PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

NIR ASSOCIADO À QUIMIOMETRIA – UMA PODEROSA FERRAMENTA ANALÍTICA

NIR ASSOCIATED WITH CHEMOMETRICS – A POWERFUL ANALYTICAL TOOL

José Antonio Martins[1], Mariana de Moraes Cardoso[2], Gabriela Marques Florencio[3], Vagner Sargentelli[4].

1- Doutor em Ciências (Instituto de Química – Unicamp), Cientista de Dados, Nanotimize Tecnologia; **2**- Engenheira Ambiental (Unesp – Sorocaba), Acadêmica de Desenvolvimento de Software Multiplataformas (Fatec – Itapira-SP), Pesquisadora, Nanotimize Tecnologia; **3**- Geóloga (Unesp – Rio Claro), Acadêmica de Desenvolvimento de Software Multiplataformas (Fatec – Itapira-SP), Pesquisadora, Nanotimize Tecnologia; **4**- Doutor em Química (Instituto de Química, Unesp – Araraquara), Cientista Sênior, Nanotimize Tecnologia.

Contato **1-4**: contato@nanotimize.com.br

RESUMO

O infravermelho próximo (NIR) é uma modalidade da espectroscopia vibracional com energia relativamente alta em que são analisadas moléculas contendo, principalmente, os grupos funcionais C–H, N–H, O–H e S–H, e pode ser utilizado em análises qualitativas e quantitativas. Os métodos analíticos resultantes do uso do NIR refletem suas características mais significativas, como: rápido (um minuto ou menos por amostra), não destrutivo, não invasivo, com alta penetração no feixe de radiação da sondagem, adequado para uso em linha, aplicação quase universal e com demandas mínimas de preparação de amostras. Apesar dessas vantagens, o uso do NIR só se tornou difundido pelo desenvolvimento da computação e pela disciplina de quimiometria. A quimiometria envolve a aplicação de métodos estatísticos e matemáticos, bem como aqueles baseados na lógica matemática, à análise química, fornecendo as ferramentas para a coleta de informações e seu uso racional. Assim, o escopo do presente trabalho é o de apresentar os principais aspectos do NIR e da quimiometria, cuja associação é imprescindível quando a finalidade é utilizar o NIR em análises quantitativas.

Palavras-Chave: NIR; Quimiometria. Análises Quantitativas.

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

ABSTRACT

Near infrared (NIR) is a modality of vibrational spectroscopy with relatively high energy in which molecules containing, mainly, the functional groups C–H, N–H, O–H and S–H are analyzed, and can be used in analyzes qualitative and quantitative. Analytical methods resulting from the use of NIR reflect its most significant characteristics, such as fast (one minute or less per sample), non-destructive, non-invasive, with high penetration into the probing radiation beam, suitable for on-line use, nearly universal and with minimal sample preparation demands. Despite these advantages, the use of NIR only became widespread through the development of computing and the discipline of chemometrics. Chemometrics involves the application of statistical and mathematical methods, as well as those based on mathematical logic, to chemical analysis, providing the tools for gathering information and its rational use. Thus, the scope of this work is to present the main aspects of NIR and chemometrics, whose association is essential when the purpose is to use NIR in quantitative analyses.

Keywords: NIR; Chemometrics; Quantitative Analysis.

INTRODUÇÃO

O infravermelho (IR – *Infrared*) é um tipo de espectroscopia vibracional que envolve a interação da radiação eletromagnética com a matéria. O IR está compreendido entre 10 – 13300 cm^{-1} (100000 a 750 nm), portanto, na região não visível do espectro. A unidade número de onda, v, em cm^{-1}, é uma grandeza física inversamente proporcional ao comprimento de onda ($v = 1/\lambda$), e é muito utilizada nos espectros de IR. Apesar disto, é difundido falar em "frequência vibracional", f, ($f = c/\lambda$; c = velocidade da luz) e apresentar o valor em "número de onda" (NAKAMOTO, 1970; SILVERSTEIN et al., 1969; TIBOLA, 2018).

No IR não ocorre transições eletrônicas nas moléculas porque a energia na região do IR não é suficiente. Na espectroscopia IR estão envolvidas vibrações moleculares. As vibrações moleculares podem ser divididas em estiramentos e

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

deformações angulares. A vibração de estiramento envolve oscilações nas distâncias interatômicas ao logo do eixo de ligação entre dois átomos e podem ser simétricas ou assimétricas. Na deformação angular o que varia é o ângulo de ligação entre dois átomos ligados e podem ocorrer no plano ou fora do plano da molécula. Para moléculas poliatômicas, cada átomo tem três graus de liberdade (nas direções x, y e z). Assim, o número de modos vibracionais de uma molécula com N átomos é 3N graus de liberdade. Nessa espectroscopia estão envolvidas apenas as vibrações moleculares (podem ser desconsiderados os movimentos translacionais e rotacionais). Deste modo, para uma molécula poliatômica os modos vibracionais de vibração são dados por 3N – 6 graus de liberdade e, para uma molécula linear, 3N – 5. Para ocorrer absorção no IR é preciso que a vibração provoque mudança no momento de dipolo elétrico da molécula. Assim, moléculas apolares (H_2, N_2, Ar, etc.) não absorvem no IR. Moléculas poliatômicas lineares que apresentem mudança no momento de dipolo induzido absorvem no IR (por exemplo: CO_2). Para uma molécula contendo três átomos e com geometria angular, os seguintes modos normais de vibração são possíveis: estiramento simétrico, estiramento assimétrico e deformação angular (NAKAMOTO, 1970; SILVERSTEIN et al., 1969; TIBOLA, 2018).

As vibrações moleculares podem ser representadas por um modelo mecânico simples, conhecido como oscilador harmônico. Neste modelo, as vibrações de estiramento são consideradas no modelo mecânico de duas massas unidas por uma mola. Portanto, a frequência de vibração de dois átomos ligados pode ser calculada considerando que a energia segue o comportamento de um oscilador harmônico conforme a lei de Hooke (NAKAMOTO, 1970; SILVERSTEIN et al., 1969):

$$f = \frac{1}{2\pi}\sqrt{(k/\mu)} \quad (1)$$

Onde:
k = constante de força da ligação (Kg s^{-2})
μ = massa reduzida = $m_1m_2/m_1 + m_2$ (massa em Kg).

PROSPECTUS
ISSN 2674-8576
DOI: 10.5281/zenodo.10045561 Artigo de Revisão

A energia potencial (E_p) pode ser dada pela equação (6) que define a amplitude da vibração:

$$Ep = \frac{1}{2} k\, x^2 \quad (2)$$

Onde:
k = constante de força da ligação (Kg s^{-2});
x = deslocamento (em metros).

Da equação (2) resulta que quanto maior k, maior é o modo normal de vibração. Por exemplo: C≡C (2150 cm^{-1}); C=C (1650 cm^{-1}) e C–C (1200 cm^{-1}). De modo análogo, quanto maior a massa reduzida, menor é o número de onda observado. Exemplos: C–Cl (750 cm^{-1}); C–Br (600 cm^{-1}) e C–I (500 cm^{-1}).

Com base em cálculos e em espectros de diversas substâncias foi possível estabelecer regiões em que determinados modos normais de vibração ocorrem, as quais foram tabeladas e são amplamente difundidas.

O modelo acima é adequado para a avaliação dos modos normais de vibração para moléculas diatômicas simples, onde os resultados calculados são coincidentes com os observados. Todavia, para moléculas com vários átomos, os números de vibrações possíveis resultam em espectros no IR que são complexos de analisar com base neste modelo, porque o mesmo não descreve totalmente o comportamento das vibrações em dimensões atômicas. Para uma melhor descrição são utilizados os princípios da mecânica quântica, nos quais os osciladores harmônicos podem ter apenas energias quantizadas (ou, em outras palavras, determinadas energias).

Para fins práticos, e de instrumentação, o IR é dividido em três regiões: distante, FIR (*Far Infrared*) 400 – 10 cm^{-1}, médio, MIR (*Middle Infrared*) 4000 – 400 cm^{-1} e próximo, NIR (*Near Infrared*), 13300 – 4000 cm^{-1}.

Os primeiros espectrofotômetros no infravermelho eram do tipo dispersivos, onde o espectro era varrido com luz monocromática. Os instrumentos eram analógicos, apresentavam baixa sensibilidade e eram necessários alguns minutos

PROSPECTUS
ISSN 2674-8576
DOI: 10.5281/zenodo.10045561 Artigo de Revisão

para se analisar uma amostra. Com o desenvolvimento da área computacional, os espectrofotômetros com Transformadas de Fourier (FT – *Fourier Transform*) passaram a ser difundidos porque permitem obter todo o espectro no IR em questão de segundos.

Um espectrômetro FTIR possui os seguintes componentes principais: fonte de radiação, interferômetro de Michelson e detector. O Interferômetro de Michelson consiste em um espelho móvel, um espelho fixo e um divisor de feixe (um espelho que permite que parte da radiação incidente seja transmitida e parte refletida). De acordo com um espectrômetro FTIR, a radiação da fonte é dividida em dois feixes de radiação, presentes no detector após percorrer caminhos diferentes. Do sinal do detector, um interferograma é obtido – registro do sinal que produz um Interferômetro de Michelson. A Transformada de Fourier é um cálculo matemático que converte o interferograma em um espectro no infravermelho. Entre as principais vantagens que os espectrômetros FTIR apresentam citam – se: alta velocidade, resolução, sensibilidade, exatidão e precisão. Menciona-se que há disponíveis equipamentos com arranjos que permitem abranger toda a faixa espectral no infravermelho, ou somente as regiões FIR e MIR, e outros que abrangem apenas a região NIR.

Uma substância pode ser analisada no estado sólido, líquido ou gasoso em um espectrofotômetro infravermelho e são necessárias poucas quantidades de amostra. No FIR e MIR, para obter os espectros nos estados líquido ou gasoso são utilizadas células especiais. Para o estado sólido podem ser feitas medições em emulsão de nujol ou diluindo-se a amostra a ser analisada em sais inorgânicos desidratados, brometo de potássio (KBr) e cloreto de sódio (NaCl), MIR, ou iodeto de césio, FIR, para produzir uma pastilha (é necessário o uso de um pastilhador que emprega altas pressões para prensar a mistura). Outras técnicas que requerem menor preparação de amostra também podem se utilizadas, tais como a Reflexão Total Atenuada (ATR - *Attenuated Total Reflection*) e Reflexão Difusa. No NIR o composto pode ser analisado sem prévia preparação colocando-o diretamente em um porta amostra. Ressalta-se que o rápido progresso na miniaturização dos espectrômetros NIR aproveitou as novas tecnologias e levou a uma redução drástica dos tamanhos e

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

pesos dos espectrômetros, permitindo um bom desempenho devido à implementação de alta precisão de elementos importantes no dispositivo final.

A divisão da espectroscopia no IR em três regiões distintas possibilita o estudo pormenorizado de diferentes aspectos das vibrações moleculares. No FIR as interações entre metal e um não metal são particularmente observadas, enquanto no MIR as moléculas orgânicas são muito investigadas e caracterizadas, tendo, inclusive, diferentes arranjos instrumentais para isto. Contudo, tanto o FIR quanto o MIR são utilizados em análises qualitativas, ou seja, as de caracterização, devido à dificuldade de ser realizada uma análise quantitativa. Já o NIR é uma região com energia relativamente alta em que são analisadas moléculas contendo, principalmente, os grupos funcionais C–H, N–H, O–H e S–H, e pode ser utilizado em análises quantitativas (NAKAMOTO, 1970; SILVERSTEIN et al., 1969; FORATO, 2010; TIBOLA, 2018).

NIR

Foram apresentadas as considerações teóricas de um oscilador harmônico clássico (Lei de Hooke) e que a mecânica quântica descreve melhor as vibrações ao nível molecular, onde existem átomos unidos por uma ligação química. Aqui, somente níveis de energia específicos podem ser assumidos para os modos normais de vibração. Todavia, os sistemas moleculares não se comportam perfeitamente harmônico porque os átomos tendem a sofrer repulsão quando se aproximam entre si e a ligação química tende a enfraquecer e sofrer ruptura quando os átomos se afastam. Assim, o modelo anarmônico quântico descreve melhor as vibrações moleculares e o tratamento mecânico-quântico dos sistemas anarmônicos produz resultados distintos (PASQUINI, 2003; PASQUINI, 2018).

Num resultado, os níveis de energia vibracionais não são igualmente espaçados, mas a diferença entre eles diminui enquanto os níveis de energia tornam-se mais altos. Portanto, isto considera que a grandes afastamentos dos átomos no momento da vibração pode conduzir ao rompimento da ligação química.

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

Um segundo aspecto é que a energia vibracional de uma molécula pode transitar entre níveis diretamente do mais baixo para outro mais alto e não necessariamente em sequência (por exemplo: uma vibração entre o nível de menor energia, nível 0 – nível fundamental, e o nível 3, 0 → 3). Esse tipo de comportamento gera o que é denominado de sobretons e é de fundamental importância para a espectroscopia NIR. A vibração 0 → 2, é o primeiro sobreton; 0 → 3, terceiro sobreton; 0 → 4, quarto sobreton, e assim sucessivamente. É preciso notar que estas transições são permitidas pela distorção do modelo harmônico, implicando, consequentemente, que suas intensidades nos espectros NIR são muito menores do que àquelas nos espectros MIR, onde as transições fundamentais são observadas (transições entre os níveis 0 → 1; 1 → 2 etc.). A diferença característica é de 10, 100 e 10000 vezes menor, à medida que se passa do primeiro, para o segundo e terceiro sobretons, respectivamente (PASQUINI, 2003; LEITÃO, 2012; PASQUINI, 2018;).

Por fim, há a possibilidade de combinação de modos normais de vibração, que são, então, observados no NIR. Por exemplo, para a molécula triatômica da água, o estiramento simétrico, assimétrico e a deformação angular, todos observados no MIR. Com certa restrição e como consequência da distorção anarmônica, estes modos podem combinar entre si, em um modo misto, cuja energia necessária para ocorrer a vibração ocorre na região do NIR.

Além dos sobretons e bandas de combinação dos modos vibracionais também são observados no NIR:

- ✓ ressonância de Fermi (acoplamento entre uma transição vibracional fundamental e um sobretom) e;
- ✓ ressonância de Darling – Dennison (ressonância entre sobretons de modos vibracionais na região NIR).

Sobretons, bandas de combinação e ressonância, quando acompanhados de uma forte mudança no momento de dipolo da molécula, produzem uma intensa banda de absorção no NIR, como acontece, por exemplo, para a molécula de água.

O grau de anarmonicidade é relevante para a absorção no NIR e, deste modo, quanto maior a anarmonicidade de uma ligação química, maior a probabilidade de

PROSPECTUS
ISSN 2674-8576
DOI: 10.5281/zenodo.10045561 Artigo de Revisão

geração de bandas no NIR (além da necessidade de alta energia dos sobretons e bandas de combinação). As ligações contendo o átomo de hidrogênio atendem os requisitos fundamentais para uma absorção no NIR e, deste modo, ligações C–H, N–H, S–H, O–H são particularmente observadas e, também, ligações de alta energia como C≡C, C=C, C=O.

Não somente a ocorrência das vibrações mencionadas no parágrafo precedente, mas também devido a efeitos relacionados com estes grupos funcionais e sobre suas interações inter e intramoleculares é que possibilitam o uso da espectroscopia NIR em medições qualitativas e quantitativas de diversos sistemas. Por exemplo, os efeitos denominados de primários são decorrentes diretamente do grupo funcional. Assim, o grupo C–H contribui para bandas em menores comprimentos de onda do que o grupo O–H em consequência da diferença de massas entre os átomos e da energia de ligação entre os mesmos. Os efeitos denominados de secundários ocorrem devido às alterações provocadas nos grupos funcionais ao nível atômico, micro e macroscópico. Estas alterações podem alterar as energias de ligação química entre os átomos dos grupos funcionais, provocando mudanças nas frequências de vibração, e, também, na anarmonicidade das ligações, conduzindo a alterações nas intensidades de absorção e nos comprimentos de onda nos quais os grupos funcionais são observados. Um exemplo de efeito secundário é a interação intermolecular de hidrogênio e, ao nível intramolecular, a simetria da molécula. Os efeitos microscópicos são decorrentes, principalmente, ao arranjo estrutural dos átomos em uma molécula. Desse modo, um composto amorfo apresentará um espectro NIR diferente de sua forma cristalina, enquanto ao nível macroscópico os efeitos secundários são correlacionados à diferença de temperatura e alterações mecânicas. O efeito da temperatura é importante, como, por exemplo, para as medições de água, haja vista que a temperatura afeta as ligações de hidrogênio intermoleculares que, por sua vez, causam alterações nos espectros NIR (fenômeno observado em outras substâncias que também apresentam interações intermoleculares). O efeito mecânico pode ser observado em polímeros antes e após dos mesmos terem sido submetidos a estresse mecânico por estiramento porque o

PROSPECTUS
ISSN 2674-8576
DOI: 10.5281/zenodo.10045561 Artigo de Revisão

estresse mecânico altera as cadeias poliméricas, modificando suas forças de interação e proporcionando um espectro NIR diferente (PASQUINI, 2003; LEITÃO, 2012; PASQUINI, 2018; TIBOLA, 2018).

A espectroscopia NIR vem sendo utilizada para análises farmacêuticas e biológicas (LEITÃO, 2012; LIMA, 2013; NEVES, 2013; JUE, 2016). Sua utilização associada à quimiometria em análises de sangue foi reportada (SARGENTELLI; MARTINS, 2020) e, também, seu emprego no estudo de folhas de plantas (MARTINS; SARGENTELLI, 2021).

O NIR também pode estar integrado à imagem hiperspectral (HSI) em que são obtidas imagens com comprimento de onda contínuo com uma resolução menor que 10 nm, onde cada “píxel” (menor unidade de uma imagem digital) da HSI apresenta um espectro de comprimento de onda contínuo. Com as imagens hiperspectrais podem ser obtidos dados espectrais e espaciais de superfícies os quais são importantes para o estudo de propriedades, como, por exemplo, a heterogeneidade dos alimentos e de formulações farmacêuticas e, para amostras biológicas, a morfologia e distribuição dos elementos biológicos ou tecidos. HSIs podem ser alcançadas empregando diferentes técnicas espectroscópicas, todavia, os métodos que empregam a região da espectroscopia NIR apresentam excelente velocidade de aquisição de dados, são não destrutivos, apresentam boa capacidade visualização espacial da composição dos analitos e são não requerem preparo laborioso das amostras (CARVALHO, 2015; SARGENTELLI; MARTINS, 2020).

A espectroscopia NIR é uma técnica com boa sensibilidade, principalmente porque possui adequada relação sinal/ruído, e vem sendo aplicada para elaboração de metodologias analíticas precisas e exatas. Os métodos analíticos resultantes do uso do NIR refletem suas características mais significativas, como: rápido (um minuto ou menos por amostra), não destrutivo, não invasivo, com alta penetração no feixe de radiação da sondagem, adequado para uso em linha, aplicação quase universal e com demandas mínimas de preparação de amostras. Apesar destas vantagens, o uso do NIR só se tornou muito difundido quando a quimiometria é aplicada, permitindo, assim,

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

um elaborado estudo matemático dos dados espectrais, obtendo resultados até então impossíveis de serem obtidos (MARTINS; SARGENTELLI, 2021).

QUIMIOMETRIA

Para compreender o que a quimiometria pode fazer através da aplicação de métodos estatísticos em química, considere um sistema químico constituído de uma mistura de cinco componentes (A, B, C, D, E), em que se quer determinar qual é a melhor razão entre os componentes A e B que maximiza uma determinada propriedade do sistema. O método clássico consiste em fixar A e variar B e, depois, fixar B e variar A, quantificar, e determinar qual é a proporção A/B que melhor traz resultado. Com o uso da quimiometria, o mesmo experimento é realizado, todavia, ocorre a variação de A e B simultaneamente e os resultados são tratados para obter a mesma conclusão.

A definição de quimiometria não é fácil de ser feita e também é complicada por diferenças em decidir o que se chama de campo verbal e comunicação escrita. Ao longo do tempo, alguns autores a consideraram como sendo uma teoria da estimativa, teoria da decisão, otimização, inteligência artificial, análise espectral e de forma de onda e cibernética. Outros a definiram como a aplicação de ferramentas matemáticas e estatísticas à química e, embora algumas técnicas em quimiometria possam ser usadas na teoria química, a quimiometria não visa cálculos teóricos, mas a extração de informações químicas úteis de dados medidos. Ainda outra definição: ciência que relaciona as medições feitas em um sistema, ou processo químico, ao estado do sistema por meio da aplicação de métodos matemáticos ou estatísticos. Provavelmente, devido a objetivos diferentes para o campo desta ciência, a definição de "quimiometria", permanece um pouco incerta ao longo dos anos. Assim, alguns autores ainda preferem defini-la, simplesmente, como uma ferramenta analítica (estatística matemática) empregada em química analítica (WOLD, 1995; BROWN, 2017).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

Devido à complexidade dos cálculos, o desenvolvimento da quimiometria ocorreu praticamente concomitantemente com o campo da ciência da computação e, embora já existissem programas computacionais utilizados em quimiometria, um significativo avanço foi feito com o *software* Matlab (1989) que é, ainda hoje, o software de escolha em quimometria e que vem se aprimorando a cada ano. Por se tratar de uma área de investigação científica, os métodos em quimiometria estão sempre em inovação e aperfeiçoamento. Assim, não é possível parar no tempo em termos de desenvolvimento de ferramentas matemáticas, pois, dependendo do sistema em estudo e o que se quer obter, novos procedimentos estão sendo continuamente elaborados (BROWN, 2017; OTTO, 2017; FERREIRA, 2016).

Dentre as áreas que abrangem a quimiometria, são citadas: calibração multivariada, planejamento e aperfeiçoamento de experimentos, processamento de sinais analíticos, reconhecimento e classificação de padrões, métodos de inteligência artificial, etc. (FERNANDES, 2013; OTTO, 2017;).

Para compreender o que vem a ser calibração multivariada, o termo é dividido, a saber:

- ✓ Calibração: processo que permite estabelecer a relação entre a resposta instrumental (sinal analítico), e uma determinada propriedade (física e/ou química) de uma amostra (analito) e,
- ✓ Multivariada: análise de várias fontes ou múltiplas respostas (por exemplo: medidas de absorbância em vários comprimentos de onda) relacionadas a uma ou mais propriedades desconhecidas de uma amostra.

É comum o fato dos modelos de calibração multivariada ser constituídos em forma de matriz. A linha corresponde a uma amostra (objetos) e cada coluna contém a informação referente ao sinal analítico (variáveis). A matriz que contém as respostas instrumentais (variáveis independentes) é denominada de matriz X, e a matriz contendo parâmetros de referência (variáveis dependentes) de Y. A calibração multivariada permite a determinação simultânea de analitos com maior sensibilidade e confiabilidade, reduzindo o tempo de análise, e pode ser realizada mesmo na

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

presença de interferentes, deste que estejam presentes na etapa de calibração. (OTTO, 2017; FERNANDES, 2013; FERREIRA et al., 1999; FERREIRA, 2016).

Os métodos de calibração multivariada podem ser divididos em "qualitativos" e "quantitativos". Os métodos qualitativos são usados para identificação e comparação de similaridades, e são denominados de Técnicas de Reconhecimento de Padrão (TRP). As TRP são divididas em: "supervisionada" e "não supervisionada". Isto é feito conforme o uso ou não de informações prévias sobre as amostras que constituem o conjunto para construção do modelo. Nas não supervisionadas, a presença de agrupamentos sem o conhecimento prévio do conjunto de amostras (ou das classes) é avaliado e utilizam medidas somente de algumas propriedades. Os segundos, como o próprio nome diz, são mais empregados para a quantificação (OTTO, 2017; FERREIRA, 2016).

Dentre os métodos qualitativos não supervisionados, merecem destaque:

- ✓ Análise de componentes principais (PCA – *Principal Component Analysis*).
- ✓ Análise de agrupamentos por métodos hierárquicos (HCA – *Hierarchical Cluster Analysis*).

Os qualitativos supervisionados englobam:

- ✓ Modelagem Independente e Flexível Por Analogia de Classes (SIMCA – *Soft Independent Modeling of Class Analogy*).
- ✓ Análise Discriminante Linear (LDA – *Linear Discriminant Analysis*).

Entre os métodos quantitativos, há os métodos de regressão:

- ✓ Regressão por quadrados mínimos parciais (PLS – *Partial Least Squares*).
- ✓ Regressão linear múltipla (MLR – *Multiple linear regression*).
- ✓ Regressão por componentes principais (PCR – *Principal Components Regression*).

E o método de Resolução Multivariada de Curvas (MCR – *Multivariate Curve Resolution*).

Já a inteligência artificial pode ser utilizada em análise multivariada tanto “qualitativa” quanto “quantitativa”, em meio a qual se encontram as redes neurais artificiais (ANN - *Artificial Neural Networks*).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

Os princípios que norteiam cada um destes métodos são apresentados resumidamente a seguir. Na sequência, são efetuadas considerações sobre seleção de variáveis intrinsecamente relacionadas com estes métodos.

MÉTODOS QUALITATIVOS NÃO SUPERVISIONADOS

Análise de componentes principais (PCA)

Na modelagem por PCA, novos sistemas de eixos, denominados de componentes principais, são construídos para representar as amostras, nos quais a natureza multivariada dos dados pode ser observada em poucas dimensões. A PCA emprega um conjunto de dados representado por uma matriz: registros e características (correlacionadas). A análise exprime o conjunto em eixos, componentes principais, que se constituem em uma combinação linear das variáveis originais. A equação matemática fundamental, e amplamente difundida, é:

$$s_i^2 = (n-1)\sum(Xij - Xi)^2 \quad (3)$$

Onde:
s_i^2 = variância de i;
n = número de amostras;
ij = valor da variável i em j objetos;
X_i = média de i.

O grau que cada variável linearmente correlacionada é dado por:

$$Cij = \left(\frac{1}{n-1}\right) \sum (X_{im} - X_i)(X_{jm} - X_j) \quad (4)$$

Onde:
C_{ij} = covariança das variáveis i e j;
im = valor da variável i em m objetos;
X_i = média de i;
X_{jm} = valor da variável j em m objetos;
X_j = média de j.

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

A PCA correlaciona os eixos do espaço dimensional para uma nova posição (eixos principais). As novas posições são ordenadas de modo que o eixo principal tem a maior variância, enquanto os demais têm a menor variância consecutivamente. As novas coordenadas no novo sistema de eixos das componentes principais são denominadas pontuações ou escores (*scores*). Cada componente principal é construída a partir da combinação linear das variáveis originais. Os coeficientes da combinação linear (quanto cada variável antiga contribui) são denominados de pesos ou carregamentos (*loadings*). A PCA é o alicerce da maioria dos métodos atuais para o tratamento de dados multivariados para o reconhecimento de padrões/tendências (OLIVEIRA, 2014).

Análise de agrupamentos por métodos hierárquicos (HCA)

A HCA objetiva o agrupamento das amostras em classes, e considera a similaridade dos elementos de uma mesma classe e nas diferenças entre os membros de classes diferentes. A representação gráfica obtida é chamada de dendrograma (dendr(o) = árvore), um gráfico bidimensional independentemente do número de variáveis do conjunto de dados. A HCA consiste, portanto, no tratamento matemático de cada amostra como um ponto no espaço multidimensional descrito pelas variáveis escolhidas. Na HCA também é possível tratar cada variável como um ponto no espaço multidimensional descrito pelas amostras. Quando uma determinada amostra é tomada como um ponto no espaço das variáveis, é possível calcular a distância deste ponto a todos os outros pontos e, assim, construindo uma matriz que descreve a proximidade entre todas as amostras estudadas. O agrupamento de amostras revela as similaridades existentes entre as mesmas, enquanto o agrupamento das variáveis indica uma correlação entre elas. O agrupamento de um conjunto de dados distribuídos em pontos está relacionado com a distância. A mais utilizada é a distância euclidiana em dados de vetores num espaço P-dimensional. Desse modo, a distância (d_{ij}) entre dois pontos x_i e x_j é dada por:

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

$$d_{ij} = ||x_i - x_j|| = [\Sigma(x_{i,k} - x_{j,k})^2]^{1/2} \quad (5)$$

Um dendrograma é construído baseado na matriz de proximidade entre as amostras. Há vários modos de aglomerar matematicamente estes pontos no espaço multidimensional para formar os agrupamentos hierárquicos. Cada um corresponde a um algoritmo específico, em que as informações da matriz são usadas gerar um dendrograma de similaridade, cuja interpretação é fundamentada no fato de que duas amostras próximas devem ter também valores semelhantes para as variáveis medidas. Consequentemente, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. Quando o dendrograma construído é de variáveis, a similaridade entre as variáveis assinala a correlação entre estas variáveis do conjunto de dados avaliado.

Como mencionado, a semelhança está relacionada com a proximidade entre as amostras em relação às variáveis num espaço multidimensional, em que um índice, denominado índice de similaridade, pode ser obtido e expresso como:

$$S_{A-B} = 1 - (d_{A-B}/d_{máx}) \quad (6)$$

Onde:
S_{A-B} = índice de similaridade entre uma amostra A e uma amostra B;
d_{A-B} = distância entre uma amostra A e uma amostra B;
$d_{máx}$ = maior distância entre duas amostras do conjunto de dados.

S_{A-B} pode apresentar valores entre 0 e 1. Quando d_{A-B} é igual a $d_{máx}$, as amostras são muito diferentes e o índice de similaridade é zero, e quando são iguais, S_{A-B} é igual a um (NETO, 2004; CORREIA, 2007; SOUZA, 2010).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

MÉTODOS QUALITATIVOS SUPERVISIONADOS

Modelagem Independente e Flexível Por Analogia de Classes (SIMCA)

A SIMCA é usada para classificar amostras em conjuntos de dados com alta dimensionalidade e utiliza componentes principais para localizar os objetos no espaço multidimensional. Assim, após a modelagem, uma amostra será classificada como pertencente a uma dada classe caso possua as características que permitam ser inserida no espaço multidimensional estabelecido.

Matematicamente, a amostra será considerada pertencente à classe caso o valor do F calculado pela variância da classe for menor que o F crítico. O F calculado é o produto da divisão de z^2 (equação 7) pela variância da classe (FERNANDES, 2013).

$$z^2 = x^2 + y^2 \quad (7)$$

Onde:
x^2 = distância entre a projeção da amostra desconhecida na direção da componente principal e a fronteira da classe;
y^2 = soma das distâncias entre a amostra desconhecida e o eixo da PC.

Análise Discriminante Linear (LDA)

A LDA realiza uma estimativa da combinação linear entre duas ou mais funções discriminantes. A discriminação é conduzida determinando os pesos (as importâncias) das variáveis independentes do melhor conjunto de variáveis. Assim, ocorrerá a minimização da variância entre as amostras pertencentes ao mesmo grupo (classes) e maximização entre as amostras pertencentes a grupos distintos. Portanto, na LDA os cálculos são dirigidos para obter uma separação máxima entre as classes avaliadas.

A equação que separa as classes é apresentada em (8):

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

$$k(x) = a_0 + a_1x_1 + a_2x_2 \quad (8)$$

A LDA está restrita a conjuntos de dados de pequenas dimensões, sendo que problemas de colinearidade podem comprometer a capacidade de generalização. Deste modo, a LDA não é o método mais adequado quando há uma quantidade muito grande de variáveis. Apesar disto, alguns algoritmos vêm sendo utilizados com êxito (PONTES, 2011; WANG, 2012; FERNANDES, 2013; KAUFMANN, 2018).

MÉTODOS QUANTITATIVOS POR REGRESSÃO

Os métodos de regressão visam desenvolver modelos capazes de quantificar uma determinada propriedade de interesse. Os dois principais tipos de modelagem são:

- Modelagem dura (*hard modeling*): é baseado na construção de modelos que descrevem o sistema por completo, e usa princípios físico-químicos e de um conjunto *reduzido* de hipóteses restritivas e,

- Modelagem suave (*soft modeling*): é fundamentada em abundância de informações sobre o comportamento do sistema estudado para gerar os algoritmos matemáticos (FERREIRA, 2016; KAUFMANN, 2018)

De modo geral, os modelos de regressão usam uma variável de resposta, uma propriedade, obtida por um equipamento analítico. Dentre os de modelagem suave, recebem destaque: PLS, MLR e PCR.

Regressão por quadrados mínimos parciais (PLS)

A PLS é um método de calibração multivariada onde não é necessário o conhecimento de todos os componentes da amostra. Como é uma calibração multivariada, a PLS permite analisar problemas complexos, onde se lida com várias variáveis X (independentes) com variáveis de resposta Y (dependentes). O principal

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

objetivo da PLS é prever as variáveis Y a partir das variáveis X, em que se encontram novas variáveis, denominadas de latentes (VL), tanto para matrizes X como para Y. As matrizes X e Y são decompostas em “K” variáveis latentes:

$$X = TP' + E_X = \Sigma^{K}_{k=1}\, t_k\, p'_k + E_x \qquad (9)$$

Onde:
P´= *loadings*;
T = matriz de *scores*;
E_x = matriz dos resíduos da matriz de dados X.

$$Y = UQ' + E_Y = \Sigma^{K}_{k=1}\, u_k\, q'_k + E_y \qquad (10)$$

Onde:
Q´= *loadings*;
U = matriz dos *scores*;
E_y = matriz dos resíduos da matriz resposta de Y.

Uma relação linear é estabelecida entre os *scores* X e Y:

$$u_K = b_K\, t_K \qquad (11)$$

Onde:
b_k = vetor dos coeficientes de regressão para cada um dos fatores.

A matriz dos coeficientes de regressão b é determinada pela equação (12):

$$Y = TBQ' + E_y \qquad (12)$$

Para obter os parâmetros de um modelo PLS estão disponíveis vários algoritmos. Entre os mais utilizados, está o Algoritmo dos Mínimos Quadrados Parciais Iterativos Não – Lineares (NIPALS – *Nonlinear Iterative Partial Least Squares*). É preciso mencionar que o mesmo resultado deveria ser obtido para qualquer algoritmo

PROSPECTUS
ISSN 2674-8576
DOI: 10.5281/zenodo.10045561 Artigo de Revisão

utilizado para obter a regressão PLS. Entretanto, foi constatado que, do ponto de vista numérico, existem diferenças no resultado. As diferenças estão associadas à natureza dos dados, ao número de fatores PLS empregados e à precisão usada nos cálculos.

A PLS pode ser utilizada para determinação de um ou de vários analitos na matriz de dados Y. Mas, a técnica realiza a regressão no domínio dos dados transformados, portanto, não possibilita uma interpretação físico-química direta dos resultados (FERNANDES, 2013).

Regressão linear múltipla (MLR)

A MLR estabelece uma relação linear entre as matrizes dos dados instrumentais localizada em X e a propriedade analisada na matriz Y. A MLR aplica o método dos mínimos quadrados, sendo que a equação básica é dada por:

$$Y = Xb_{MLR} + E \quad (13)$$

Onde:
y = matriz da propriedade analisada (por exemplo: concentração);
X = matriz dos dados instrumentais (por exemplo: espectros);
b = vetor, calculado no estágio da calibração e utilizando os métodos dos mínimos quadrados ordinários. É a matriz dos coeficientes de regressão.
E = resíduo não modelado em Y (matriz de erros).

A matriz b é dada por:

$$b_{MLR} = (X'X)^{-1}X'y \quad (14)$$

A equação (14) necessita do procedimento de inversão da matriz (X´X). Esta operação apresenta alguns fatores limitantes:

- O número de amostras necessita ser no mínimo igual (ou superior) ao número de variáveis. Caso isto não ocorra, acarretará indeterminação do sistema;

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

- As variáveis em X precisam ser vetores que sejam linearmente independentes, pois, de outro modo, conduziria a uma matriz singular.

Uma das vantagens da MLR é que não requer a obrigatoriedade do conhecimento de todas as propriedades das espécies investigadas, sendo que os interferentes podem ser tratados. A principal desvantagem é que o número de amostras limita o número de variáveis. Assim, necessita o emprego prévio de seleção de variáveis (FERNANDES, 2013; KAUFMANN, 2018).

Regressão por componentes principais (PCR)

A PCR usa os fundamentos da PCA e, de modo diferente da MLR, não requer a seleção de variáveis porque utiliza uma transformação ortogonal da matriz X a fim de obter um novo conjunto de variáveis linearmente independentes. A matriz X é decomposta em duas outras matrizes menores: *scores* (T) e *loadings* (P) e a equação fundamental é dada por:

$$X = T*P' + E \quad (15)$$

Para obter uma equação de regressão entre a propriedade que se quer determinar (y) e a matriz de *scores* (T), a equação abaixo é utilizada:

$$y = Tb_{PCR} + F \quad (16)$$

Onde:
T = *scores* (anteriormente obtido pela equação (15));
b_{PCR} = matriz dos coeficientes de regressão;
F = resíduos não modelados.

Um algoritmo que pode ser utilizado para obter T é o NIPALS (FERNANDES, 2013).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

MÉTODO QUANTITATIVO POR RESOLUÇÃO MULTIVARIADA DE CURVAS (MCR)

Outra técnica quimiométrica é a Resolução Multivariada de Curvas (MCR – *Multivariate Curve Resolution*). É um método de processamento de sinais analíticos com o intuito de resolver misturas de sinais. O método de MCR apresenta o pré-requisito de que, em um conjunto de dados, a resposta do sistema seja linear em relação à quantidade de analito. Os principais objetivos do MCR são o isolamento, a resolução e a quantificação das fontes de variação do conjunto de dados. O modelo bilinear MCR decompõe uma matriz de dados de uma amostra conforme a Equação 17 (TAULER, 1995):

$$\boldsymbol{X} = \boldsymbol{CS^T} + \boldsymbol{E} \quad (17)$$

Na equação 17, tem-se o modelo bilinear MCR. X (i x j) é a matriz de dados obtida experimentalmente, a matriz C (i x f) é, geralmente, associada com perfis de concentração, S (j x f) é a matriz associada com os perfis experimentais puros para cada analito, E é a matriz de resíduos e f o número de componentes matemáticos, o qual, idealmente, deve ser igual ao número de ingredientes presentes na amostra.

Os parâmetros do modelo são estimados utilizando os mínimos quadrados alternantes (ALS – *Alternating Least Squares*), que iterativamente ajusta as matrizes C e S ao conjunto de dados X, utilizando um número de fatores f pré-definidos e uma estimativa inicial de C ou S, a qual pode ser obtida de diferentes maneiras, entre as quais, a análise de fatores evolucionários (EFA – *Evolving Factors Analysis*), a decomposição de valores singulares (SVD – *Singular Value Decomposition*) e conhecimento prévio do sistema. O número de analitos pode ser estimado utilizando o conhecimento prévio do sistema em estudo ou a partir dos resultados de análise por SVD da matriz de dados constituída das amostras utilizadas na etapa de calibração (IZQUIERDO-RIDORSA, 1997).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

A imposição de uma ou mais restrições ao MCR, entre as quais, não-negatividade, unimodalidade, posto local, trilinearidade e igualdade, pode ser utilizada para ajudar na convergência do algoritmo, resolver problemas de ambiguidade rotacional e garantir que os resultados apresentem significado químico. A restrição de trilinearidade consiste em forçar que os perfis instrumentais/espectrais e de concentração não variem de amostra para amostra, ou seja, no caso de análise cromatográfica, por exemplo, não poderiam ocorrer variações nos tempos de retenção. Na restrição de igualdade, o perfil instrumental ou de concentração é fixado durante a otimização do modelo, esta restrição pode ser aplicada no caso de conhecer previamente o perfil experimental ou de concentração do sistema em estudo. A restrição de posto local é utilizada quando há o conhecimento prévio que determinado analito está ausente em uma ou mais amostras. Neste caso, esta informação é passada ao modelo fazendo com que o perfil de concentração deste analito seja igual a zero nas respectivas amostras em que ele estiver ausente. Esta restrição é muito útil nos casos em que amostras contendo interferentes são decompostas conjuntamente como amostras de calibração, onde é requerido que o perfil de concentração do interferente seja zero.

O MCR funciona, portanto, modelando a matriz de dados como o produto de duas matrizes de fatores, uma que contém as informações espectrais de cada componente e outra que contém as informações de concentração ou pureza de cada componente em cada amostra.

O MCR é um método poderoso que permite a análise de dados altamente complexos e é amplamente utilizado em quimiometria para análise de amostras em muitas áreas, incluindo química ambiental, ciência dos alimentos, farmacologia e análise de polímeros.

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

REDES NEURAIS ARTIFICIAS (ANN)

Redes neurais artificiais (ANN – *Artificial Neural Networks*) são uma classe de modelos de aprendizado de máquina que têm sido amplamente utilizados em quimiometria para resolver problemas de regressão e classificação. Esses modelos são inspirados no funcionamento do cérebro humano e se constituem em camadas de neurônios artificiais interconectadas.

Sob a ótica computacional, pode-se dizer que em um neurônio é realizado o processamento sobre uma ou, geralmente, várias entradas, de modo a gerar uma saída. O neurônio artificial é uma estrutura lógico-matemática que simula a forma, o comportamento e as funções de um neurônio biológico. (Os neurônios são compostos de três partes principais: dendritos, um corpo celular e um axônio. Os sinais são recebidos através dos dendritos, viajam para o corpo celular e continuam para o axônio até atingir a sinapse (o ponto de comunicação entre dois neurônios)).

Pode-se, simplificadamente, associar o dendrito à entrada, a soma ao processamento e o axônio à saída; portanto, o neurônio é considerado uma unidade fundamental processadora de informação. Os dendritos são as entradas, cujas ligações com o corpo celular artificial são realizadas por de canais de comunicação que estão associados a um determinado peso (simulando as sinapses). Os estímulos captados pelas entradas são processados pela função do soma, e o limiar de disparo do neurônio biológico é substituído pela função de transferência (FURTADO, 2019).

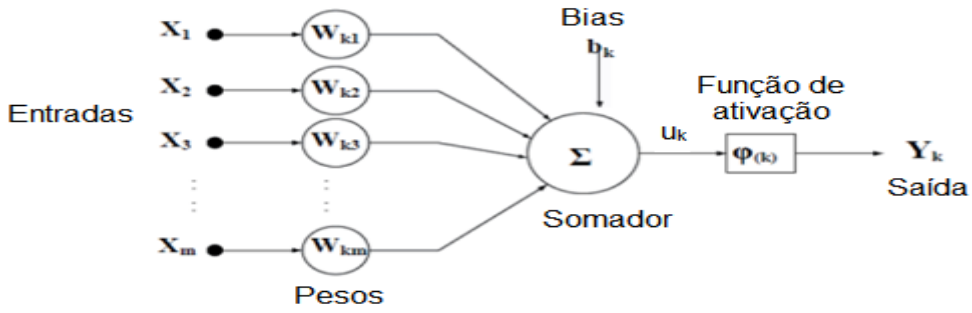
Foi apresentado em 1943 o primeiro modelo matemático do neurônio (McCULLOCH; PITTS, 1943). Esse modelo, conhecido como MCP (McCulloch-Pitts), é uma simplificação do neurônio biológico, que considera o neurônio como uma unidade de processamento de informações binárias, com várias entradas binárias e uma única saída binária. O modelo MCP foi um marco importante no desenvolvimento das ANN, porque mostrou que um sistema de neurônios interconectados, com apropriadas regras de conexão, poderia ser usado para realizar tarefas de processamento de informações. O modelo foi posteriormente aprimorado, em que foi

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

denominado de perceptron (ROSENBLATT, 1958). A Figura 1 ilustra o modelo de neurônio artificial (HAYKIN, 2001; SOARES; SILVA, 2011; ARAÚJO et al., 2012):

Figura 1. Modelo de neurônio artificial.



Fonte: (HAYKIN, 2001; SOARES; SILVA, 2011; ARAÚJO et al., 2012).

No modelo de neurônio, os termos X_m são as entradas da rede; os W_{km} são os pesos, ou pesos sinápticos, associados a cada entrada; u_k é a combinação linear dos sinais de entrada; $\varphi_{(k)}$ é a função de ativação e Y_k é a saída do neurônio. Pode-se dizer que é nos pesos que reside todo o conhecimento adquirido pela rede. Para uma maior fidelidade ao modelo biológico e flexibilidade computacional, além da excitação vinda das entradas da rede ou saídas de outros neurônios, cada neurônio é também excitado por uma polarização constante chamada "*bias*", b_k, constante de valor 1, transmitida ao neurônio através da sinapse (SOARES; SILVA, 2011; FURTADO, 2019). O modelo pode ser representado matematicamente por (FURTADO, 2019):

$$u_k = \sum_{j=1}^{m} W_{kj} \cdot X_j \qquad (18)$$

$$Y_k = \varphi (u_k) \qquad (19)$$

Onde:
X_j = sinais de entrada;
W_{kj} = pesos sinápticos ou pesos;
Y_k = sinais de saída;
$\Phi_{(k)}$ = função de ativação.

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

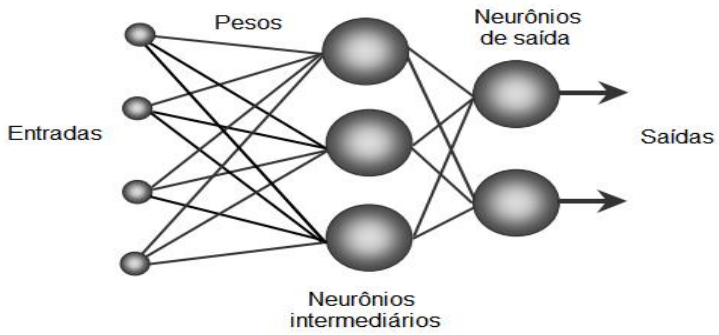
Com o termo b_k, a equação matemática (19) torna-se na (20):

$$Y_k = \varphi (u_k + b_k) \quad (20)$$

O modelo se apresenta constante para quase todas as ANN, variando somente a função de ativação, que limita a amplitude do sinal de saída do neurônio. Normalmente a faixa de saída está em um intervalo fechado [0, 1] ou alternativamente em [-1, 1], podendo também este intervalo de saída estar entre (- ∞, + ∞). Entre os diversos tipos de funções de ativação, as mais comuns são: sinal (que produz uma saída binária), linear (saída linear contínua) e sigmoide (saída não-linear contínua) (FERNANDES, 1999; BARRETO, 2002; FURTADO, 2019).

Assim, a junção de vários neurônios artificiais em camadas interligadas constitui as ANN, nas quais as conexões transformam o sinal de saída de um neurônio em entrada de outro, ou orientam o sinal de saída para o exterior. Um tipo de rede muito utilizado é a Perceptron de Múltiplas Camadas (MLP – *Multi-Layer Perceptron*), uma modificação da Perceptron de Camada Única (*Single-Layer Perceptron*). Uma rede MLP contém várias camadas "alimentada a diante" (*feedfoward*) (ARAÚJO et al., 2012; FURTADO, 2022). Uma representação de ANN-MLP é apresentada na Figura 2, na qual os neurônios intermediários também são conhecidos como camada oculta (ARAÚJO et al., 2012; FURTADO, 2022).

Figura 2. ANN-MLP.



Fonte: (FURTADO, 2022).

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

As ANN podem ser usadas para modelar relações não lineares entre entradas e saídas, o que é particularmente útil em quimiometria, onde os relacionamentos entre as variáveis podem ser complexos e não lineares. Além disso, esses modelos conseguem lidar com dados com alta dimensionalidade, incompletos e/ou ruidosos, incluindo, ainda, associação com a PCA (FERREIRA, 2022).

SELEÇÃO DE VARIÁVEIS

Não obstante, é preciso mencionar ainda que, para a obtenção de um modelo robusto para tratamento dos dados em quimiometria, é necessária uma seleção de variáveis. Isto porque as dimensões do conjunto de dados são extensas, apresentam poucas amostras ou muitas variáveis.

De modo geral, as técnicas de seleção de variáveis consideram que um pequeno número de variáveis é suficiente para os cálculos e, deste modo, removem variáveis que não trazem informação significativa do sistema em estudo. Assim, minimizam a multicolinearidade (que prejudica tanto o reconhecimento de padrões quanto as regressões). Deste modo, a seleção de variáveis contribui para a exatidão dos métodos quimiométricos.

Algoritmos são utilizados para selecionar variáveis, dentre os quais:

- ✓ Algoritmo de Projeções Sucessivas (SPA – *Successive Projections Algorithm*). É um algoritmo que faz uso da seleção de frente (*forward*), em que se inicia com uma variável e, em seguida, a cada interação uma nova variável, com uma menor multicolinearidade possível em relação às investigadas, é incorporada. Utilizado principalmente em MLR. Neste método, primeiro, subconjuntos de variáveis considerando o conceito de menor multicolinearidade são selecionados. Numa segunda etapa, o subconjunto que apresente o melhor resultado em relação ao critério utilizado é escolhido, e a Raiz Quadrada do Erro Médio Quadrático de Validação (RMSEV – *Root Mean Square Error of Validation*) é utilizada. Por fim, o subconjunto escolhido é submetido a uma

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

análise para averiguar a possibilidade de alguma variável ser eliminada sem prejuízo da capacidade de predição.

- ✓ Intervalo Parcial Mínimo Quadrado (iPLS – *Interval Partial Last Square*). Muito utilizado em espectroscopias. Neste algoritmo, os espectros são fragmentados em regiões distantes entre si. Cada região é tratada por PLS, em que são obtidos modelos em que se calcula o RMSEV (e também para todo o espectro). A região que obtiver menor RMSEV apresenta menor erro para a predição da variável de interesse e tem, portanto, as variáveis selecionadas.

- ✓ Algoritmo *Stepwise*. Este algoritmo é baseado no princípio: "Seleção para frente" (*Forward Selection*) e "eliminação para trás" (*Backward Elimination*). Em *Forward Selection*, inicia-se com uma variável x que apresente a melhor correlação com y, sendo outras variáveis adicionadas em seguida. Após cada adição de uma variável, um teste F é executado. A variável que apresentar o melhor F calculado (F_{cal}) fica no modelo. Em *Backward Elimination*, inicia-se com todas as variáveis e, pouco a pouco, variáveis são retiradas. A cada retirada, um teste F é realizado. A variável que apresentar o menor valor de F_{cal} é retirada do modelo. Os cálculos continuam até que não exista variável com F_{cal} maior que o F crítico (F_c). (F_c pode ser um tabelado ou obtido empiricamente do sistema em estudo). Para evitar que cada variável seja de novo recolocada no cálculo, o algoritmo inclui uma fase de inclusão e outra de exclusão, ambas guiadas por teste F, que é calculado em que o Erro Quadrático Médio de Predição (RMSEP – *Root Mean Square Error of Prediction*) é utilizado (FERNANDES, 2013; KAUFMANN, 2018).

As equações para RMSEV e RMSEP podem ser expressas como (KAUFMANN, 2018):

$$RMSEV = (\Sigma^{n}_{i=1} (y_{previsto} - y_{medido})^2)^{1/2}/n \quad (21)$$

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

$$RMSEP = (\Sigma^{k}_{i=1} (y_{previsto} - y_{medido})^2)^{1/2}/k \quad (22)$$

Onde:
$y_{previsto}$ = valor da propriedade de interesse prevista pelo modelo;
y_{medido} = valor da propriedade de interesse medida para a amostra;
n = número de amostras utilizado na calibração;
k = número de amostras utilizado na validação.

CONSIDERAÇÕES FINAIS

O NIR é uma espectroscopia vibracional com energia relativamente alta em que são analisadas moléculas contendo, principalmente, os grupos funcionais C–H, N–H, O–H e S–H, e é mais utilizada em análises quantitativas. Os métodos analíticos resultantes do uso do NIR refletem suas características mais significativas, como: rápido (um minuto ou menos por amostra), não destrutivo, não invasivo, com alta penetração no feixe de radiação da sondagem, adequado para uso em linha, aplicação quase universal e com demandas mínimas de preparação de amostras. Apesar dessas vantagens, o uso do NIR só se tornou difundido pelo desenvolvimento da computação e pela disciplina de quimiometria. A quimiometria envolve a aplicação de métodos estatísticos e matemáticos, bem como aqueles baseados na lógica matemática, à análise química, fornecendo as ferramentas para a coleta de informações e seu uso racional.

Dentro da quimiometria, o HCA e PCA permitem a visualização gráfica de todo o conjunto de dados, mesmo quando o número de amostras e variáveis é elevado. O uso desses algoritmos visa principalmente aumentar a compreensão do conjunto de dados, examinando a presença ou ausência de agrupamentos naturais entre as amostras. Ambos são classificados como exploratórios ou não supervisionados, visto que nenhuma informação com relação à identidade das amostras é considerada. A PCA e a HCA são técnicas de análise multivariada com fundamentos teóricos distintos, e podem ser aplicadas independentemente, podendo, até mesmo, ser

PROSPECTUS
ISSN 2674-8576

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

complementares. A apresentação dos resultados experimentais na forma de gráficos facilita a interpretação dos dados e, consequentemente, a identificação de grupos de amostras com características parecidas. É possível, também, verificar quais parâmetros são responsáveis pela formação dos grupos de amostras. Os métodos supervisionados empregam o conceito de componentes principais, sendo que, dentre os que foram aqui abordados, a SIMCA encontra maior utilização.

Entre os métodos quantitativos citados, cada qual apresenta vantagens e desvantagens. Como visto, o MLR tem alguns fatores limitantes como: o número de amostras precisa ser no mínimo igual (ou superior) ao número de variáveis e, estas, em X, precisam ser vetores que sejam linearmente independentes. Isto limita um pouco o uso do MLR, especialmente em espectroscopia NIR. Já a PCR usa os fundamentos da PCA e, de modo diferente da MLR, não requer a seleção de variáveis porque utiliza uma transformação ortogonal da matriz X a fim de obter um novo conjunto de variáveis linearmente independentes. Do ponto de vista da espectroscopia NIR, quando todas as substâncias que apresentam sinais espectroscópicos não são conhecidas, onde é amplo o número de variáveis, é comum utilizar PLS na construção de modelos de regressão, pois este fornece os melhores resultados e não há necessidade do emprego da PCA. Entretanto, o PLS considera que a relação espectro/propriedade seja linear, o que nem sempre é garantido em dados dessa natureza e o que pode influenciar no rigor do modelo. Alternativamente, as ANN associadas à PCA possuem a vantagem de serem eficientes em lidar com dados não lineares, incompletos e/ou ruidosos. Uma metodologia quimiométrica adequada é, ao ter dúvidas em relação ao espectro/propriedade, tratar os dados via PLS e checá-los, empregando as ANN associadas à PCA, embora isso seja trabalhoso.

REFERÊNCIAS

ARAÚJO, M. S. S.; FARIAS, C. T. T.; OLIVEIRA, A. S. A. Detecção e classificação de defeitos em chapas de aço carbono utilizando ondas ultrassônicas guiadas de Lamb

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

e redes neurais artificiais, **VII Congresso Norte Nordeste de Pesquisa e Inovação**, Palmas-TO, 2012.

BARRETO, J. M. **Introdução às redes neurais artificiais.** Laboratório de Conexionismo e Ciências Cognitivas, Universidade Federal de Santa Catarina – UFSC, Departamento de Informática e de Estatística, Florianópolis-SC, 2002.

BROWN, S. D. The chemometrics revolution re-examined, **Journal of Chemometrics**, v. 31, e2856, p. 1 – 23, 2017.

CARVALHO, M. A. **Utilização da análise de imagem hiperspectral no infravermelho próximo para a identificação de marcadores luminescentes a base de redes metalorgânicas MOF**. Dissertação de Mestrado. Universidade Federal de Pernambuco – UFPE, Centro de Ciências Exatas e da Natureza, Recife-PE, 91 p., 2015.

CORREIA, P. R. M., FERREIRA, M. M. C. Reconhecimento de padrões por métodos não supervisionados: Explorando procedimentos quimiométricos para tratamento de dados analíticos, **Química Nova**, v. 30, n. 2, 481–487, 2007.

FERNANDES, M. A. C. **Redes Neurais Artificiais Aplicadas à Detecção Inteligente de Sinais**. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte – UFRN, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Laboratório de Engenharia de Computação e Automação, Natal- RN, 100 p., 1999.

FERNANDES, D. D. S. **Espectroscopia UV-VIS para avaliação de biodiesel e misturas de biodiesel/diesel**. Dissertação de Mestrado. Universidade Estadual da Paraíba, UEPB, Programa de Pós – Graduação em Ciências Agrárias, Campina Grande-PB, 70 p., 2013.

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

FERREIRA, M. M. C.; ANTUNES, A. M.; MELGO, M. S.; VOLPES, P. L. O. Quimiometria I: Calibração multivariada, um tutorial, **Química Nova**, 22, 5, 724–731, 1999.

FERREIRA, M. M. C. **Quimiometria**. 1.ª Edição. Campinas – SP. Editora da Unicamp, 2016.

FERREIRA, R. A. **Redes neurais artificiais com componentes principais para a construção de modelos de predição em dados de espectroscopia NIR**. Tese de Doutorado (Doutorado em Estatística Aplicada e Biometria) – Universidade Federal de Viçosa, Viçosa-MG, 72 p., 2022.

FORATO, L. A., et al. **A espectroscopia na região do Infravermelho e algumas aplicações**, São Carlos-SP, Embrapa Instrumentação Agropecuária, 2010.

FURTADO, M. I. V. **Redes neurais artificiais:** uma abordagem para sala de aula, Ponta Grossa-PR, Editora Atena, 2019.

HAYKIN, S. **Redes neurais:** princípios e prática. 2.ª Edição, Porto Alegre-RS, Bookman, 2001.

JUE, T., MASUDA, K. **Application of near infrared spectroscopy in biomedicine**, Canada, Kobo Editions, 2016.

KAUFMANN, K. C. **Utilização de espectrofotômetro NIR portátil para avaliar a qualidade e a autenticidade de óleos vegetais comestíveis empregando métodos quimiométricos**. Dissertação de Mestrado. Universidade Estadual de Campinas – UNICAMP, Faculdade de Engenharia de Alimentos, Campinas-SP, 118 p., 2018.

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

LEITÃO, T. M. D**. Aplicações da espectroscopia de infravermelho próximo em Ciências Farmacêuticas**. Dissertação de Mestrado. Universidade Fernando Pessoa – UFP, Faculdade de Ciências da Saúde, Porto – Portugal, 71 p., 2012.

LIMA, L. A. S. **Estudo da potencialidade da espectroscopia de infravermelho próximo na análise de cabelo utilizando ferramentas quimiométricas para diferenciar fumantes de não fumante**. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte – UFRN, Centro de Ciências Exatas e da Terra, Instituto de Química, Natal- RN, 42 p., 2013.

McCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 4, p. 115–133, 1943.

MARTINS, J. A.; SARGENTELLI, V. Infravermelho próximo na avaliação quantitativa foliar, **Prospectus**, v. 3, n. 1, p. 33– 55, 2021.

NAKAMOTO, K. **Infrared spectra of inorganic coordination compounds**, 2nd. Edition. Wiley-Interscience, 1970.

NETO, J. M. M. **Estatística Multivariada –** Uma visão didática – metodológica, **Crítica**, 1–13, 2004.

NEVES, A. C. O. **Espectroscopia no infravermelho próximo e métodos de calibração multivariada aplicados à determinação simultânea de parâmetros bioquímicos em plasma sanguíneo**. Dissertação de Mestrado. Universidade Federal do Rio Grande do Norte – UFRN, Centro de Ciências Exatas e da Terra, Instituto de Química, Natal- RN, 107 p., 2013.

OLIVEIRA, A. D. P. **Utilização de métodos quimiométricos para análise quantitativa de glibenclamida comprimido utilizando as espectroscopias de**

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

infravermelho próximo e Raman – Desenvolvimento e validação de uma estratégia PAT. Dissertação de Mestrado. Universidade Federal de Pernambuco – UFPE, Centro de Ciências da Saúde, Departamento de Ciências Farmacêuticas, Pós – Graduação em Ciências Farmacêuticas, Recife-PE, 99 p., 2014.

OTTO, M. **Chemometrics:** Statistics and computation application in analytical chemistry, 3rd. Edition, Weinheim, Wiley-VHC Verlag GmbH & Co. KGaA, 2017.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, practical aspects and analytical applications, **Journal Brazilian Chemical Society**, v. 14, n. 2, p. 198–219, 2003.

PASQUINI, C. Near Infrared Spectroscopy: A mature analytical technique with new perspective – A review, **Analytical Chemistry**, v. 1026, p. 8–36, 2018.

PONTES, M. J. C. et al. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectroscopy and multivariate classification, **Talanta**, v. 85, n. 4, p. 2159–2161, 2011.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, 1958.

SARGENTELLI, V., MARTINS, J. A. Near infrared and chemometrics applied to non-invasive in vivo blood analysis: Overview of last twenty years of development, **International Journal of Engineering Development and Research**, v. 8, n. 1, p. 570–576, 2020.

SARGENTELLI, V., MARTINS, J. A. Near infrared hyperspectral imaging spectroscopy applied to investigation of plant leaves: A brief review, **International Journal of Science and Research**, v. 9, n. 4, p. 1055–1062, 2020.

DOI: 10.5281/zenodo.10045561 Artigo de Revisão

SILVERSTEIN, R. M., BASSLER, C. C., MORRILL, T. C. **Spectrometric Identification of Organic Compounds**, New York, John Wiley & Sons, 1969.

SOARES, P. L. B.; SILVA, J. P. Aplicação de Redes Neurais Artificiais em Conjunto com o Método Vetorial da Propagação de Feixes na Análise de um Acoplador Direcional Baseado em Fibra Ótica, **Revista Brasileira de Computação Aplicada**, v. 3, n. 2, p. 58–72, 2011.

SOUZA, G. S. **Avaliação da bacia hidrográfica do rio Paraguaçu utilizando análise multivariada**. Dissertação de Mestrado. Universidade Federal da Bahia – UFBA, Instituto de Química, Programa de Pós – Graduação em Química, Salvador-BA, 112 p., 2010.

TIBOLA CS, et al. **Espectroscopia no infravermelho próximo para avaliar indicadores de qualidade tecnológica e contaminantes em grãos**. Brasília – DF: Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA, 2018.

WANG, S. et al. Application of hybrid image features for fast and non – invasive classification of raisin, **Journal of Food Engineering**, v. 109, p. 531–537, 2012.

WOLD, S. Chemometrics; what do we mean with it, and what do we want from it? **Chemometrics and Intelligent Laboratory Systems**, v. 30, p. 109–115, 1995.

Os autores declararam não haver qualquer potencial conflito de interesses referente a este artigo.