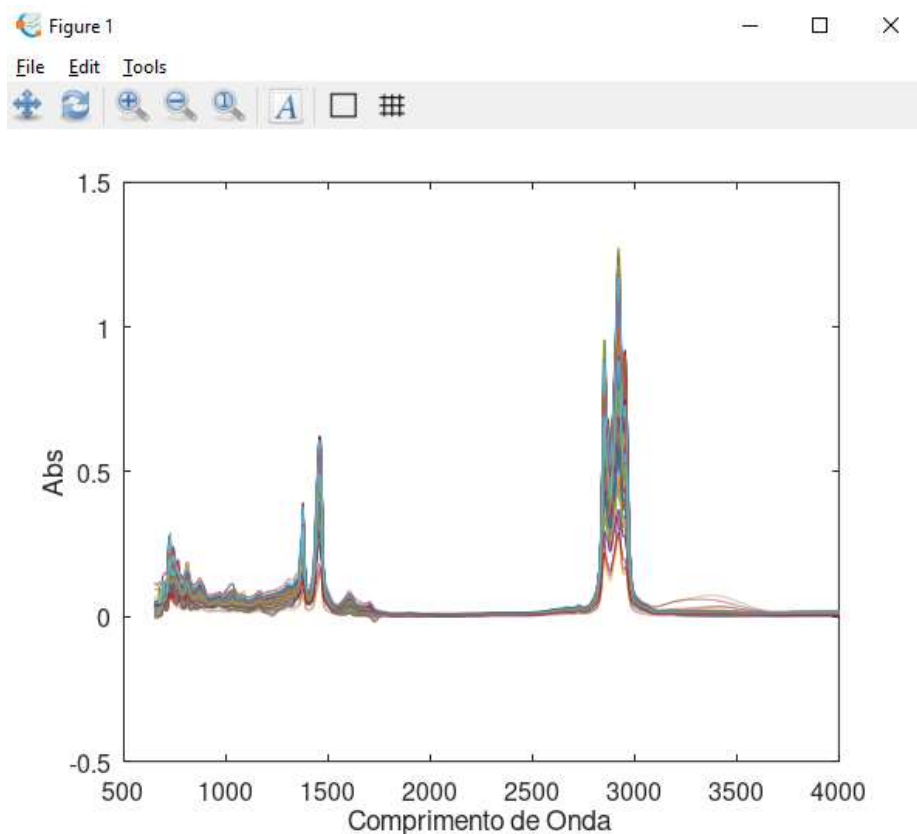


Figura 9. Espectro das amostras de densidade API.

Como desejamos desenvolver um modelo de PLS, precisaremos começar separando as amostras em conjunto calibração e teste.

```
>> 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 = [];
>> options.Xpretreat = {'center'};
>> options.vene = 5;
>> options.vl = 20;
```

A primeira linha serve para garantir que o arquivo esteja vazio, na segunda definimos qual será o pré-tratamento interno, caso queira utilizar nenhum, coloque “{‘none’}” no lugar, na terceira linha definimos o tamanho da janela de calibração cruzada e na última linha a quantidade de variáveis latentes que iremos usar na otimização, neste caso usaremos 20 para criar o gráfico RMSECV x LV. Para verificar se o “options” está programado da forma desejada, vai na subjanela “Ambiente de Trabalho” e dá duplo clique no “options”, a tela principal será transferida para a subjanela “Editor de Variáveis”, ‘Variables’ no Matlab, e poderá visualizar o que dentro do arquivo, como apresentado na **Figura 10**.

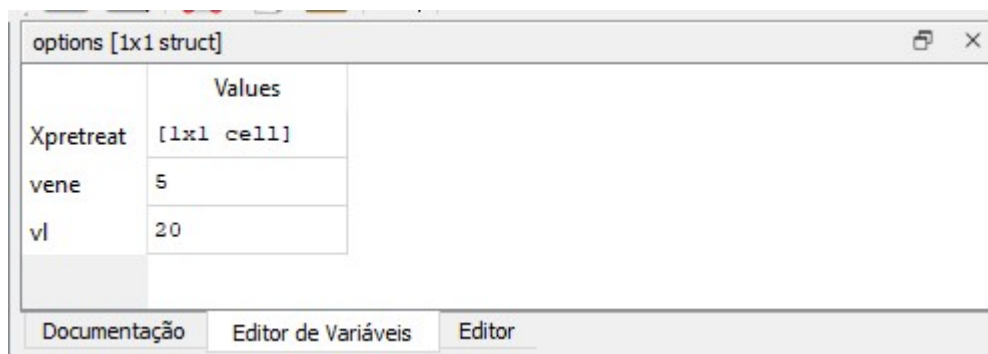


Figura 10. Editor de Variáveis do Octave.

Com este “options” podemos rodar o PLS no modo otimização, na forma que é apresentado abaixo;

```
>> modelo=plsmodel2(Xcal,ycal,options)
```

Ao realizar este comando software fara os cálculos de otimização e a seguinte imagem irá abrir, **Figura 11**, este é o gráfico RMSECV por LV, nele 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.

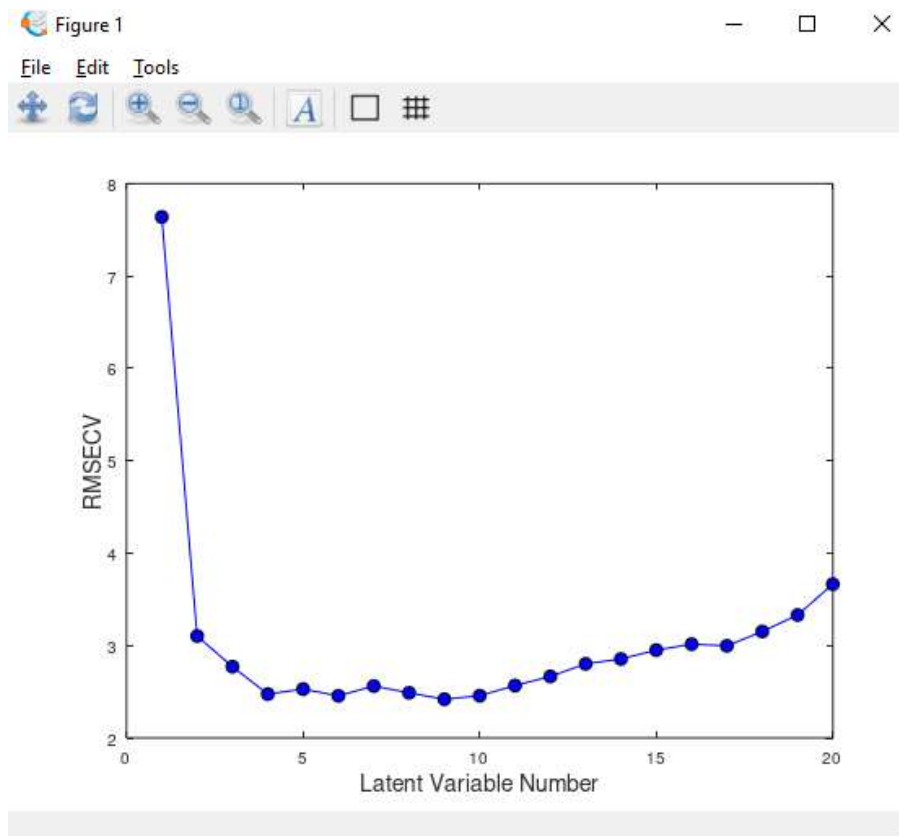
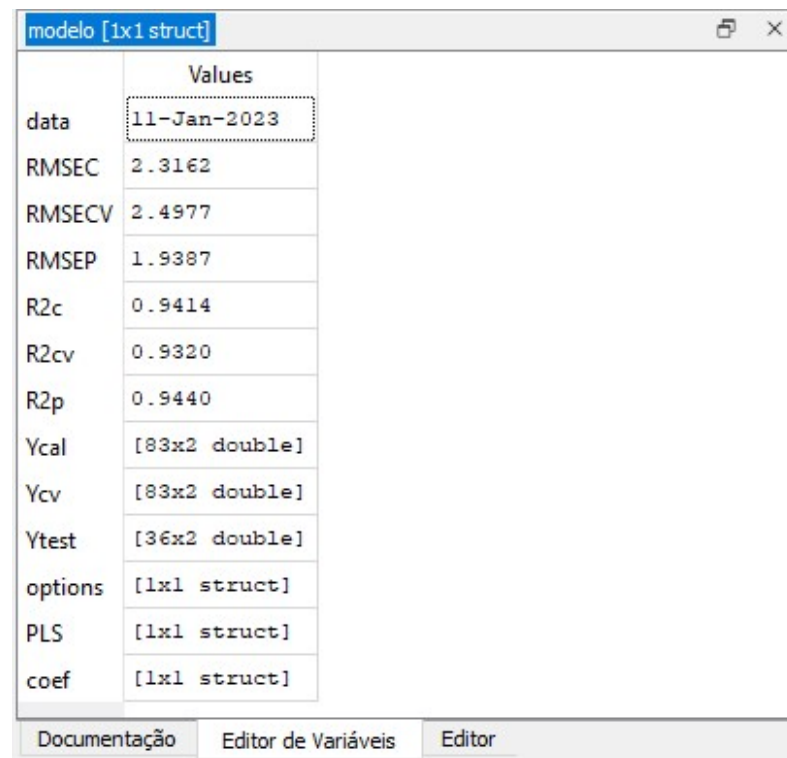


Figura 11. Gráfico de RMSECV x LV

Quando utilizamos uma função temos inputs, que são arquivos de entrada, e outputs, que são arquivos de saída. Na última linha de comando utilizada, a que criou a **Figura 11**, nós usamos a função “`plsmodel2`” que tem três inputs, “`Xcal`”, “`ycal`” e “`options`” e tem como output “`modelo`”. Usando uma analogia para fins didáticos, podemos imaginar que os inputs são reagentes, os outputs são produtos e a função são as ferramentas, métodos e condições necessárias para a reação ocorrer. Nesse modo, entregamos os inputs para a função e ela retorna os outputs. Continuando a lição, escolhendo o número 4 de LV, utilizamos os seguintes comandos;

```
>> options.vl = 4;
>> modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);
```

Note, que agora estamos utilizando não só as amostras de calibração (`Xcal` e `ycal`) como inputs, mas também as de teste (`Xtest` e `ytest`), além do “`options`”, com isso a função não irá otimizar o modelo e sim criar um modelo de PLS com ambos conjuntos. Após o Octave parar o cálculo, vamos dar duplo clique no input “`modelo`”.



	Values
data	11-Jan-2023
RMSEC	2.3162
RMSECV	2.4977
RMSEP	1.9387
R2c	0.9414
R2cv	0.9320
R2p	0.9440
Ycal	[83x2 double]
Ycv	[83x2 double]
Ytest	[36x2 double]
options	[1x1 struct]
PLS	[1x1 struct]
coef	[1x1 struct]

Figura 12. Estrutura interna no modelo, no Octave.

Dentro da estrutura do “modelo” temos tudo que iremos precisar para avaliar e utilizar neste modelo PLS. Começamos analisando os Parâmetros de Avaliação, temos um RMSEC de 2,3, um RMSEP de 1,9, R^2c de 0,9414 e R^2p de 0,9440. Essas métricas são boas para os parâmetros da regressão multivariada, contudo, dá para conseguir modelos melhores? Vamos testar o modelo com 5 LV.

```
>> options.vl      = 5;
>> modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);
>> modelo.R2c; % 0.9537
>> modelo.R2p  % 0.9500
```

Com esse conjunto de comando o cálculo terminou com “ans = 0.9500”, isso ocorreu porque na última linha eu pedi para ele me informar o valor contido no “modelo.R2p” e não coloquei “;” no final, ao fazer isso o OM entende que eu quero saber o valor daquele arquivo. Sobre o modelo, percebe-se que ele apresentou uma melhora nos parâmetros de avaliação, vamos testar agora a LV 6.

```
>> options.vl      = 6;
>> modelo = plsmodel2(Xcal,ycal,Xtest,ytest,options);
>> modelo.RMSEC    % 1.9061
>> modelo.RMSEP    % 1.4480
>> modelo.R2c      % 0.9614
>> modelo.R2p      % 0.9582
```

Agora o Octave irá informar os quatro parâmetros de avaliação, o modelo com 6 LV tem parâmetros melhores que os demais, mas isso não é o suficiente para confirmarmos que ele é o melhor. Vamos comparar, para fins didáticos, os modelos LV 4 e 6 e iremos utilizar os seguintes comandos;

```
>> options.vl      = 4;
>> modelo4 = plsmodel2(Xcal,ycal,Xtest,ytest,options);
>> options.vl      = 6;
>> modelo6 = plsmodel2(Xcal,ycal,Xtest,ytest,options);
```

Nestas linhas de comando em vez de utilizar o nome “modelo” para o arquivo preferi utilizar dois nomes diferentes, desse modo, os dois ficam armazenado na subjanela “Ambiente de Trabalho”. Agora, vamos avaliar os modelos em conjunto utilizando o “Gráfico de Medido x Predito”, para isso fazemos o seguinte comando;

```
>> close all %Fechando imagens ja 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','Location','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','Location','southeast');  
>> title('Modelo 6');  
>> set(gca,'FontSize',12);xlabel('Reference','fontsize',12);  
>> ylabel('Predicted','fontsize',12);
```

Ao utilizar essas linhas de comando, todas juntas, o resultado é a **Figura 13**, peça que manualmente ajuste o tamanho da janela para que os gráficos fiquem quadrados, isso facilita a interpretação.

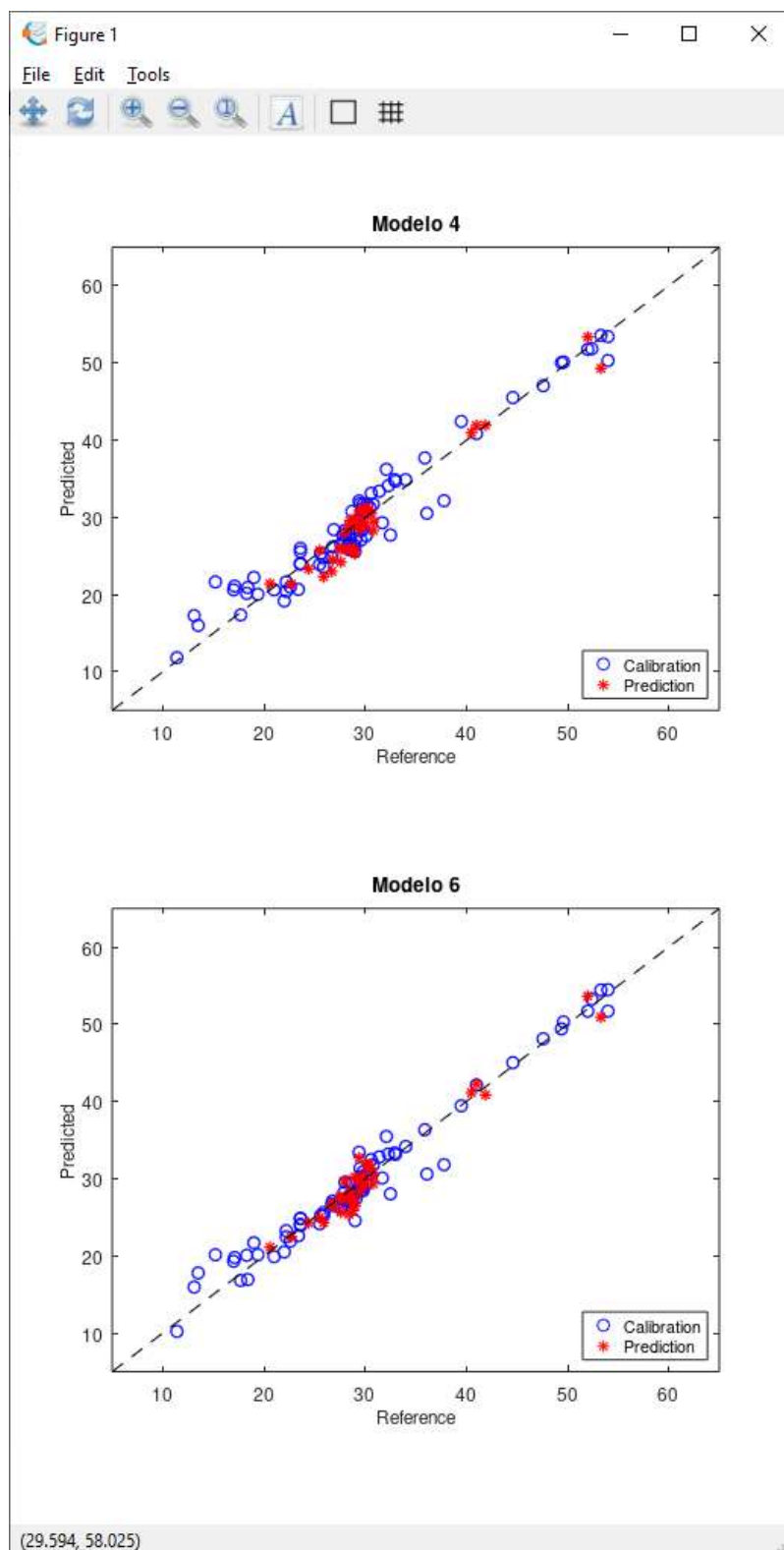


Figura 13. Gráfico de Medido x Predito.

Esta análise, como já explicado, 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 13** podemos perceber que as amostras, em sua maioria, ficaram bem próximas da linha de tendência, o que é um bom indicador, repara-se que ambos modelos tiveram o mesmo comportamento na faixa 5 a 37, e o modelo 6 obteve uma melhor performance após essa faixa, contudo, só isso não é o suficiente para afirmarmos que o modelo 6 é melhor que o 4.

Para confirma que o modelo 6, vamos utilizar uma comparação entre modelos, neste caso o teste randômico de exatidão, no caso a função “accuracy_test”. Esta função compara modelos multivariados, utilizando teste randômicos, para verificar a diferença estatística dos modelos, utilizaremos os seguintes comandos;

```
>> tic
>> [pvalue,dist_tt,meandiff] =
accuracy_test(modelo4.Ytest(:,1),modelo4.Ytest(:,2),modelo6.Ytest(:,2),'randbi',500000
,0.05);
>> toc %{1068 sec }
```

Aqui utilizamos a função tic/toc, uma função e dois comandos, esta função informa o tempo, em segundos, entre o comando tic e o toc, interessante para saber o tempo de processamento, em um teste utilizando um computador pouco potente, essa função demorou aproximadamente 17 min, então, não estranhe caso demore este tempo.

A função “accuracy_test” foi programada para dar a resposta assim que termina os cálculos, neste caso “Modelos com DIFERENCAS 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
```

Todas funções desenvolvidas neste estudo vem com uma explicação mais elaborada na parte interna. Este é o final da Pater 01 do tutorial, nessa parte o foco era aprender o básico da função plsmodel2 e como avaliar um modelo, na próxima etapa vamos aprender como aprimorar o modelo utilizando os mesmos dados dessa parte. Recomenda-se que salve o “Ambiente de Trabalho” utilizando o seguinte comando;


```
>> save("Parte 01");
```

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”.

5.4. Parte 02 - Aprimorando o modelo.

Nesta parte iremos aprender como aprimorar um modelo, mais especificamente, utilizando pré-tratamento nas fontes analíticas e depois como identificar outlier, um método de avaliação de modelo que também pode aprimora-lo. Usaremos o mesmo conjunto de dados utilizados na Parte 01.

```
>> load("Parte 01");
```

```
>> clearvars -except ycal ytest Xcal Xtest objetos Nmir
```

A função ‘clearvars’ combinada com ‘-except’ deleta todos arquivos internos do Ambiente de Trabalho exceto os nomeados em seguida, é bom utiliza-lo para não haver confusão de arquivos. Como dito anteriormente, o pré-tratamento é um método de aperfeiçoamento e com isso tem o objetivo de melhorar os parâmetros do modelo, neste caso é feito um aprimoramento na fonte analítica de modo a evidenciar informações relevantes e ofuscar informações irrelevantes, como ruídos. Para isso utilizaremos o seguinte comando;

```
>> [Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'auto'});
```

Nessa função temos três inputs (Xcal, Xtest e method) e dois outputs (Xcal2 e Xtest2). O method escolhido, no caso do último comando, foi {'auto'}, mas poderia ser: {'center'}, {'snv'}, {'msc'} e {'deriv',[7,2,1]}. Tome cuidado ao utilizar pré-tratamentos, pois pode acabar destacando ruídos e escondendo informações. Recomendo sempre adicionar “2” nos espectros tratados, pois assim se mantém a fonte analítica original intacta. Agora, vamos utilizar os seguintes comandos para ver o nosso espectro tratado;

```
>> subplot(2,1,1)
>> plot(Nmir,Xcal);
>> subplot(2,1,2)
>> plot(Nmir,Xcal2);
```

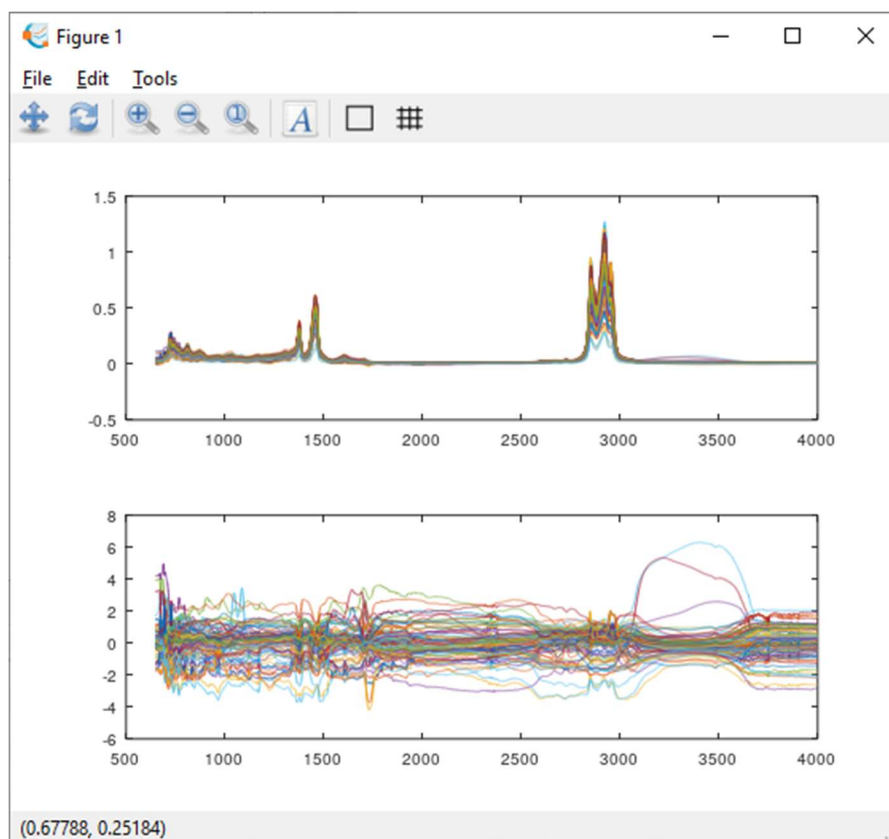


Figura 14. Gráfico bruto e tratado.

Como podemos ver na **Figura 14**, o perfil espectral das amostras se modificou completamente. O Autoescalamento pode ser aplicado em espectros de infravermelho, contudo é preferível utilizar em fontes analíticas de dimensões diferentes, o que não é o caso. No infravermelho é recomendado utilizar primeira e segunda derivada, da forma abaixo:

```
>> [Xcal2,~]=pretrat(Xcal,Xcal,{'deriv',[15,2,1]});
```

Diferentes de outros ‘method’ a derivada precisa de um segundo input acompanhando, o primeiro número represente o número de janelas que serão

consideradas para fazer a derivada, o segundo número o grau do polinômio, ou grau da derivada, e o terceiro a ordem da derivada. No caso do último comando temos, uma janela de 15 variáveis, segundo grau de polinômio e primeira derivada. Nota-se também que eu utilizei somente o 'Xcal' como input, essa função foi criada para tratar dois conjuntos de fontes ao mesmo tempo, caso queira tratar somente uma, irá precisar usar esse pequeno truque, mas lembre-se o conjunto teste SEMPRE deve ser tratado no segundo input e com o conjunto calibração no primeiro.

```
>> close all  
>> subplot(2,1,1)  
>> plot(Nmir,Xcal);  
>> subplot(2,1,2)  
>> plot(Nmir,Xcal2);
```

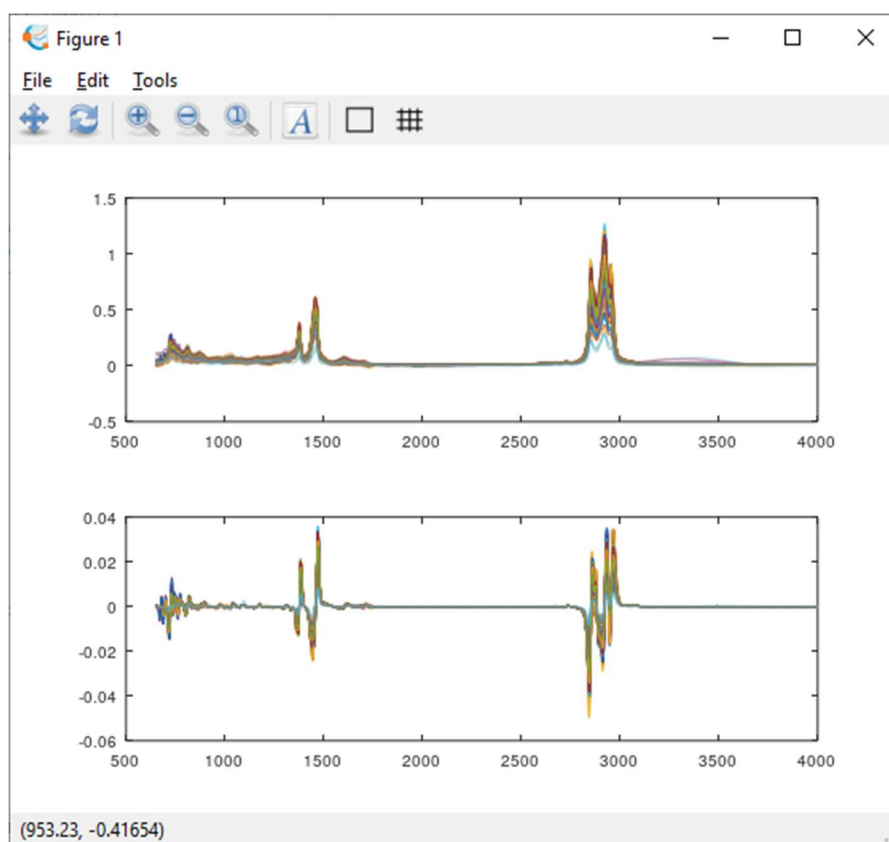


Figura 15. Gráfico bruto e tratado com derivada.

Na **Figura 15** podemos ver que a derivada fez uma diferença significativa no

perfil espectral destacando os pontos do espectro de maior absorbância e alisado o ruído do espectro. Vamos verificar se esta mudança aprimorou o nosso modelo;

```
>> options=[];
>> options.Xpretreat = {'center'};
>> options.vene      = 5;
>> options.vl        = 20;
>> modelo=plsmodel2(Xcal2,ycal,options);
```

Ao utilizar estes comandos se percebe que o gráfico de RMSECV e LV é feito por cima da última imagem, isso ocorre porque não especificamos a janela que iríamos fazer a figura e nem fechamos o anterior. Todavia, isso não atrapalha na nossa decisão de LV, o 6 e 8 parecem os mais promissores, nesse modo, vamos testar dois modelos um com cada LV, antes disso, precisamos tratar o conjunto de teste.

```
>> [Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});
>> options.vl      = 6;
>> modelo6 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);
>> options.vl      = 8;
>> modelo8 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);
```

Ao analisar os modelos, percebemos que o R^2_p e RMSEP de ambos ficaram bem próximos, o modelo com 8 LV obteve R^2_p 0,9727 e RMSEP 1,4 enquanto o modelo com 6 LV 0,9759 e 1,4, então escolhemos o modelo com LV 6 pelo princípio da parcimônia, um modelo mais simples, com uma menor quantidade de LV, é melhor que um modelo maior, com uma quantidade maior de LV.

```
>> clear modelo8
```

Quando se é especificado um arquivo após o comando ‘clear’ só este arquivo é apagado. Além disso, não precisamos nós limitar a utilizar somente um único pré-

tratamento, podemos fazer uma combinação dos tratamentos dominados. Vamos testar:

```
>> [Xcal2,Xtest2]=pretrat(Xcal,Xtest,{ 'deriv',[15,2,1]});
>> [Xcal2,Xtest2]=pretrat(Xcal2,Xtest2,{ 'msc'});

>> options=[];
>> options.Xpretreat = {'center'};
>> options.vene      = 5;
>> options.vl        = 20;
>> modelo=plsmodel2(Xcal2,ycal,options);
```

Vamos testar o LV 9;

```
>> options.vl        = 9;
>> modelo9 = plsmodel2(Xcal2,ycal,Xtest2,ytest,options);
```

Com base nos parâmetros de avaliação, R^2_p de 0,9816 e RMSEP de 1,1, podemos considerar o modelo com LV 9 e duplo tratamento melhor que o outro modelo, todavia, como já dito, só isso não é o suficiente para a confirmação, nesse modo, vamos analisar o gráfico de medido x predito.

```
>> close all %Fechando imagens já criadas.
>> subplot(2,1,1)
>> 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','Location','southeast');
>> title('Modelo 6');
>> set(gca,'FontSize',12);xlabel('Reference','fontsize',12);
>> ylabel('Predicted','fontsize',12);
```

```
>> subplot(2,1,2)
>> 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([5 65]); xlim([5 65]);
>> plot(xlim, ylim, '--k'); legend('Calibration','Prediction','Location','southeast');
>> title('Modelo 9');
>> set(gca,'FontSize',12); xlabel('Reference','fontsize',12);
>> ylabel('Predicted','fontsize',12);
```

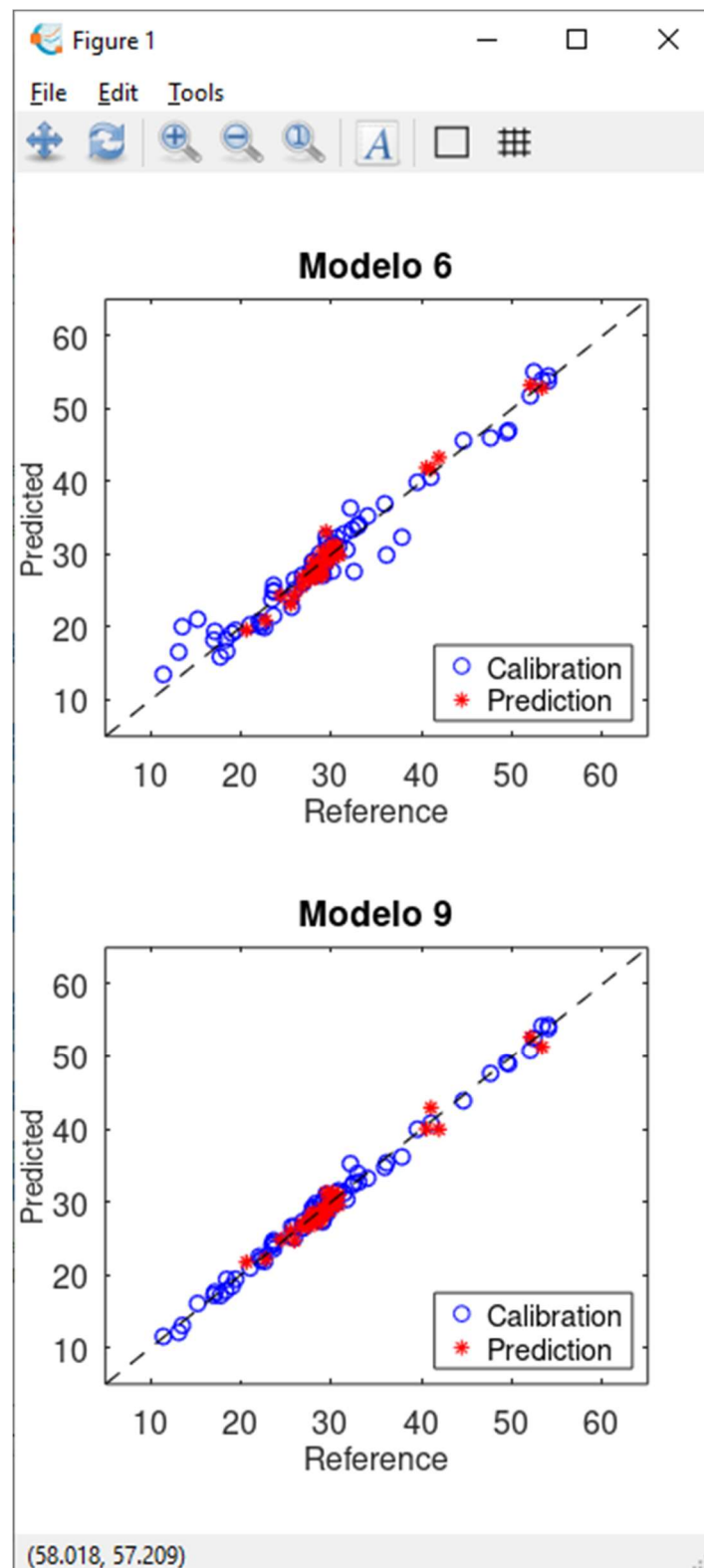


Figura 16. Gráfico de Medido x Predito do modelo 6 e 9.

Quando analisamos os gráficos percebemos que o modelo 9 conseguiu amostras mais próximas da linha de referência, principalmente na faixa 10 a 30. Vamos fazer o teste randômico de exatidão;

```
>> tic

>> [pvalue,dist_tt,meandiff] =
accuracy_test(ytest,modelo6.Ytest(:,2),modelo9.Ytest(:,2),'randbi',500000,0.05);

>> toc % 890 Sec
```

O resultado foi estatisticamente semelhante, não podemos dizer que o modelo 9 é o melhor, todavia, podemos escolher ele como o melhor modelo. Será que ainda dá para melhorar? Vamos verificar se existe algum Outlier no modelo 9. Os Outlier, ou amostras anômalas, podem ocorrer por diversos motivos, seja má calibração do aparelho, seja anotar os dados com sono, dessa forma, elas não representam o que deveriam e muitas vezes distorcem o modelo, dessa forma, torna-se interessante remove-las.

```
>> clearvars -except ycal ytest Xcal Xtest objetos Nmir modelo9 modelo6

>> [Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});

>> [Xcal2,Xtest2]=pretrat(Xcal2,Xtest2,{'msc'});

>> modelo9 = lev_res(modelo9,Xcal2,ycal,Xtest2,ytest);

>> close all
```

Adicionei um “close all” no final do comando para podermos analisar as duas métricas de forma isolada primeiro. A função `lev_res` calcula o leverage e o resíduo do modelo, o ideal é analisar ambos em conjunto, todavia compreender como cada um funciona é fundamental.

O Leverage, não tem uma tradução própria, trata-se de uma métrica que determinar a influência de uma amostra na construção do modelo de regressão. Um baixo leverage significa pouca influência e alto, grande influência. Quando uma amostra é anômala, a tendência é que ela tenha um alto valor de leverage e assim se destacar neste teste. Vamos analisar o gráfico Leverage isoladamente;

```
>> figure(1)

>> plot(1:size(modelo9.lev_res.lev_cal,1),modelo9.lev_res.lev_cal,'bo'); hold
```


on;

```
>> plot(1:1:size(modelo9.lev_res.lev_test,1),modelo9.lev_res.lev_test,'r*'); hold
```

on;

```
>> hline(modelo9.lev_res.lev_limite,'k');
```

```
>> set(gca,'FontSize',16);
```

```
>> title('Leverage');
```

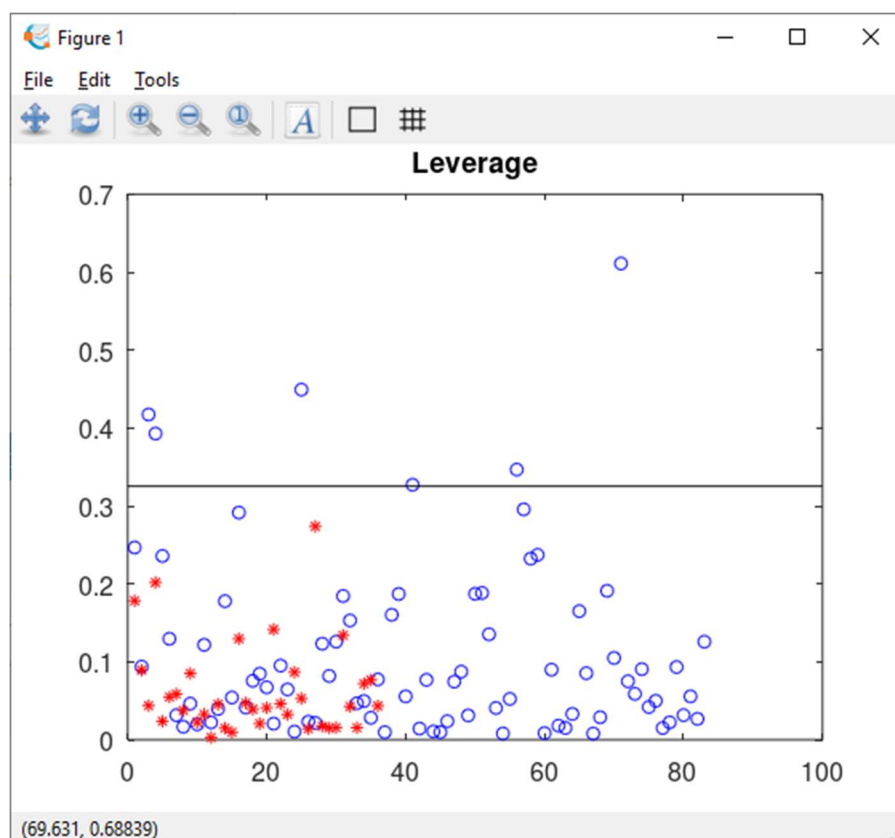


Figura 17. Gráfico de Leverage

Nesse gráfico, onde amostras azuis são de calibração e vermelhas teste, a amostra tem a suspeita de ser outlier quando ela ultrapassa a linha, em 0,3253, nesse caso tivemos seis amostras com indicativo, além disso, podemos perceber que só amostras de calibração estão acima da linha. Agora iremos analisar o Resíduo.

```
>> figure(2)
```

```
>> for qi=1:1:size(modelo9.lev_res.res_cal,1)
```

```
>> plot(ycal(qi),modelo9.lev_res.res_cal2(qi),'bo'); hold on;
```

```

>> end

>> for qi=1:1:size(modelo9.lev_res.res_test,1);%qi=1
>> plot(ytest(qi),modelo9.lev_res.res_test2(qi),'r*'); hold on;
>> end

>> title('Residue')
>> hline(0,'k:');

```

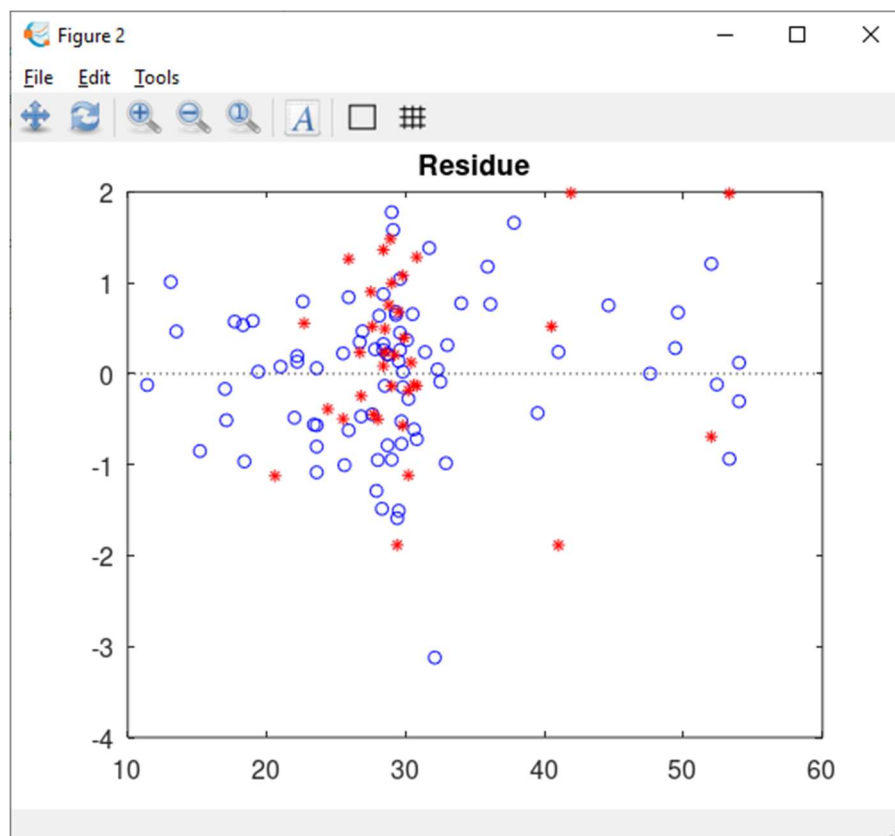


Figura 18. Gráfico de Resíduos.

O Resíduo, como o nome sugere, trata-se da diferença entre a medida original e a predita pelo modelo, caso essa medida seja muito alta, maior proporcionalmente, que de outras amostras do modelo, é um indicio de outlier. Quando analisamos o gráfico, **Figura 18**, onde bolas azuis são amostras de calibração e asteriscos vermelhos são amostras de teste, procuramos por tendências de resíduo, amostras só tendo um erro para um lado em uma determinada faixa, amostras muito afastadas da linha de referência, no caso, encontramos uma amostra bastante afastada negativamente perto da faixa 30 a 35. Agora, vamos analisar ambos os gráficos em conjunto;

```
>> modelo9 = lev_res(modelo9,Xcal2,ycal,Xtest2,ytest);
```

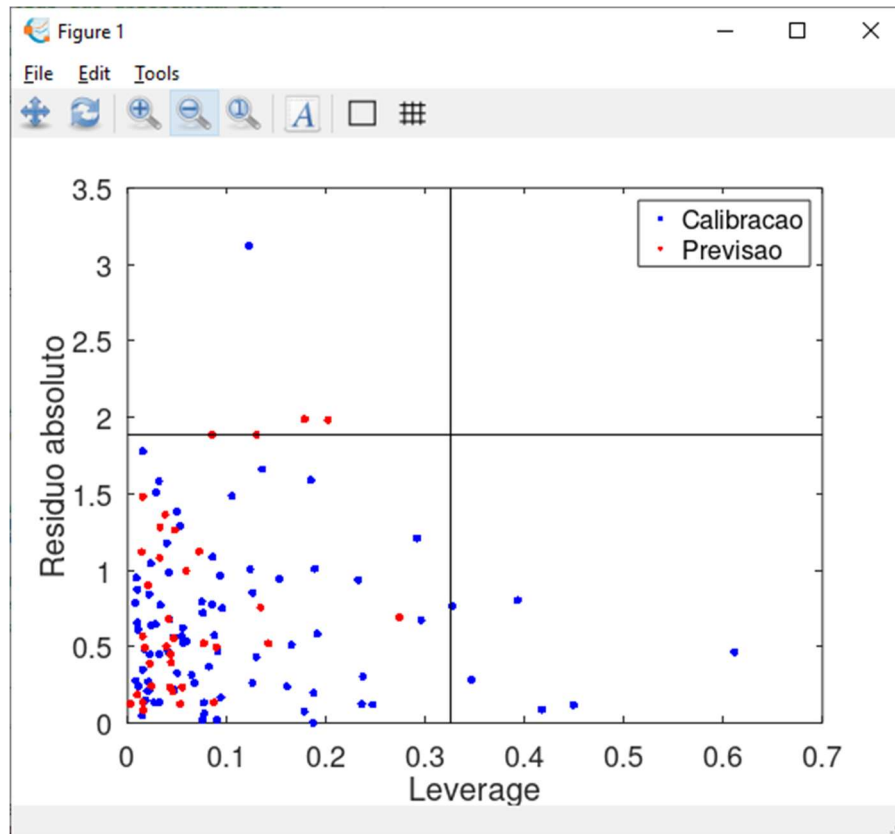


Figura 19. Gráfico Lev_res.

Esta figura, **Figura 19**, combina os gráficos de leverage e resíduos, também conhecido como Gráfico de Lev_Res, com ele podemos analisar quais amostras tem tendência em ambos em conjunto, quando esta amostra fica localizada no primeiro quadrante (Quadrado superior direito). Neste caso, nenhuma amostra foi considerada suspeita por ambos os testes ao mesmo tempo, então concluímos que não existe outlier neste modelo. Para fins didáticos, vamos analisar o modelo com 6 VL e um pré-tratamento?

```
>> close all
>> [Xcal2,Xtest2]=pretrat(Xcal,Xtest,{'deriv',[15,2,1]});
>> modelo6 = lev_res(modelo6,Xcal2,ycal,Xtest2,ytest);
>> legend('off') % Vamos remover a legenda para melhor visualização.
```

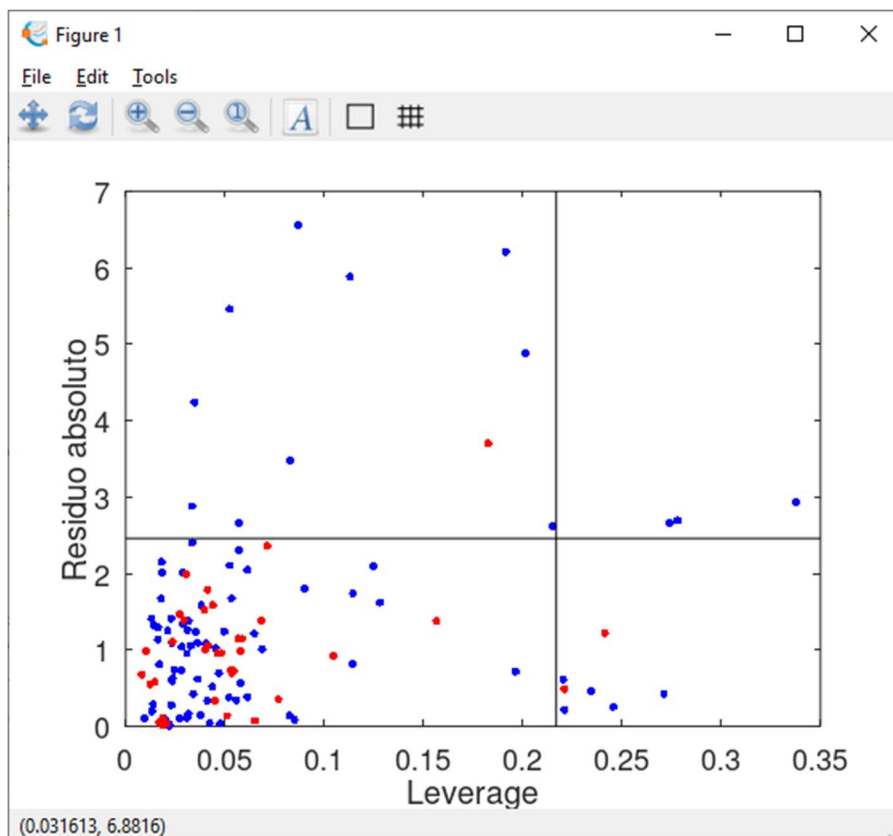


Figura 20. Gráfico de Lev_res do modelo 6.

Pelo gráfico temos três amostras da calibração que se encaixam como dupla suspeita, com isso, podemos considera-las outlier. Como se trata de amostras do conjunto de calibração isso compromete o modelo todo, então, além de remove-las teremos que refazer todo o processo de modelagem.

```
>> A = modelo6.lev_res.Sample_cal
>> Xcal_2 = Xcal; ycal_2 = ycal;
>> Xcal_2(A,:) = []; ycal_2(A,:) = [];
```

A função 'lev_res' informa quais são as amostras que estão no primeiro quadrante, na primeira linha de comando no arquivo 'A' estamos captando quais amostras do conjunto calibração foram considerados outlier, na segunda linha estamos duplicando a matriz Xcal e o vetor ycal, é sempre recomendado manter os dados originais, e na última linha removemos as amostras outlier do conjunto. Após isso, vamos refazer o modelo;

```

>> [Xcal2_2,Xtest2]=pretrat(Xcal_2,Xtest,{'deriv',[15,2,1]});
>> options=[];
>> options.Xpretreat = {'center'};
>> options.vene      = 5;
>> options.vl        = 20;
>> modelo6_2=plsmodel2(Xcal2_2,ycal_2,options);
>> options.vl        = 7;
>> modelo6_7=plsmodel2(Xcal2_2,ycal_2,Xtest2,ytest,options);

```

Ao analisar os parâmetros de avaliação percebemos que os modelos estão bem semelhante um ao outro, vamos refazer o teste de outlier;

```

>> modelo6_7 = lev_res(modelo6_7,Xcal2_2,ycal_2,Xtest2,ytest);
>> legend('off')

```

Uma amostra do conjunto teste ficou no primeiro quadrante, dessa vez podemos remove-la e refazer o modelo com a mesma LV;

9. >> A = modelo6_7.lev_res.Sample_test % Encontrando a amostra teste. Amostra

```

>> Xtest_2 = Xtest; ytest_2 = ytest;
>> Xtest_2(9,:) = []; ytest_2(9,:) = [];
>> [Xcal2_2,Xtest2_2]=pretrat(Xcal_2,Xtest_2,{'deriv',[15,2,1]});
>> options.vl      = 7;
>> modelo6_7=plsmodel2(Xcal2_2,ycal_2,Xtest2_2,ytest_2,options);

```

O novo modelo tem uma mudança significativa nos parâmetros de avaliação, todavia não poderemos utilizar o ‘accuracy_test’ para confirmar que seja o melhor modelo.

```
>> tic

>> [pvalue,dist_tt,meandiff] =
accuracy_test(ytest,modelo6.Ytest(:,2),modelo6_7.Ytest(:,2),'randbi',500000,0.05);

>> toc
```

Ao utilizar a função ocorrerá um erro. Como essa função é um teste paramétrico ela depende que as amostras testes dos dois modelos tenham a mesma dimensão, ou quantidade de amostras, podemos resolver isso ao utilizar um teste não paramétrico, contudo isso foge do escopo desse tutorial, então não iremos abordar neste tutorial.

5.5. Parte 03 - Exemplo real.

Nesta parte iremos simular um dia comum de um Quimiometrista, recebendo as amostras para análise, a separação de conjuntos, analisando qual o melhor pré-tratamento, verificando outlier e decidindo o melhor modelo. Iremos começar com a extração de dados numa planilha, este é o modo mais prático e fácil de passar informações de um químico para o outro;

```
>> cd('C:\Users\Exemplo\...\Monografia\PLS_Model');
>> % Como extrair dados de planilha excel.
>> [y,~,~]=xlsread('Oleos_Adulterados.xlsx','Plan1','B2:B230'); % Vetor de
regressão
```

A função `xlsread`, trata-se de uma função bem prática de entender e aplicar, ela é 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.xlsx','Plan1','C2:DW230');
>> [~,Sample,~]=xlsread('Oleos_Adulterados.xlsx','Plan1','A2:A230');
```

Essas amostras são infravermelho próximo portátil, MicroNir, de azeite adulterado com óleo comercial, essas amostras foram utilizadas em um artigo publicado e pegadas com autorização dos autores,⁵ além disso, por medida de segurança, algumas alterações foram feitas. O Azeite é o óleo extraído da azeitona, ao contrário dos outros óleos, ele tem propriedades boas para a saúde do consumidor, com isso, o seu preço comercial é mais elevado que os demais. Neste modo, pessoas maliciosas utilizam óleos baratos para adulterar o azeite e assim ter algum lucro. Assim, estudos que conseguem identificar adulterantes de modo rápido, como o feito por Folli *et al*⁵ é de suma importância para a indústria.

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 e etc. Então vamos analisar o espectro com o seguinte comando;

```
>> plot(num,X);
```

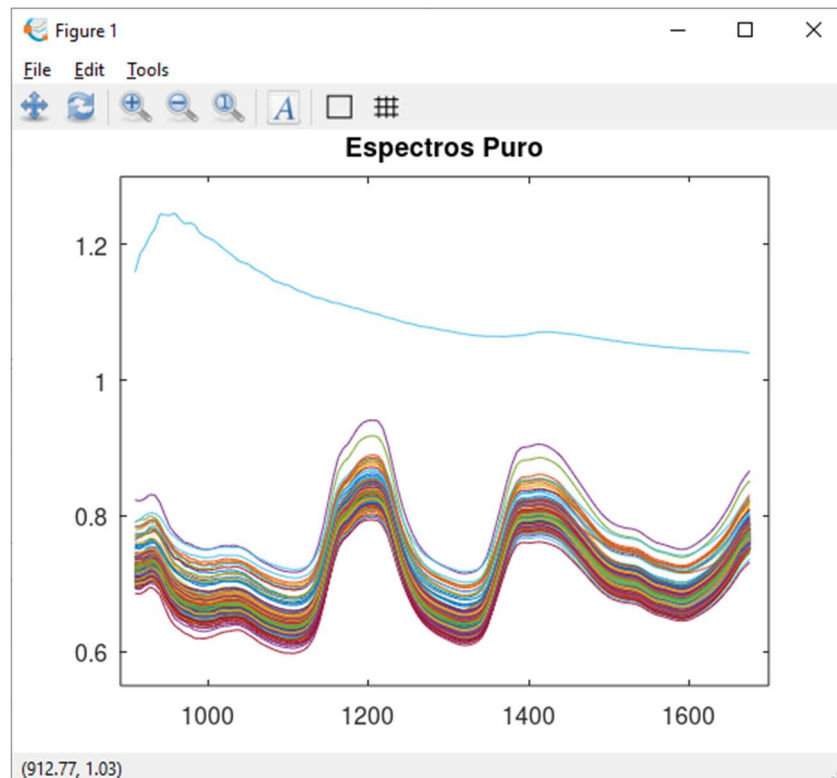


Figura 21. Gráfico microNIR bruto.

Ao analisar o gráfico das amostras, **Figura 21**, percebemos que tem uma amostra com um perfil anômalo, só isso já é o suficiente para podermos remove-la como outlier. Para isso utilizaremos a função `find`;

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

A função ‘`find`’ tem o objetivo de encontrar uma amostra em uma condição específica, no caso estamos procurando uma amostras no ‘`X(:,1)`’, todas amostras do `X` e olhando na coluna 1, e a condição é ‘`> 1`’, valor maior que um, essa condição foi definida ao observar o espectro. Agora vamos separar em conjunto calibração e teste;

```
>> [objetos,Xcal,Xtest,ycal,ytest]=caltest(X,y,70,'k',0,{'center'});
```

A função ‘`caltest`’ foi desenvolvida para separar de forma rápida e fácil as amostras em conjunto calibração e teste, vamos entender melhor dando o comando;

[A,B,C,D,E]=caltest(XXX,YYY,ZZZ,WWW,VVV,UUU);

A funcao xlsread e utilizada para extrair informação de planilhas como xls, xlsx e csv.

INPUT:

XXX = Fonte analítica das amostras.

YYY = Vetor informativo dos dados de regressão.

ZZZ = Percentagem que deve está no conjunto calibração. (Recomendo 70)

WWW = Algoritmo desejado. (Recomendo 'k' kenston)

VVV = Checar repetição. (0 = Sem checagem 1 = Com checagem)

UUU = Método de pré-tratamento utilizado antes da separação. (Recomendo 'none')

OUTPUT:

A = Objetos, conjunto estrutural com os dados da separação. (Recomendo sempre salvar)

B = Fonte Analítica conjunto calibração.

C = Fonte Analítica conjunto teste.

D = Vetor informativo conjunto calibração.

E = Vetor informativo conjunto teste.

Após uma separação é importante verificar se as amostras foram corretamente separadas, então vamos utilizar os seguintes comandos:

`>> close all`

`>> plot(1:1:size(yca1,1),yca1,'bo'); hold on;`

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

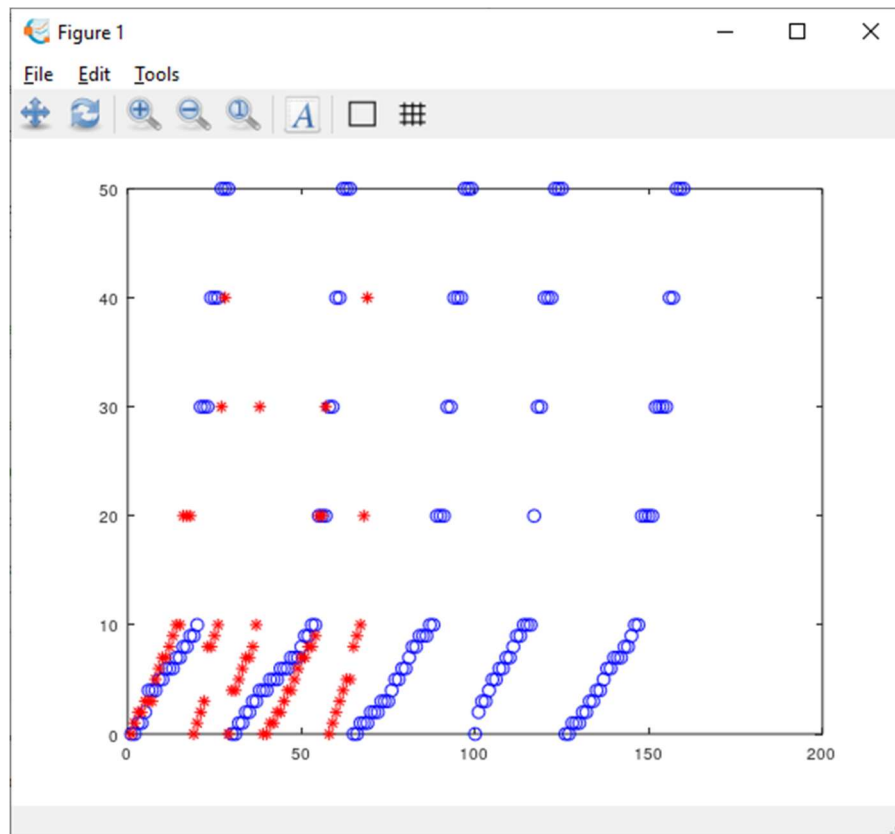


Figura 22. Separação caltest.

A separação de calibração e teste foi bem feita, podemos dizer isso porque as amostras de calibração, bola azul, espalharam-se por faixa espectral e as amostras testes, asterisco vermelho, ficaram dentro dessa faixa. Contudo, não é só isso que é importante para validarmos a separação, este conjunto amostral está na forma de triplicata, como o comando a seguir demonstra;

```
>> Sample([1 2 3],:)
```

Vemos que as três primeiras amostras são amostras de algodão 1, uma triplicata não real, por regra, triplicata não real devem ficar no mesmo conjunto, ou seja, as amostras 1,2 e 3 devem ficar ou no conjunto calibração, ou no teste, caso isso não seja feito estaremos provocando uma viés no modelo, criando um resultado falso. Vamos verificar aonde as três primeiras amostras estão;

```
>> objetos.cal([1 2 3],:)
```

Ao analisar as primeiras amostras do conjunto calibração percebemos que são 1, 3 e 5, assim, percebemos que as triplicatas não reais não ficaram no mesmo conjunto, assim, essa separação não é aceitável. Então, usaremos a mesma separação que o artigo original utilizou, com base na concentração e para isso usaremos ao seguinte comando;

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

A terceira linha de comando é conectada diretamente a décima linha, através da função ‘for’ combinado com o comando ‘if’ conseguimos verificar amostra por amostra e dependendo da resposta, linha 6, colocamos a amostra no conjunto correto. Contudo, planejamos criar um tutorial extra no futuro explicando esses dois comandos. Vamos começar a fazer o modelo;

```
>> close all
>> options=[];
>> options.Xpretreat = {'center'};
>> options.vene = 5;
>> options.vl = 20;
>> modelo=plsmodel2(Xcal,ycal,options)
>> options.vl = 9;
```

```
>> modelo9 = plsmodel2(Xcal,ycal,Xtest,ytest,options);
>> options.vl      = 15;
>> modelo15 = plsmodel2(Xcal,ycal,Xtest,ytest,options);
```

Quando analisamos os parâmetros de avaliação percebemos que o modelo com 15 LV conseguiu R^2p e RMSEP melhores, 0,7937 e 5 m/m respectivamente, nessa forma poderíamos considerar este modelo melhor, todavia, como estamos trabalhando em determinar a quantidade de adulterante em uma amostra de alimentos é importante analisarmos a capacidade de detecção e quantificação dos modelos. Então utilizamos uma nova função;

```
>> modelo9=eparamenter(modelo9,Xcal,Xtest,ycal,ytest);
>> modelo15=eparamenter(modelo15,Xcal,Xtest,ycal,ytest);
```

A função ‘eparamenter’ foi desenvolvida para podermos analisar outros parâmetros de avaliação em um modelo criado pela função ‘plsmodel’, neste caso iremos analisar o limite de detecção (LoD) e limite de quantificação (LoQ), falamos de ambos em 272.2.3. Avaliação, contudo, lembre-se que o LoD determinar qual a concentração mínima que o modelo consegue detectar e o LoQ a concentração mínima para quantificação, nesse modo, eles são consideravelmente importantes para determinação de adulterantes.

Quando olhamos estes parâmetros o modelo de 9 LV se destaca com 6 e 18 m/m de LoD e LoQ, respectivamente, contra 7 e 24 m/m do modelo de 15 LV. Agora, vamos analisar o gráfico de medido e predito das amostras;

```
te
>> 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','Location','southeast');
>> title('Modelo 9');
>> set(gca,'FontSize',12); xlabel('Reference','fontsize',12);
>> ylabel('Predicted','fontsize',12);
```

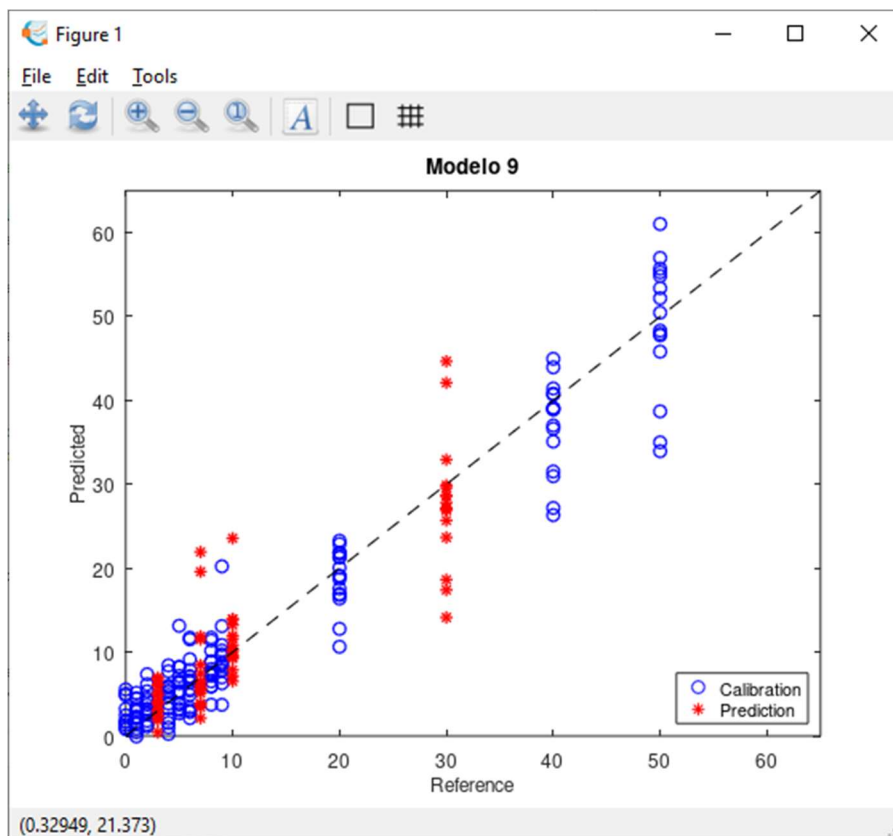


Figura 23. Gráfico de medido e predito modelo azeite.

Quando analisamos o gráfico de medidos e preditos, **Figura 23**, 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. Vamos analisar os outlier do modelo;

```
>> modelo9 = lev_res(modelo9,Xcal,ycal,Xtest,ytest);
```

Como podemos ver ao analisar o ‘modelo9.lev_res.Sample_test’, temos quatro amostras do conjunto teste como outlier e nenhuma do conjunto calibração, assim, devemos remover as amostras teste e refazer o modelo utilizando a mesma LV;

```
>> A = modelo9.lev_res.Sample_test;
>> Xtest_2 = Xtest; ytest_2 = ytest;
>> Xtest_2(A,:) = []; ytest_2(A,:) = [];
>> options.vl      = 9;
>> modelo9_9 = plsmodel2(Xcal,ycal,Xtest_2,ytest_2,options);
```

```
>> modelo9_9 = eparameter(modelo9_2,Xcal,Xtest,ycal,ytest);
```

Houve uma melhora nos parâmetros R^2p e RMSEP, o que corrobora a ideia que eram outliers, nesse sentido, escolhemos o modelo9_9 como o melhor modelo, contudo, com o aprendizado nesse tutorial ainda podemos testar alguns pré-tratamentos e outras metodologias.

5.6. Parte 04 - Exercício para praticar.

Na parte final, deixaremos um exercício para casa. 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 obtidos pela equipe do laboratório, então, se sinta desafiado a tentar encontrar os modelos. O Gabarito se encontra no final do tutorial.

6. CONCLUSÃO

Neste estudo foi apresentado um tutorial didático focado no PLS aplicável no software gratuito Octave e no mais utilizado Matlab, com foco secundário divulgar e facilitar o aprendizado de Quimiometria. As rotinas desenvolvidas nesse tutorial estão disponíveis no Github, além das funções e rotinas extras, além da disponibilidade dos dados químicos para praticas em sala de aula e individuais.

7. REFERÊNCIAS

1. Barros Neto B de, Scarminio IS, Bruns RE. 25 anos de quimiometria no Brasil. *Quim Nova*. 2006;**29**(6):1401-1406. doi:10.1590/S0100-40422006000600042.
2. Pereira F, Pereira-Filho E. APLICAÇÃO DE PROGRAMA COMPUTACIONAL LIVRE EM PLANEJAMENTO DE EXPERIMENTOS: UM TUTORIAL. *Quim Nova*. 2018;**2018**(9):1061-1071. doi:10.21577/0100-4042.20170254.
3. Chen B, Lu Y, Pan W, et al. Support vector machine classification of nonmelanoma skin lesions based on fluorescence lifetime imaging microscopy. *Anal Chem*. 2019;**91**(16):10640-10647. doi:10.1021/acs.analchem.9b01866.
4. da Cunha PHP, de Paulo EH, Silveira Folli G, Nascimento MHC, Moro MK, Filgueiras PR. Variable selection by permutation applied in support vector regression models. *J Chemom*. 2022;**36**(10):1-14. doi:10.1002/cem.3444.
5. Folli GS, Santos LP, Santos FD, et al. Food analysis by portable NIR spectrometer. *Food Chem Adv*. 2022;**1**(March):100074. doi:10.1016/j.focha.2022.100074.
6. Santana F, Souza A, Almeida M, et al. EXPERIMENTO DIDÁTICO DE QUIMIOMETRIA PARA CLASSIFICAÇÃO DE ÓLEOS VEGETAIS COMESTÍVEIS POR ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO COMBINADO COM ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS: UM TUTORIAL, PARTE V. *Quim Nova*. 2020;**43**(3):371-381. doi:10.21577/0100-4042.20170480.
7. De Souza AM, Poppi RJ. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: UM tutorial, parte I. *Quim Nova*. 2012;**35**(1):223-229. doi:10.1590/S0100-40422012000100039.
8. Moro MK, Neto ÁC, Lacerda V, et al. FTIR, ¹H and ¹³C NMR data fusion to predict crude oils properties. *Fuel*. 2020;**263**(November 2019):116721. doi:10.1016/j.fuel.2019.116721.
9. Ferreira MMC. *Quimiometria: Conceitos, Métodos e Aplicações*. São Paulo: Editora da Unicamp; 2015. doi:10.7476/9788526814714.
10. IUPAC. Chemometrics. In: *IUPAC Compendium of Chemical Terminology*. Vol 69. Research Triangle Park, NC: IUPAC; 1997:1140. doi:10.1351/goldbook.CT06948.

11. Spencer Lima L. Lei de Lambert–Beer. *Rev Ciência Elem.* 2013;**1**(1):1-2. doi:10.24927/rce2013.047.
12. Wu B, Guo S, Zhang L, et al. Spatial variation of residual total petroleum hydrocarbons and ecological risk in oilfield soils. *Chemosphere.* 2022;**291**(November 2021):132916. doi:10.1016/j.chemosphere.2021.132916.
13. Morais CLM, Costa FSL, Lima KMG. Variable selection with a support vector machine for discriminating: *Cryptococcus* fungal species based on ATR-FTIR spectroscopy. *Anal Methods.* 2017;**9**(20):2964-2970. doi:10.1039/c7ay00428a.
14. Santos FD, Santos LP, Cunha PHP, et al. Discrimination of oils and fuels using a portable NIR spectrometer. *Fuel.* 2021;**283**:118854. doi:10.1016/j.fuel.2020.118854.
15. Bruns RE, Faigle JFG. Quimiometria. *Quim Nova.* 1985;**8**(2):84-99.
16. de Paulo EH, dos Santos FD, Folli GS, et al. Determination of gross calorific value in crude oil by variable selection methods applied to ¹³C NMR spectroscopy. *Fuel.* 2022;**311**(July 2021):122527. doi:10.1016/j.fuel.2021.122527.
17. de Araújo Gomes A, Azcarate SM, Diniz PHGD, de Sousa Fernandes DD, Veras G. Variable selection in the chemometric treatment of food data: A tutorial review. *Food Chem.* 2022;**370**(July 2021):131072. doi:10.1016/j.foodchem.2021.131072.
18. Yang Y, Wang Y. Characterization of *Paris polyphylla* var. *yunnanensis* by Infrared and Ultraviolet Spectroscopies with Chemometric Data Fusion. *Anal Lett.* 2018;**51**(11):1730-1742. doi:10.1080/00032719.2017.1385618.
19. Kennard, R.W. Stone LA, Kennard RW, Stone LA. Computer Aided Design of Experiments. *Technometrics.* 1969;**11**(1):137-148. doi:doi.org/10.1080/00401706.1969.10490666.
20. Bin J, Ai F, Fan W, et al. An efficient variable selection method based on variable permutation and model population analysis for multivariate calibration of NIR spectra. *Chemom Intell Lab Syst.* 2016;**158**:1-13. doi:10.1016/j.chemolab.2016.08.006.
21. Lemos MFMF, Perez C, da Cunha PHPPHP, et al. Chemical and sensory profile of new genotypes of Brazilian *Coffea canephora*. *Food Chem.* 2020;**310**(October 2019):125850. doi:10.1016/j.foodchem.2019.125850.
22. Liu J, Fan X, Zhang C, et al. Moisture Diagnosis of Transformer Oil-Immersed

- Insulation with Intelligent Technique and Frequency-Domain Spectroscopy. *IEEE Trans Ind Informatics*. 2021;**17**(7):4624-4634. doi:10.1109/TII.2020.3014224.
23. Santana F, Souza A, Almeida MR, et al. EXPERIMENTO DIDÁTICO DE QUIMIOMETRIA PARA CLASSIFICAÇÃO DE ÓLEOS VEGETAIS COMESTÍVEIS POR ESPECTROSCOPIA NO INFRAVERMELHO MÉDIO COMBINADO COM ANÁLISE DISCRIMINANTE POR MÍNIMOS QUADRADOS PARCIAIS: UM TUTORIAL, PARTE V. *Quim Nova*. 2020;**43**(3):371-381. doi:10.21577/0100-4042.20170480.
 24. Rocha WF de C, Sheen DA. Determination of physicochemical properties of petroleum derivatives and biodiesel using GC/MS and chemometric methods with uncertainty estimation. *Fuel*. 2019;**243**(July 2018):413-422. doi:10.1016/j.fuel.2018.12.126.
 25. Souza AM De, Breitreitz MC, Filgueiras PR, Rohwedder JJR, Poppi RJ. Experimento didático de quimiometria para calibração multivariada na determinação de paracetamol em comprimidos comerciais utilizando espectroscopia no infravermelho próximo: um tutorial, parte II. *Quim Nova*. 2013;**36**(7):1057-1065. doi:10.1590/S0100-40422013000700022.
 26. Zhou X, Sun J, Tian Y, Lu B, Hang Y, Chen Q. Development of deep learning method for lead content prediction of lettuce leaf using hyperspectral images. *Int J Remote Sens*. 2020;**41**(6):2263-2276. doi:10.1080/01431161.2019.1685721.
 27. Correia PRM, Ferreira MMC. Reconhecimento de padrões por métodos não supervisionados: Explorando procedimentos quimiométricos para tratamento de dados analíticos. *Quim Nova*. 2007;**30**(2):481-487. doi:10.1590/S0100-40422007000200042.
 28. Mohammadi M, Khanmohammadi Khorrami M, Vatanparast H, Karimi A, Sadrara M. Classification and determination of sulfur content in crude oil samples by infrared spectrometry. *Infrared Phys Technol*. 2022;**127**(June):104382. doi:10.1016/j.infrared.2022.104382.
 29. Valderrama P, Braga JWB, Poppi RJ. Estado da arte de figuras de mérito em calibração multivariada. *Quim Nova*. 2009;**32**(5):1278-1287. doi:10.1590/s0100-40422009000500034.
 30. Ferreira MMC, Antunes AM, Melgo MS, Volpe PLO. Quimiometria I: calibração multivariada, um tutorial. *Quim Nova*. 1999;**22**(5):724-731. doi:10.1590/S0100-40421999000500016.

31. ASTM E1655. Standard Practices for Infrared Multivariate Quantitative Analysis. 2012. doi:10.1520/E1655-05.
32. van der Voet H. Comparing the predictive accuracy of models using a simple randomization test. *Chemom Intell Lab Syst.* 1994;**25**(2):313-323. doi:10.1016/0169-7439(94)85050-X.
33. Filgueiras PR, Portela NA, Silva SRC, et al. Determination of Saturates, Aromatics, and Polars in Crude Oil by ¹³C NMR and Support Vector Regression with Variable Selection by Genetic Algorithm. *Energy and Fuels.* 2016;**30**(3):1972-1978. doi:10.1021/acs.energyfuels.5b02377.
34. Cruz-Tirado JP, Lima Brasil Y, Freitas Lima A, et al. Rapid and non-destructive cinnamon authentication by NIR-hyperspectral imaging and classification chemometrics tools. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2023;**289**(December 2022):122226. doi:10.1016/j.saa.2022.122226.
35. Nascimento MHC, Oliveira BP, Rainha KP, Castro EVR, Silva SRC, Filgueiras PR. Determination of flash point and Reid vapor pressure in petroleum from HTGC and DHA associated with chemometrics. *Fuel.* 2018;**234**(July):643-649. doi:10.1016/j.fuel.2018.07.050.
36. Baldo MA, Oliveri P, Fabris S, Malegori C, Daniele S. Fast determination of extra-virgin olive oil acidity by voltammetry and Partial Least Squares regression. *Anal Chim Acta.* 2019;**1056**:7-15. doi:10.1016/j.aca.2018.12.050.
37. Correia RM, Tosato F, Domingos E, et al. Portable near infrared spectroscopy applied to quality control of Brazilian coffee. *Talanta.* 2018;**176**(August 2017):59-68. doi:10.1016/j.talanta.2017.08.009.
38. Chen H, Tan C, Lin Z, Wu T. Classification and quantitation of milk powder by near-infrared spectroscopy and mutual information-based variable selection and partial least squares. *Spectrochim Acta Part A Mol Biomol Spectrosc.* 2018;**189**:183-189. doi:10.1016/j.saa.2017.08.034.
39. Zhang W, Zhou R, Yang P, et al. Determination of chlorine with radical emission using laser-induced breakdown spectroscopy coupled with partial least square regression. *Talanta.* 2019;**198**(January):93-96. doi:10.1016/j.talanta.2019.01.102.
40. Mahesh S, Jayas DS, Paliwal J, White NDG. Comparison of Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) Methods for Protein and Hardness Predictions using the Near-Infrared (NIR) Hyperspectral Images of Bulk Samples of Canadian Wheat. *Food Bioprocess*

- Technol.* 2015;**8**(1):31-40. doi:10.1007/s11947-014-1381-z.
41. Gredilla A, Fdez-Ortiz de Vallejuelo S, Elejoste N, de Diego A, Madariaga JM. Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review. *TrAC Trends Anal Chem.* 2016;**76**:30-39. doi:10.1016/j.trac.2015.11.011.
 42. Zhou F, Yang K, Li D, Shi X. Acid Number Prediction Model of Lubricating Oil Based on Mid-Infrared Spectroscopy. *Lubricants.* 2022;**10**(9):205. doi:10.3390/lubricants10090205.
 43. Mariani NCT, Costa RC, Lima KMG, Nardini V, Cunha Júnior LC, Teixeira GHA. Predicting soluble solid content in intact jaboticaba [*Myrciaria jaboticaba* (Vell.) O. Berg] fruit using near-infrared spectroscopy and chemometrics. *Food Chem.* 2014;**159**:458-462. doi:10.1016/j.foodchem.2014.03.066.
 44. Vieira AP, Portela NA, Neto ÁC, et al. Determination of physicochemical properties of petroleum using ¹H NMR spectroscopy combined with multivariate calibration. *Fuel.* 2019;**253**(December 2018):320-326. doi:10.1016/j.fuel.2019.05.028.
 45. Lovatti BPO, Nascimento MHC, Neto ÁC, Castro EVR, Filgueiras PR. Use of Random forest in the identification of important variables. *Microchem J.* 2019;**145**(December 2018):1129-1134. doi:10.1016/j.microc.2018.12.028.
 46. Geladi P. Notes on the history and nature of partial least squares (PLS) modelling. *J Chemom.* 1988;**2**(4):231-246. doi:10.1002/cem.1180020403.
 47. Brereton RG. Introduction to multivariate calibration in analytical chemistry. *Analyst.* 2000;**125**(11):2125-2154. doi:10.1039/b003805i.
 48. Zerlia T, Pinelli G. Asphaltenes determination in heavy petroleum products by partial least squares analysis of u.v. data. *Fuel.* 1992;**71**(5):559-563. doi:10.1016/0016-2361(92)90154-G.
 49. Wilt BK, Welch WT, Graham Rankin J. Determination of asphaltenes in petroleum crude oils by fourier transform infrared spectroscopy. *Energy and Fuels.* 1998;**12**(5):1008-1012. doi:10.1021/ef980078p.
 50. Long J, Wang K, Yang M, Zhong W. Rapid crude oil analysis using near-infrared reflectance spectroscopy. *Pet Sci Technol.* 2019;**37**(3):354-360. doi:10.1080/10916466.2018.1547754.
 51. Moro MK, dos Santos FD, Folli GS, Romão W, Filgueiras PR. A review of chemometrics models to predict crude oil properties from nuclear magnetic

- resonance and infrared spectroscopy. *Fuel*. 2021;**303**(June):121283.
doi:10.1016/j.fuel.2021.121283.
52. Du Y, Huang H, Peng Y, Wang J, Gao Z. Rapid determination of *Staphylococcus aureus* enterotoxin B in milk using Raman spectroscopy and chemometric methods. *J Raman Spectrosc*. 2022;**53**(4):709-714.
doi:10.1002/jrs.6296.
 53. Wang Y, Cao H, Zhou Y, Zhang Y. Nonlinear partial least squares regressions for spectral quantitative analysis. *Chemom Intell Lab Syst*. 2015;**148**:32-50.
doi:10.1016/j.chemolab.2015.08.024.
 54. Matlab. <https://www.mathworks.com/products/matlab.html>.
 55. ISO 12185. Crude petroleum and petroleum products – determination of density – oscillating U-tube method. 1996.