SELEÇÃO DE COMPRIMENTOS DE ONDA PARA PREDIÇÃO DE PROPRIEDADES DE BIODIESEL/DIESEL APOIADA NA LEI DE LAMBERT-BEER

Juliana de Abreu Fontes*
Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Av. Osvaldo Aranha, 99, Porto Alegre, RS, Brasil.
juliana.abreu.fontes@gmail.com

João Batista Gonçalves de Brito
Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Av. Osvaldo Aranha, 99, Porto Alegre, RS, Brasil.
joao.brito@ufrgs.br

Michel José Anzanello
Departamento de Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Av. Osvaldo Aranha, 99, Porto Alegre, RS, Brasil.
michel.anzanello@gmail.com

Guilherme Brandelli Bucco
Escola de Administração, Universidade Federal do Rio Grande do Sul, Rua Washington Luiz, 855 - Centro Histórico, Porto Alegre – RS, Brasil.
gbbucco@gmail.com

RESUMO

O objetivo deste artigo é propor um método de seleção de COs com vistas à predição de propriedades de biodiesel baseado em conceitos da lei de Lambert-Beer, a qual é utilizada para ranquear os COs de acordo com sua relevância. A ferramenta preditiva utilizada é a rede neural artificial (Artificial Neural Network – ANN). Para tanto, foram analisados três bancos de dados espectrais de diesel descritos por espectroscopia no infravermelho próximo (NIRS). Tais dados foram usados para predizer propriedades do combustível que impactam na sua qualidade. O desempenho de predição foi avaliado por intermédio do erro quadrático médio (Root Mean Square Error – RMSE) no conjunto de teste. O método proposto reduziu significativamente o número de COs necessários para geração do modelo e melhorou o desempenho preditivo, quando comparado à utilização da totalidade de COs disponíveis.

PALAVRAS CHAVE. Dados espectrais, Diesel, Seleção de comprimentos de onda.

- Amostras descritas por um grande número de comprimentos de onda tendem a prejudicar o desempenho de algoritmos de classificação.
- Propõe-se um índice de importância de variáveis com base na em conceitos da lei de Lambert-Beer;
- A estrutura proposta apresentou resultados significativamente superiores aos métodos tradicionais de seleção trazidos pela literatura.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

ABSTRACT

The objective of this article is to propose a wavelengths selection method with a view to predicting biodiesel properties based on concepts from the Lambert-Beer law, which is used to rank wavelengths according to their relevance. The predictive tool used is the artificial neural network (Artificial Neural Network - ANN). For this purpose, three diesel spectral databases analyzed by near infrared spectroscopy (NIRS) were analyzed. Such data were used to make predictions of fuel properties that impact the determination of its quality. The prediction performance of the method was evaluated by means of the mean square error (Root Mean Square Error - RMSE) in the test set. The proposed method significantly reduced the number of COs and improved the predictive performance, when compared to the use of all available COs.

KEYWORDS. Spectral data, Diesel, wavenumber selection.

- Samples described by a large number of wavelengths tend to reduce the performance of classification algorithms;
- We proposed a variable importance index based on the Lambert-Beer Law;
- The proposed framework outperformed traditional methods for wavenumber selection from the literature.

1 INTRODUÇÃO

No Brasil, o óleo diesel é o derivado de petróleo mais consumido por conta de sua predominância no transporte rodoviário de passageiros e de carga. A grande demanda do mercado tem instigado esforços por lucros ilegais por meio de adulterações deste produto. Os efeitos da falta de conformidade deste combustível estão fortemente ligados ao desempenho dos motores e ao aumento nas emissões de poluentes (CÂMARA et al., 2017; CUNHA et al., 2019; KRISHNA et al., 2019; LI et al., 2005). Cunha et al. (2019) enumeram diversos problemas oriundos da adulteração de combustíveis: falha repentina do motor, dificuldade de partida, aumento do consumo de combustível, baixa taxa de pulverização de combustível na câmara de combustão e aumento das emissões de material particulado.

Ponto de ebulição, número de cetano e ponto de fulgor são algumas das propriedades físico-químicas utilizadas pela indústria para monitorar a qualidade do diesel e garantir a queima adequada nos motores (ALEME; BARBEIRA, 2012; LI et al., 2005). Nos últimos anos, técnicas analíticas como espectroscopia no infravermelho próximo (*Near-Infrared Spectroscopy* – NIR) e transformada de Fourier (*Fourier-transform infrared* – FTIR) foram amplamente adotadas como ferramentas analíticas em diferentes campos e com diversas finalidades (SOARES et al., 2017). Estas técnicas podem ser usadas para identificar um composto ou investigar a composição de uma amostra, a fim de identificar possíveis adulterações (CÂMARA et al., 2017; GAYDOU; KISTER; DUPUY, 2011). Palou et al. (2017) descrevem a técnica como um atrativo recurso para a indústria do petróleo, uma vez que contribui para o alcance da qualidade exigida em seus produtos.

Entretanto, técnicas analíticas apoiadas em espectroscopia usualmente geram conjuntos de dados compostos por centenas ou milhares de comprimentos de onda (COs) altamente correlacionados.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Quando tais dados são usados como preditores em modelos matemáticos, percebe-se redução no desempenho preditivo de técnicas multivariadas no que diz respeito à velocidade e taxa de acerto (KAHMANN et al., 2018). Cada variável passa a ser vista como uma coordenada *d*-dimensional no espaço; à medida que o número de dimensões aumenta, as instâncias ficam mais dispersas no espaço. Quando se tem menos instâncias (amostras) por região, a tarefa de aprendizado se torna mais difícil. Isso ocorre porque os algoritmos de aprendizado de máquina constroem preditores com base nas proporções estimadas de instância em cada classe por regiões do espaço. Diante disso, ocorre um aumento da demanda por técnicas de seleção de COs que permitam executar a análise dos dados sem a perda de informações relevantes, diminuindo o risco de inferências não confiáveis e do custo computacional de análises (GAUCHI; CHAGNON, 2001).

Este artigo apresenta um novo método para determinação das regiões mais relevantes do espectro com vistas à predição de propriedades de amostras de diesel. O método propõe o ranqueamento dos COs baseado na lei de *Lambert-Beer*, o que reduz significativamente o número de possibilidades de combinações de variáveis a serem testadas em cada subconjunto (uma vez que, para n variáveis, tem-se $2^n - 1$ combinações possíveis). A Lei de *Lambert-Beer* estabelece uma relação linear entre a absorbância e a concentração de substâncias químicos presentes na amostra (GOBRECHT et al., 2015). A absorbância é uma medida da "quantidade" de luz absorvida pela amostra, sendo influenciada por variáveis como natureza do solvente, pH da solução, temperatura, concentração de eletrólitos e presença de substâncias interferentes. Por esse motivo, quando dados de origem espectral com o objetivo de prever características como concentração e temperatura são analisados, buscam-se COs com maior variabilidade para explicar a variável dependente. De tal forma, em vez de analisar os dados (amostras) em sua totalidade, este artigo propõe considerar apenas duas faixas das amostras: o intervalo que contém as observações cuja variável dependente apresenta os menores valores, e aquele cujas observações apresentam variável dependente com os maiores valores. As diferenças entre tais faixas são usadas na geração de um índice de importância dos COs. Na sequência, subconjuntos de variáveis são gerados usando a inserção *forward*, uma de cada vez, iniciando a partir de um conjunto vazio, de acordo com uma estratégia de seleção de busca sequencial. Nesta estratégia, as variáveis são adicionadas sequencialmente a um conjunto de candidatas de acordo com a ordem sugerida pelo índice de importância até que a adição de outras variáveis não melhore o critério de avaliação (YUN et al., 2019). Nas proposições deste artigo, ANN são usadas como ferramenta de predição. Finalmente, o método seleciona as regiões mais apropriadas do espectro para predição de três importantes propriedades do diesel: ponto de ebulição, número de cetano e ponto de fulgor.

Embora a literatura ofereça várias estruturas para identificar as variáveis mais informativas, muitas delas normalmente não consideram a não linearidade das relações entre as variáveis. De acordo com Lee et al. (2015), as métricas comumente usadas em quimiometria são baseadas principalmente em PLS. Embora o PLS seja um método robusto usado normalmente para propósitos preditivos ou monitoramento de controle, apresenta uma desvantagem importante em relação à forma da equação preditiva resultante. O PLS tradicional produz apenas modelos preditivos lineares, que podem produzir desempenhos preditivos ruins na presença de uma forte relação não linear (NL) entre variáveis independentes e dependentes (LAVOIE; MUTEKI; GOSSELIN, 2019). Por outro lado, as ANNs podem modelar várias variáveis e suas relações não lineares, além de resolver problemas complexos (HAYKIN, 2001). Além disso, também encontram-se na literatura métodos que se baseiam em técnicas de otimização, como Kahmann et al. (2018). Em geral, os métodos de otimização apresentam dificuldades relacionadas à não convergência, existência de vários pontos locais ótimos da função objetivo e aumento do tempo computacional (TELES, MATEUS LEMBI; GOMES, 2010). O método proposto inova ao tomar como base conceitos advindos da química. Desse modo, desconsidera as amostras com valores intermediários na composição do índice de importância de COs. Isto torna a abordagem proposta

mais eficiente e direta, além de gerar um índice com maior poder de separação. Além disso, utiliza ANN para a geração dos subconjuntos tornando o método capaz de se adaptar bem a presença de não-linearidades nos dados.

2 MATERIAL E MÉTODOS

2.1 Bases de dados espectrais

Três bancos de dados foram utilizados a fim de avaliar o desempenho da abordagem proposta. Todos os bancos utilizados consistem em espectros NIR de combustível diesel, incluindo várias propriedades do combustível que impactam na determinação da sua qualidade. A tabela 1 apresenta uma síntese de cada banco, assim como uma breve descrição.

Tabela 1 – Bancos de dados de espectroscopia de infravermelho próximo usados neste estudo

Banco	COs	Nº de amostras	Descrição da VR	Refs	Fonte
Ponto de ebulição (ºC)	401	113	Temperatura na qual o diesel muda do estado líquido para o gasoso em toda a sua massa	(SOARES; ANZANELLO, 2018)	http://www.eigenvector.com/data/
Número de cetano	401	113	Indicador da velocidade de combustão do combustível diesel e da compressão necessária para a ignição.	(SOARES; ANZANELLO, 2018)	http://www.eigenvector.com/data/
Ponto de fulgor (ºC)	1738	85	Temperatura na qual o vapor de combustível se inflamará em forma de flash quando entrar em contato com o fogo.	(FERRÃO et al., 2011)	-

2.2 Fundamentos de Redes Neurais Artificiais (ANN)

Uma estrutura típica de um modelo ANN é composta por uma camada de entrada, que apresenta os padrões à rede (variáveis que alimentam o modelo), *n* camadas escondidas consideradas como extratoras de características, e uma camada de saída onde o resultado final é apresentado. Cada camada consiste de um número de neurônios conectados a todos os neurônios da camada seguinte e cada conexão tem um peso numérico w_{ij} atribuído, onde *i* identifica o neurônio da camada seguinte ao qual o neurônio *j* está ligado (ANTONOPOULOS et al., 2019; WU; DANDY; MAIER, 2014).

O número de camadas e a maneira como os neurônios artificiais são agrupados em cada uma destas são características que definem a arquitetura de uma ANN. Não existe uma metodologia estabelecida para a definição da arquitetura (ANTONOPOULOS et al., 2019; JAIN; NAYAK; SUDHEER, 2008). Em geral, experimentos com ANN são realizados através de repetidos testes

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

com diferentes topologias até serem obtidos resultados satisfatórios. Após estabelecer a arquitetura da rede, uma constante, denominada Bias, é inserida no modelo, sendo considerada pesos de uma entrada extra (IGEL; HÜSKEN, 2000), a fim de ajudá-lo a se adaptar melhor aos dados fornecidos.

O processamento em cada neurônio se dá pela função de ativação, definida como a transformação linear ou não linear feita ao longo do sinal de entrada. Ela basicamente decide se a informação que o neurônio está recebendo é relevante ou deve ser ignorada, determinando assim, a saída dos neurônios (JIMENEZ-MARTINEZ; ALFARO-PONCE, 2019). Neste estudo, a função de ativação escolhida foi a função sigmoide ou (logística) devido ao seu amplo uso em redes multicamadas e em redes com sinais contínuos. A inicialização dos pesos ocorre de forma aleatória e com valores diferentes de zero. Dentre os algoritmos de aprendizado existentes, o resiliente *backpropagation* (RPROP) foi o utilizado neste estudo por conta de sua rapidez computacional (PRASAD; SINGH; LAL, 2013). Este algoritmo funciona de maneira similar ao *backpropagation* tradicional, o qual se baseia na retro propagação dos erros para realizar os ajustes de pesos das camadas intermediárias. Objetiva otimizar os pesos sinápticos para que a rede neural possa aprender associações ente as entradas e as saídas (KRISHNAKUMAR, 1993).

A diferença entre esses dois algoritmos é que, no RPROP, a atualização de pesos é feita de forma independente por meio de uma taxa de aprendizado dinâmico. A taxa de aprendizado é atualizada para cada conexão de neurônio, reduzindo o erro independentemente para cada neurônio. Em outras palavras, cada peso tem seu próprio valor de atualização, que muda com o tempo (DE ALMEIDA, 2003; OLIVEIRA; BARBAR; SOARES, 2015).

3 MÉTODO PARA SELEÇÃO DOS COs MAIS INFORMATIVOS PARA PREDIÇÃO DAS PROPRIEDADES DE AMOSTRAS DE DIESEL

O método proposto compreende quatro passos principais: (*i*) pré-processamento dos dados com o ajuste de uma escala única para todas as variáveis, (*ii*) divisão do banco de dados em conjuntos de treino e teste, (*iii*) ranqueamento dos COs e (*iv*) Seleção iterativa dos COs mais relevantes. Estes passos são ilustrados na figura 1 e detalhados em seguida.

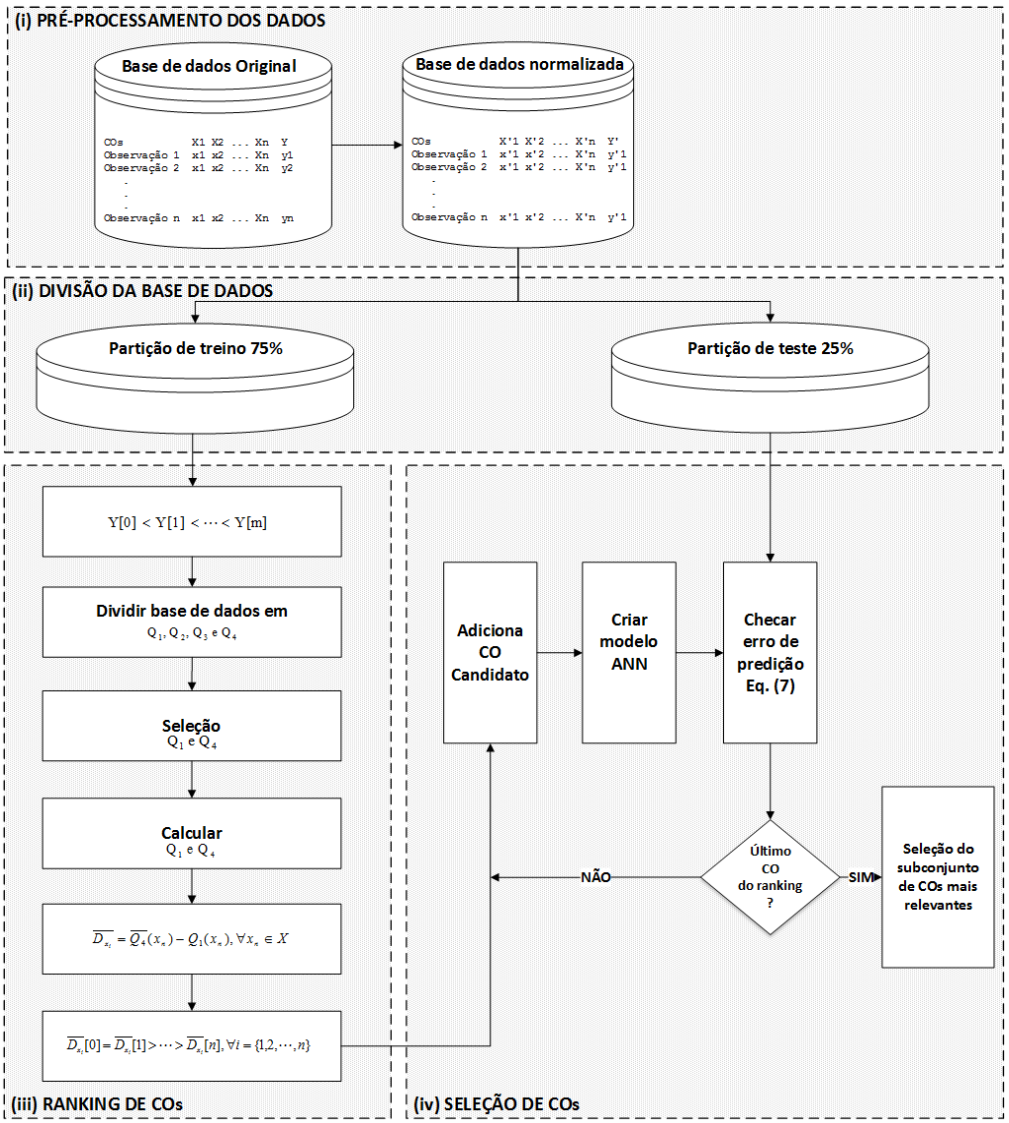


Figura 1 – Método proposto

3.1 Pré-processamento dos dados

Todos os bancos foram previamente normalizados utilizando o método do mínimo-máximo [0,1], conforme a equação (1).

$$x_{norm} = \frac{(x - x_{\min})}{(x_{max} - x_{\min})} \tag{1}$$

onde x_{norm} é o valor escalonado correspondente ao valor original x, e $x_{\min}$e x_{max} são os valores mínimo e máximo de cada CO. A normalização dos dados serve para evitar que o algoritmo interprete níveis de importância errôneos para cada variável em razão de sua magnitude.

3.2 Divisão do banco de dados em conjuntos de treino e teste

O conjunto de dados original é particionado em dois conjuntos: treino e teste, onde 75% das amostras são inseridas aleatoriamente no conjunto de treino e os 25% restantes no conjunto de teste. Tal particionamento é realizado 200 vezes, gerando subconjuntos aleatórios de treino e teste. O conjunto de treino é utilizado para ranquear os COs e selecionar os subconjuntos de COs de maior relevância, enquanto que o conjunto de teste avalia o desempenho do modelo ANN do subconjunto selecionado.

3.3 Ranqueamento dos COs de acordo com a sua relevância

De acordo com os conceitos da lei de *Lambert Beer*, infere-se que os COs que apresentam maiores variações podem auxiliar na explicação da variabilidade da variável dependente (propriedade em análise). Para identificar os COs com maior variabilidade, ordenam-se as observações (amostras) em ordem crescente em relação à variável dependente y; os dados são então divididos em quatro quartis (Q1, Q2, Q3 e Q4). Na sequência, seleciona-se o primeiro e o quarto quartil, uma vez que estes apresentam os valores extremos da variável dependente (e, de acordo com a lei de *Lambert-Beer*, as diferenças desses valores estão associadas a diferenças nas variáveis independentes – os COs que descrevem cada amostra). Calcula-se então a média de cada CO em Q1 e Q4 e, em seguida, calcula-se a diferença entre tais médias, conforme a equação (2).

$$\bar{D}_{x_n} = \bar{Q}_4(x_n) - \bar{Q}_1(x_n) \ , \forall\, x_n \in X \tag{2}$$

onde $\bar{D}_{x_i}$ é a diferença entre médias de quartis da variável x_i e $\bar{Q}_4(x_i)$ e $\bar{Q}_1(x_i)$ são as médias do quartil 4 e do quartil 1 para a mesma variável x_i. Dessa maneira, gera-se um vetor $\bar{D}$ que contém as diferenças médias entre os quartis extremos de cada CO. Os elementos do vetor são então colocados em ordem decrescente (dando origem a um índice de importância); variáveis com maiores distâncias são entendidas como mais relevantes na explicação da variabilidade da variável dependente.

3.4 Seleção dos COs mais relevantes para predição da variável dependente

A partir do *ranking* gerado anteriormente, COs são inseridos um a um em modelo ANN. O desempenho de predição do modelo ANN é avaliado no conjunto de teste por meio da métrica raiz do erro quadrático médio (RMSE), medida de erro amplamente reportada na literatura, conforme a equação (3). A cada inserção de uma nova variável no subconjunto, o RMSE é reavaliado.

$$RMSE = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(Y_i - D_i)^2} \tag{3}$$

onde N é o número de observações da amostra, Y_i se refere aos valores estimados e D_i aos valores reais.

O subconjunto que apresentar o menor erro médio (RMSE) será o selecionado. Porém, quando a diferença entre as médias dos erros de dois ou mais subconjuntos for significativamente pequena, verificam-se os valores do desvio-padrão; o subconjunto que obtiver um erro reduzido com menor variabilidade (sugerindo maior robustez) será selecionado.

4 Resultados e discussão

4.1 Resultados da seleção dos COs

Calculou-se o RMSE médio das 200 replicações, gerando uma medida de precisão preditiva mais confiável do que o resultado de uma única partição do banco de dados em treino e teste. O aumento da confiabilidade do resultado é alcançado com a amostragem aleatória pois, além de representar a diversidade de uma população avaliando os modelos gerados por cada um dos *n* conjuntos de treino e teste, permite a aplicação de procedimentos de inferência estatística, o que gera maior segurança nas análises (OLIVEIRA; GRÁCIO, 2005).

Durante o processo de adição de COs no subconjunto submetido ao modelo ANN, observou-se que os 3 bancos analisados possuem comportamentos similares em relação ao erro de predição. O erro de predição diminui significativamente quando COs de maior relevância são inseridos no modelo (isso é ilustrado no extremo esquerdo da Figura 2). Na sequência, a inserção de variáveis menos relevantes reduz a habilidade preditiva do modelo, aumentando, consequentemente, o seu erro. A fim de ilustrar tal comportamento, a Figura 2 apresenta o perfil do erro por quantidade de COs retidos para a base ponto de ebulição. Perfis semelhantes são obtidos para as demais propriedades avaliadas (perfis não apresentados por limitações de espaço).

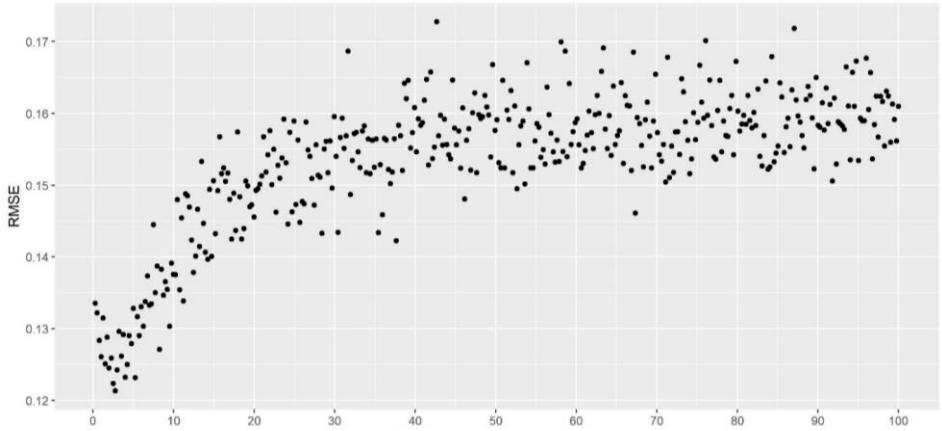


Figura 2 – Perfil do erro por quantidade de COs retidos para o Ponto de ebulição

A tabela 3 apresenta os três melhores resultados obtidos para o conjunto de teste em cada banco de dados analisado com destaque para o número e o percentual de COs retidos, RMSE médio e desvio-padrão. O melhor resultado do banco ponto de ebulição reteve cerca de 2,74% dos COs. Para o número de cetano, embora o menor RMSE médio seja obtido quando 0,5% das variáveis são retidas, o conjunto composto por cerca de 1% destas apresentou incremento de 0,68% no erro médio, porém um desvio-padrão do erro menor; por esta razão, este último foi o conjunto selecionado como o melhor resultado deste banco de dados. Já para o ponto de fulgor, reteve-se 0,69% das variáveis.

Tabela 2 – Desempenho do modelo para conjunto de teste

Descrição	Nº de COs retidos	% comprimentos de onda retidos	RMSE Médio	Sd
Ponto de ebulição	**11**	**2.7431**	**0.1214**	**0.0306**
	10	2.4938	0.1224	0.0329
	21	5.2369	0.1232	0.0409

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

Número de cetano	2	0.4988	0.1305	0.0208
	4	**0.9975**	**0.1314**	**0.0176**
	1	0.2494	0.1321	0.0190
Ponto de fulgor	**12**	**0.6904**	**0.0886**	**0.0399**
	21	1.2083	0.0912	0.0343
	15	0.8631	0.0915	0.0354

4.2 Comparação com outros métodos de seleção de COs da literatura

Os resultados da abordagem proposta foram comparados com os resultados das pesquisas de Anzanello et al. (2015) e Soares e Anzanello (2018). Esses autores utilizaram as bases de dados analisadas neste trabalho e a mesma medida para avaliação do desempenho dos modelos, RMSE.

Anzanello et al. (2015) propuseram um método que aplica *Partial Least Squares* (PLS) *regression* no espectro e quatro índices de importância derivados de parâmetros do PLS para eliminar iterativamente COs ruidosos e irrelevantes na base de dados Ponto de fulgor. Os autores apresentaram duas abordagens para a seleção do subconjunto de COs após o procedimento de eliminação iterativa das variáveis. Uma delas é o Mínimo RMSE (MR), em que o subconjunto selecionado é aquele que apresentar o menor RMSE, e a outra é a Distância Euclidiana Mínima (MED), onde um gráfico associando o RMSE calculado após a remoção de cada CO e o percentual destas variáveis retidas é construído. Cada ponto (Percentual de COs retidos; RMSE) é compreendido como uma coordenada no espaço. Além disso, os autores estabeleceram um ponto denominado ótimo, com coordenadas próximas à zero (0.01, 0.01), uma vez que, o ideal seria obter um erro próximo à zero retendo-se o menor percentual possível de variáveis. Os autores então selecionaram a coordenada que gerou a menor distância Euclidiana entre do ponto ótimo.

Já Soares e Anzanello (2018) analisam doze bancos, dentre os quais Número de cetano e Ponto de fulgor, analisados neste artigo. Os autores propõem o uso do método *Support Vector Regression – Recursive Feature Elimination* (SVR-RFE), um algoritmo para selecionar os COs mais informativos para utilizar em modelos SVR. Além disso, os autores construíram outros modelos SVR e PLS utilizando outros métodos de seleção de COs, tais como *interval* PLS (iPLS), *backward interval* PLS (biPLS), *synergy interval* PLS (siPLS) e *Successive Projection Algorithm* PLS (SPA-PLS). O estudo revelou SVR como uma robusta alternativa ao PLS, especialmente quando SVR-RFE é empregado para seleção de COs. Para as bases analisadas no presente artigo (Ponto de ebulição e Número de cetano) SVR-RFE apresentou os melhores resultados.

A tabela 4 exibe os melhores resultados encontrados em cada trabalho, assim como os melhores resultados do método proposto. O método proposto neste artigo apresentou resultados substancialmente melhores do que os métodos propostos por Anzanello et al. (2015) e Soares e Anzanello (2018). Tal superioridade pode estar associada a dois aspectos: (*i*) o índice de importância de CO aqui proposto apoia-se somente nas amostras com valores extremos da variável dependente (amostras inseridas em Q1 e Q4), dando origem a um índice com maior poder de separação ao desconsiderar as amostras intermediárias (amostras inseridas em Q2 e Q3) na composição do índice; e (*ii*) a ferramenta de predição ANN mostrou-se mais robusta do que PLS e SVM nos bancos analisados, o que pode sugerir a presença de não-linearidades nos dados (condição em que a ANN apresenta melhor desempenho quando comparada a outras técnicas de regressão).

Tabela 3 – Comparação dos resultados entre a abordagem proposta e métodos propostos por outros autores.

Base	Autor	Abordagem	Nº de COs	% COs retidos	RMSE Médio
Ponto de Ebulição	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**11**	**2.7431**	**0.1214**
	Soares e Anzanello (2018)	SVR-RFE	15	3.7406	2.6462
Número de Cetano	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**4**	**0.9975**	**0.1314**

Base	Autor	Abordagem	Nº de COs	% COs retidos	RMSE Médio
Número de Cetano	Soares e Anzanello (2018)	SVR-RFE	4	0.9975	1.7906
Ponto de Fulgor	**Abordagem proposta**	Índice de importância de COs baseado na lei de *Lambert Beer* + ANN	**12**	**0.6904**	**0.0886**
	Anzanello, Fogliatto e Ferrão (2015)	Índice de importância de COs baseado em parâmetros PLS + Regressão PLS	89	5.13	1.021

5 CONCLUSÃO

Tendo em vista que os dados gerados por técnicas de espectrometria podem reduzir o poder preditivo de modelos matemáticos em virtude de conjuntos de dados compostos por milhares de COs altamente correlacionados, este artigo propôs um método de seleção de COs com vistas à redução de dimensionalidade. Tal método integra os conceitos da lei de Lambert-Beer à técnica de classificação de redes neurais. Para a validação do método foram preditas propriedades de três conjuntos de dados NIR referentes a diesel: ponto de ebulição, número de cetano e ponto de fulgor.

A abordagem aqui proposta superou outros métodos para seleção de COs com propósitos de predição de propriedade de produtos. Os resultados obtidos neste artigo demonstram a robustez do método e sugerem que os métodos de seleção de variáveis podem ser simplificados usando uma estratégia de análise de parte de um banco de dados (somente quartis extremos), ao invés de utilizar todas as amostras na construção de um índice de importância de COs. Para pesquisas futuras recomenda-se um estudo do grau de interrelação entre as variáveis indepedentes a fim de otimizar a geração dos subconjuntos.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

6 Referências

ALEME, H. G.; BARBEIRA, P. J. S. Determination of flash point and cetane index in diesel using distillation curves and multivariate calibration. **Fuel**, v. 102, p. 129–134, 2012.

ANTONOPOULOS, V. Z. et al. Solar radiation estimation methods using ANN and empirical models. **Computers and Electronics in Agriculture**, v. 160, p. 160–167, 2019.

ANZANELLO, M. J. et al. HATR-FTIR wavenumber selection for predicting biodiesel/diesel blends flash point. **Chemometrics and Intelligent Laboratory Systems**, v. 145, p. 1–6, 2015.

CÂMARA, A. B. F. et al. MCR-ALS and PLS coupled to NIR/MIR spectroscopies for quantification and identification of adulterant in biodiesel-diesel blends. **Fuel**, v. 210, p. 497–506, 2017.

CUNHA, D. A. et al. Application of time-domain NMR as a methodology to quantify adulteration of diesel fuel with soybean oil and frying oil. **Fuel**, v. 252, p. 567–573, 2019.

DE ALMEIDA, M. B. **Um estudo comparativo de técnicas conexionistas na implementação de um sistema de reconhecimento de padrões para um nariz artificial**. Recife: Universidade Federal de Pernambuco (UFPE), 2003.

GAUCHI, J. P.; CHAGNON, P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 171–193, 2001.

GAYDOU, V.; KISTER, J.; DUPUY, N. Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil. **Chemometrics and Intelligent Laboratory Systems**, v. 106, p. 190–197, 2011.

GOBRECHT, A. et al. Combining linear polarization spectroscopy and the Representative Layer Theory to measure the Beer-Lambert law absorbance of highly scattering materials. **Analytica Chimica Acta**, v. 853, p. 486–494, 2015.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman ed. Porto Alegre: [s.n.].

IGEL, C.; HÜSKEN, M. Improving the Rprop Learning Algorithm. **Proceedings of the second international ICSC symposium on neural computation (NC 2000)**, p. 115–121, 2000.

JAIN, S. K.; NAYAK, P. C.; SUDHEER, K. P. Models for estimating evapotranspiration using artificial neural networks, and their physical interpretation. **Hydrological Processes**, v. 22, p. 2225–2234, 2008.

JIMENEZ-MARTINEZ, M.; ALFARO-PONCE, M. Fatigue damage effect approach by artificial neural network. **International Journal of Fatigue**, v. 124, p. 42–47, 2019.

KAHMANN, A. et al. Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples. **Journal of Pharmaceutical and Biomedical Analysis**, v. 152, p. 120–127, 2018.

KRISHNA, S. M. et al. Performance and emission assessment of optimally blended biodiesel-diesel-ethanol in diesel engine generator. **Applied Thermal Engineering**, v. 155, p. 525–533, 2019.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings

KRISHNAKUMAR, K. Optimization of the neural net connectivity pattern using a backpropagation algorithm. **Neurocomputing**, v. 5, p. 273–286, 1993.

LAVOIE, F. B.; MUTEKI, K.; GOSSELIN, R. A novel robust NL-PLS regression methodology. **Chemometrics and Intelligent Laboratory Systems**, v. 184, p. 71–81, 2019.

LEE, J. et al. Kernel-based calibration methods combined with multivariate feature selection to improve accuracy of near-infrared spectroscopic analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 147, p. 139–146, 2015.

LI, D. G. et al. Physico-chemical properties of ethanol-diesel blend fuel and its effect on performance and emissions of diesel engines. **Renewable Energy**, v. 30, n. 6, p. 967–976, 2005.

OLIVEIRA, E. F. T. DE; GRÁCIO, M. C. C. Análise a respeito do tamanho de amostras aleatórias simples: uma aplicação na área de Ciência da Informação. **Revista de Ciência da Informação**, v. 6, n. 3, p. 1–11, 2005.

OLIVEIRA, T. P.; BARBAR, J. S.; SOARES, A. S. Predição do tráfego de rede de computadores usando redes neurais tradicionais e de aprendizagem profunda. **Revista de Informática Teórica e Aplicada**, v. 22, n. 1, p. 10–28, 2015.

PALOU, A. et al. Calibration sets selection strategy for the construction of robust PLS models for prediction of biodiesel/diesel blends physico-chemical properties using NIR spectroscopy. **Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy**, v. 180, p. 119–126, 2017.

PRASAD, N.; SINGH, R.; LAL, S. P. **Comparison of back propagation and resilient propagation algorithm for spam classification**. Fifth International Conference on Computational Intelligence, Modelling and Simulation. **Anais**...2013

SOARES, F. et al. A non-equidistant wavenumber interval selection approach for classifying diesel/biodiesel samples. **Chemometrics and Intelligent Laboratory Systems**, v. 167, p. 171–178, 2017.

SOARES, F.; ANZANELLO, M. J. Support vector regression coupled with wavelength selection as a robust analytical method. **Chemometrics and Intelligent Laboratory Systems**, v. 172, p. 167–173, 2018.

TELES, M. L.; GOMES, H. M. Comparação de algoritmos genéticos e programação quadrática seqüencial para otimização de problemas em engenharia. **Teoria e Prática na Engenharia Civil**, v. 10, n. 15, p. 29–39, 2010.

WU, W.; DANDY, G. C.; MAIER, H. R. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. **Environmental Modelling and Software**, v. 54, p. 108–27, 2014.

YUN, Y. H. et al. An overview of variable selection methods in multivariate analysis of near-infrared spectra. **TrAC - Trends in Analytical Chemistry**, v. 113, p. 102–115, 2019.

Galoá { Este trabalho foi publicado utilizando o **Galoá** proceedings