UNIVERSIDADE FEDERAL DA PARAÍBA – CAMPUS I

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

CURSO DE GRADUAÇÃO EM QUÍMICA – BACHARELADO

Jainny Rityelle Batista dos Santos

Avaliação do modelo LASSO na determinação de metano, etano e propano em gás natural e biogás utilizando um espectrômetro de baixo custo e portátil

JOÃO PESSOA – PB

2022

Jainny Rityelle Batista dos Santos

Avaliação do modelo LASSO na determinação de metano, etano e propano em gás natural e biogás utilizando um espectrômetro de baixo custo e portátil

Trabalho de Conclusão de Curso, requisito parcial para obtenção do grau de Bacharel em Química, submetido ao Curso de Graduação em Química – Bacharelado, da Universidade Federal da Paraíba.

Orientador: **Prof. Dr. Mário César Ugulino de Araújo**

Coorientador: Dr. Sófacles Figueredo Carreiro Soares

JOÃO PESSOA – PB

2022

Catalogação na publicação
Seção de Catalogação e Classificação

S237a Santos, Jainny Rityelle Batista dos.
Avaliação do modelo LASSO na determinação de metano, etano e propano em gás natural e biogás utilizando um espectrômetro de baixo custo e portátil / Jainny Rityelle Batista Dos Santos. - João Pessoa, 2022.
43 p. : il.

Orientação: Mário César Ugulino de Araújo.
Coorientação: Sófacles Figueredo Carreiro Soares.
TCC (Graduação/Bacharelado em Química) - UFPB/CCEN.

1. LASSO. 2. Gás natural. 3. NanoNIR. I. Araújo, Mário César Ugulino de. II. Soares, Sófacles Figueredo Carreiro. III. Título.

UFPB/CCEN CDU 54(043.2)

Elaborado por Ana Roberta Sousa Mota - CRB-15/101

Jainny Rityelle Batista dos Santos

Avaliação do modelo LASSO na determinação de metano, etano e propano em gás natural e biogás utilizando um espectrômetro de baixo custo e portátil

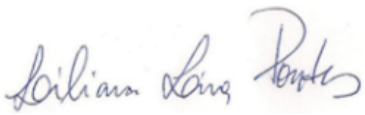
Trabalho de Conclusão de Curso, requisito parcial para obtenção do grau de Bacharel em Química, submetido ao Curso de Graduação em Química – Bacharelado, da Universidade Federal da Paraíba.

Data da avaliação: 15/12/2022

BANCA EXAMINADORA:

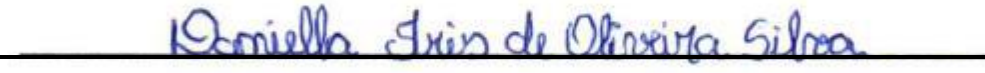
Prof. Dr. Mário César Ugulino de Araújo

(Presidente/Orientador)



Profª. Drª. Liliana de Fátima Bezerra Lira de Pontes

(Examinadora)



Ma. Daniella Iris de Oliveira Silva

(Examinadora)

Dedico àqueles que vêm me apoiando ao longo da vida facilitando a minha caminhada. Em especial, aos meus pais, Jair e Rita.

AGRADECIMENTOS

Primeiro quero agradecer a Deus por ter me concedido força e sabedoria para poder trilhar o caminho acadêmico e saber lidar com os momentos difíceis. Agradeço também por todas as dádivas e oportunidades que traçaram o meu caminho.

Dedico esse trabalho aos meu pais, Jair e Rita, por terem acreditado em mim, até em momentos que eu mesma não acreditava, me incentivando, me apoiando e dando todo carinho do mundo. Aos meus irmãos, João Rafael e Joana, por tornarem os meus dias mais leves.

Quero agradecer aos amigos que de alguma forma acrescentaram momentos especiais e únicos durante a caminhada, como Jayne, João de Jesus, Lucas Santana, Ingrid, Ângela, André, Edson, Anne, Carol, Jordan, Alice, Julyanna, entre outros que tive a oportunidade de conhecer.

Quero agradecer aos amigos que fizeram parte da minha estadia no LAQA por todo incentivo, amizade, por momentos incríveis e as melhores tardes de café, Nayara, Kelly, Samantha, Matheus, Tássio, Larissa, Ruth, Laila e João Batista. Em especial aos amigos que sempre tiveram uma palavra pra acalmar e um colo amigo nos dias difíceis, Ana Rosa, Juliana Cruz, Carla Priscila e Wallis. Um agradecimento mais que especial ao Thyago por todo apoio, companheirismo, incentivo e carinho ao longo desses anos.

Por fim, agradeço ao professor Dr. Mário César Ugulino por todo conhecimento transmitido a mim durante os anos que fiz parte do LAQA, fazendo ter acesso a muitas oportunidades. Agradeço também a Mayara e ao coorientador do trabalho, Sófacles, por compartilharem parte de seus conhecimentos, que foram de grande importância, me auxiliando no desenvolvimento e conclusão desse trabalho. Um agradecimento especial aos membros da banca, a professora Drª. Liliana Pontes e a Ma. Daniella Oliveira.

.

“Jamais considere seus estudos como uma obrigação, mas como uma oportunidade invejável para aprender a conhecer a influência libertadora da beleza do reino do espírito, para seu próprio prazer pessoal e para proveito da comunidade à qual seu futuro trabalho pertencer.”

(Albert Einstein)

RESUMO

O gás natural (GN) é um combustível fóssil, composto majoritariamente por metano, etano e propano, que é utilizado como fonte calorífica e energética em residências e indústrias. O controle de qualidade do GN é realizado para determinar os teores de metano, etano e propano, além do controle das características do gás. Para o controle de qualidade utiliza-se comumente a cromatografia a gás. No entanto está técnica possui algumas desvantagens como tempo de análise, custos com manutenção do equipamento, demanda de reagentes, entre outros. Uma alternativa para determinação desses analitos em misturas gasosas é a espectroscopia no Infravermelho próximo (NIR). Contudo, tais espectros obtidos necessitam de ferramentas quimometricas para que possam ser usadas para fins quantitativos. Diante disso, o método Least Absolute Shrinkage and Selection Operator (LASSO), que pode ser empregado para seleção de variáveis e predição de amostras futuras, foi usado para determinar metano, etano e propano em amostras de gás natural e biogás. Os resultados assim obtidos foram comparados com os obtidos pelos métodos *full spectrum* Partial Least Squares (PLS) e pelo modelo MLR com as variáveis selecionadas pelo Algoritmo das Projeções Sucessivas (SPA). Os dados foram pré-processados usando segunda derivada com suavização Savitzky-Golay ajustados a janelas de 21 pontos e polinômio de 2º ordem. Os parâmetros obtidos na predição, r_{Pred} e RMSEP, para o LASSO foram comparados com os valores obtidos para os modelos PLS e SPA-MLR. Para o metano, os valores obtidos pelo LASSO foram similares ao PLS e SPA-MLR, com RMSEP de 4,45% mol mol^{-1} e r de 0,96. Para o etano, o modelo LASSO apresentou um resultado similar ao PLS e superior ao SPA-MLR, com os valores de RMSEP de 3,98% mol mol^{-1} e r de 0,95. Considerando o propano, foi obervado que os valores do LASSO não divergiram de modo significativo do PLS e do SPA-MLR, com RMSEP de 1,46% mol mol^{-1} e r de 0,97. O teste F a um nivel de 95% de confiança mostrou que não ocorreram diferenças estatiscamente significativas entre as metodologias estudadas. Portanto, o estudo antingiu seu objetivo inicial que é demonstrar que o LASSO pode ser utiliado para predição dos analitos de modo satisfatório, da mesma forma que os modelos PLS e SPA-MLR.

Palavras chaves: LASSO; gás natural; NanoNIR.

ABSTRACT

Natural gas is a fossil fuel, composed mainly of methane, ethane and propane, which is used as a heat and energy source in homes and industries. NG quality control is carried out to determine the methane, ethane and propane contents, in addition to controlling the gas characteristics. For quality control, gas chromatography is commonly used.. However, this technique has some disadvantages such as analysis time, equipment maintenance costs, reagent demand, among others. An alternative for determining these compounds in gaseous mixtures is Near Infrared (NIR) spectroscopy. However, such spectra need chemometric tools so that they can be used for quantitative purposes. Therefore, the Least Absolute Shrinkage and Selection Operator (LASSO) method, which can be used for selection of variables and prediction of future samples, was used to determine methane, ethane and propane in natural gas and biogas samples. The results thus obtained were compared with those obtained by the full spectrum Partial Least Squares (PLS) methods and by the MLR model with the variables selected by the Successive Projection Algorithm (SPA). Data were pre-processed using second derivative with Savitzky-Golay smoothing fitted to 21 point windows and 2nd order polynomial. The parameters obtained in the prediction, rPred and RMSEP, for LASSO were compared with the values obtained for the PLS and SPA-MLR models. For methane, the values obtained by LASSO were similar to PLS and SPA-MLR, with RMSEP of 4,45% mol mol^{-1} and r of 0,96. For ethane, the LASSO model presented a similar result to PLS and superior to SPA-MLR, with RMSEP values of 3,98% mol mol^{-1} and r of 0,95. Considering propane, it was observed that LASSO values did not differ significantly from PLS and SPA-MLR, with RMSEP of 1,46% mol mol^{-1} and r of 0,97. The F test at a 95% confidence level showed that there were no statistically significant differences between the studied methodologies. Therefore, the study achieved its initial objective, which is to demonstrate that LASSO can be used to predict the analytes in a satisfactory way, in the same way as the PLS and SPA-MLR models.

Keywords: LASSO; natural gas; NanoNIR.

LISTA DE FIGURAS

Figura 1: Região do infravermelho no espectro eletromagnético **.....20**

Figura 2: Imagem de modelos de espectrômetros NIR miniaturizados. a) MicroPHAZIR (Thermo Fisher Scientific); b) Módulo de avaliação DLP NIRscan Nano (EVM; Texas Instruments); c) NeoSpectra (Sistemas Si-Ware); d) nanoFTIR NIR (Tecnologia SouthNest); e) Sensor NIRONE S (Motores Espectrais); e f) MicroNIR Pro ES 1700 (VIAVI)**.....22**

Figura 3: Gráfico dos contornos da soma de quadrados dos resíduos para a regressão ridge e o lasso, e o conjunto solução para λ com 2 covariáveis.**.....26**

Figura 4: Diagrama esquemático do sistema de análise das misturas preparadas, gás natural e biogás.**.....28**

Figura 5: Espectros NIR das misturas gasosas preparadas, gás natural e biogás do equipamento miniaturizado NIR.**.....30**

Figura 6: Espectros com as variáveis selecionadas pelo modelo LASSO, onde corresponde: A) metano, B) etano e C) propano.**.....32**

Figura 7: Espectros com as variáveis selecionadas pelo SPA, onde corresponde: A) metano, B) etano e C) propano.**.....34**

Figura 8: Gráficos do valor predito pelo valor de referência para o modelo LASSO, onde corresponde: A) metano, B) etano e C) propano.**.....37**

Figura 9: Gráficos do valor predito pelo valor de referência para o modelo SPA-MLR, onde corresponde: A) metano, B) etano e C) propano.**.....38**

LISTA DE TABELAS

Tabela 1: Especificações técnicas do gás natural.**.....18**

Tabela 2: Resultados de validação dos modelos**,** SPA-MLR, LASSO e PLS, obtidos para a determinação de metano, etano e propano. Entre parênteses estão os valores das variáveis selecionadas e latentes para cada modelo.**.....31**

Tabela 3: Resultados de predição dos modelos**,** SPA-MLR, LASSO e PLS, obtidos para a determinação de metano, etano e propano. Entre parênteses estão os valores das variáveis selecionadas e latentes para cada modelo.**.....35**

Tabela 4: Resultados do teste F a um nível de 95% de confiança obtidos na comparação entre o SPA-MLR, LASSO e PLS**.....36**

.

LISTA DE EQUAÇÕES

Equação 1– Equação de obtenção do coeficiente de regressão**.....24**

Equação 2– Equação do teste F**.....30**

LISTA DE ABREVIAÇÕES

ANP Agência Nacional do Petróleo, Gás Natural e Biocombustíveis

FID Detector de ionização em chama (do inglês, *flame ionization detector*)

FIR Infravermelho distante (do inglês, *Far Infrared*)

FT Transformada de Fourier (do inglês, *Fourier Transform*)

GC Cromatografia em fase gasosa (do inglês, *Gas Chromatography*)

GN Gás Natural

IR Infravermelho (do inglês, *Infrared*)

LASSO Least Absolute Shrinkage and Selection Operator

MIR Infravermelho Médio (do inglês, *Middle Infrared*)

MLR Regressão Linear Múltipla (do inglês, *Multiple Linear Regression*)

NIR Infravermelho Próximo (do inglês, *Near Infrared*)

$\mathbf{r_{cv}}$ Coeficientes de correlação da validação cruzada

$\mathbf{r_{pred}}$ Coeficientes de correlação da predição

RMSECV Raiz Quadrada do Erro Médio Quadrático de Validação Cruzada (do inglês, *Root Mean Square Error of Cross Validation*)

RMSEP Raiz Quadrada do Erro Médio Quadrático de Predição (do inglês, *Root Mean Square Error of Prediction*).

SUMÁRIO

1. INTRODUÇÃO **14**
2. OBJETIVOS **16**
2.1 Gerais **16**
2.2 Específicos **16**
3. FUNDAMENTAÇÃO TEÓRICA **17**
3.1 Gás Natural **17**
3.1.1 Especificações Técnicas do Gás Natural **18**
3.1.2 Biogás **19**
3.2 Espectroscopia de Infravermelho Próximo **19**
3.2.1 Equipamentos utilizados na Espectroscopia NIR **20**
3.3. Técnicas Quimiométricas **22**
3.3.1 Calibração Multivariada **23**
3.3.1.1 Regressão Linear Múltipla **23**
3.3.2 Métodos de Seleção de Variáveis **24**
3.3.2.1 Algoritmo das Projeções Sucessivas (SPA) **24**
3.3.3 LASSO **25**
4. METODOLOGIA **27**
4.1 Padrões e Amostras **27**
4.1.1 Equipamentos **28**
4.2 Análise no Espectrômetro Portátil **28**
4.3 Cromatografia Gasosa **29**
4.4 Métodos Quimiométricos **29**
5. RESULTADOS **30**
5.1 Avaliação dos Modelos de Calibração **31**
5.2 Método de Seleção de Variáveis **31**
5.3 Desempenho dos Modelos na Predição **35**
6. CONCLUSÃO **39**
REFERÊNCIAS **41**

1. INTRODUÇÃO

Atualmente é inegável a importância socioeconômica do setor energético para o mundo e é de extrema importância que a matriz energética utilizada seja diversificada e a mais limpa possível, pensando em questões ambientais e de segurança energética. Composto principalmente por hidrocarbonetos gasosos, metano, etano e propano, o gás natural (GN) é utilizado como fonte energética em indústrias, residências e meios de transportes, além de ser utilizado como matéria prima em indústrias químicas. Durante a sua queima é liberado um alto teor de energia e quando comparado com outros combustíveis fósseis libera menores teores de gases responsáveis pelo efeito estufa consequentemente auxiliando assim na diminuição de problemas associados (CAMPOS, 2017; ANP, 2022).

Por ser um combustível com uma demanda elevada de consumo é necessário que exista formas de realizar o controle de qualidade desse produto. No Brasil, o controle dos parâmetros de qualidade do gás natural, tanto nacionalmente quanto importado a ser comercializado, é realizada pela Agência Nacional de Petróleo, Gás e Biocombustíveis (ANP) através da resolução de número 16 de 17 de junho de 2008. Essa norma especifica a análise em cromatografia a gás para avaliação da composição química do GN, contudo essa técnica possui algumas desvantagens como, tempo de análise, custos com manutenção do equipamento, custos das operações de coleta, entre outros (BRASIL, 2008).

Diante disso, uma alternativa que pode ser utilizada para análise desses hidrocarbonetos é a Espectroscopia na região do Infravermelho Próximo (NIR). Além de possuir uma instrumentação simples e de baixo custo, ela não é invasiva e não necessita de pré-tratamento nas amostras. Outra vantagem importante é que ela pode ser utilizada *online* para determinar os analitos em amostras de gases com alta precisão e com resposta rápida (PASQUINI, 2018).

Contudo, as medidas no NIR são dependentes de ferramentas quimiometricas, usadas para extrair informações e possibilitar a predição de metano, etano e propano em amostras comerciais de gás natural (BARBOSA et al., 2021).

O método de Regressão Linear Múltipla (MLR) é uma das ferramentas quimiometricas usadas na construção de modelos lineares que relaciona uma ou mais

variáveis independentes com uma resposta que é dependente. Contudo, este método é altamente prejudicado quando há multicolinearidade entre as variáveis independentes e/ou número de variáveis é muito maior que a quantidade de amostras usadas para construir o modelo. Na literatura, podem ser encontrado diversas formas de corrigir estes problemas, entra elas se destacam o método de seleção de variáveis (ROY, 2016).

O Algoritmo das projeções sucessivas (SPA) tem sido utilizado com objetivo de selecionar variáveis afim de obter menores erros, modelos mais simples e fáceis de interpretar (SOARES et al.,2013).

Para demonstrar a importância do SPA-MLR, serão apresentados alguns trabalhos encontrados na literatura, como o artigo que utilizou o SPA-MLR para predição dos teores de fósforo e potássio em folhas de plantas de chá usando imagens hiperespectrais (HSI) aliado a quimiometria (WANG et al., 2020). Outro exemplo é o trabalho de Li e Shangxing que utilizou o algoritmo a fim de melhorar a precisão da previsão e a velocidade de cálculo da detecção de pH usando espectroscopia VIS-NIR (infravermelho próximo visível) (LI e SHANGXING, 2021). Já outro estudo investigou a aplicação da espectroscopia de infravermelho próximo (NIR) miniaturizada e portátil para monitoramento rápido de ácido tiobarbitúrico (TBARS) em carne de porco (KUCHA e NGADI, 2020). Por fim, outro exemplo é o trabalho que apresentou dois métodos analíticos para quantificar Óxidos de silício, alumínio e ferro (SiO_2, Al_2O_3 e Fe_2O_3) em amostras de solo usando visão computacional (COMPVIS) e espectroscopia no infravermelho médio (MIR) e como um dos métodos de calibração multivariada o SPA-MLR (MORAIS et al., 2021).

Um método que utiliza estratégias de regularização proposto por Robert Tibshirani (1996) pode ser aplicado, o Least Absolute Shrinkage and Selection Operator (LASSO), que pode ser empregado para seleção de variáveis e predição de amostras futuras. Tem por objetivo minimizar a soma dos quadrados residuais do modelo utilizando um parâmetro de ajuste, considerando que a soma dos valores absolutos dos coeficientes sejam menores que esse parâmetro, através de uma penalização L1. Com isso, alguns coeficientes de regressão são zerados e os não nulos podem ser considerados com maior relevância para construção do modelo (TIBSHIRANI,1996).

Na literatura estão expostos alguns trabalhos que relatam a eficiência do Lasso, como por exemplo: a utilização do Lasso para prever o rendimento de grãos de trigo de diferentes cultivos e linhagens (SHAFIEE et al., 2021). Em outro trabalho o método foi utilizada para identificar preditores de prevalência da doença da malária em um grupo de crianças em Gana, identificando fatores de maior relevância para o desenvolvimento da doença (AHETO et al. 2021). O Lasso também foi utilizado para realizar a predição da viscosidade do bisfenol-A sem a necessidade de correção de variações de temperatura (LUAN et al., 2020).

Nesse trabalho foi avaliado o uso do LASSO para predição de metano, etano e propano em amostras de gás natural e biogás. Os resultados foram comparados com os obtidos utilizando modelos MLR com a variáveis selecionadas pelo Algoritmo das Projeções Sucessivas e usando os espectros completos pelo método PLS (Partial Least Squares).

2. OBJETIVOS

2.1 Gerais

Avaliar a eficiência do modelo LASSO na predição de metano, etano e propano em amostra de gás natural e biogás, utilizando os espectros completos pelo método PLS, e comparar os resultados com os modelos SPA-MLR e o PLS.

2.2 Específicos

- Desenvolver os modelos de calibração com os espectros obtidos usando um nanoNIR;
- Utilizar os modelos de calibração multivariada construídos na determinação de metano, etano e propano em amostras certificadas e comerciais de gás natural;
- Avaliar os parâmetros de desempenho obtidos, tais como: os coeficientes de correlação da validação cruzada e predição, RMSECV e RMSEP;
- Comparar o modelo LASSO com o SPA – MLR e PLS a fim de saber se a modelagem obteve resultados satisfatórios na predição dos analitos.

3. FUNDAMENTAÇÃO TEÓRICA

3.1 Gás natural

O Gás Natural (GN) é um combustível fóssil, encontrado principalmente associado aos reservatórios de petróleo, mas também pode ser encontrado em reservatórios subterrâneos livres de óleo e água (ANP, 2022). O GN bruto encontrado na natureza é constituído por diversos hidrocarbonetos gasosos, no qual seu componente majoritário é o metano (CH_4), com teores acima de 70%, mas ainda pode ser encontrado outros, como o etano (C_2H_6), o propano (C_3H_8) e o butano (C_4H_{10}). Além dos hidrocarbonetos, fazem parte da sua composição, o dióxido de carbono (CO_2), o nitrogênio (N_2), sulfeto de hidrogênio (H_2S) e outros compostos sulfurados, como, as mercaptanas, além de impurezas (ANP, 2022).

Em seu estado puro, o GN, é inodoro, incolor e sem forma. É um gás que libera uma quantidade significativa de energia durante sua queima, além fornecer benefícios ambientais, quando comparado a outros combustíveis fósseis, como, por exemplo, o petróleo, pois as taxas de dióxido de carbono e óxido nitroso, liberados durante a sua combustão, são menores e as de dióxido de enxofre são insignificantes, consequentemente, auxiliando na diminuição de problemas associados ao efeito estufa (ANP, 2022; CHEN et al., 2019). Além disso, fatores como a oscilação nas fontes de energias não fósseis, como a energia solar e eólica e a insegurança com fontes de energia nucleares, favorecem os investimentos em gás natural (CHEN et al., 2019).

Mundialmente, mesmo durante a crise da pandemia, o gás natural foi responsável pelo consumo de 24% da energia primária utilizada durante o ano de 2021 e sua demanda global cresceu em 5,3% (BP, 2022). O Brasil foi o 24º consumidor de gás natural em 2021, onde os maiores consumidores foram os Estados Unidos e a Rússia (IBP, 2022). Sendo empregado como fonte na geração de calor e eletricidade em residências e indústrias, bem como matéria-prima para as indústrias de petroquímica, como por exemplo, na produção de etileno e também nas indústrias de fertilizantes, na produção de amônia (CHEN et al., 2019; IEA, 2022).

3.1.1 Especificações técnicas do gás natural

No Brasil, o controle de qualidade do gás natural, tanto nacional quanto importado a ser comercializado em território nacional, é realizado através da resolução número 16 da ANP de 17 de junho de 2008. Através dela é estabelecido o controle do poder calorífico superior (PCS), número de metano, bem como os pontos de orvalho da água (POA) e os pontos de orvalho dos hidrocarbonetos (POH). Na tabela 1 estão descritas algumas especificações do GN para comercialização no país (BRASIL, 2008).

Os componentes potencialmente corrosivos, como o dióxido de carbono, sulfeto de hidrogênio e água têm suas concentrações limitadas, como apresentado na tabela 1, devido a segurança e a integridade dos equipamentos. Apesar de não ser produzido naturalmente, o oxigênio (O_2), pode estar presente no GN, como consequência de vazamentos nas tubulações e válvulas, sendo assim necessário, o controle do mesmo, devido ao alto poder corrosivo (FARAMAWY et al., 2016).

Tabela 1: Especificações técnicas do gás natural.

CARACTERÍSTICAS	**Unidade**	**Limite**		
		Norte	Nordeste	Centro-Oeste/ Sul/Sudeste
Poder Calorífico Superior	kJ/m^3	34.000 a 38.400	35.000 a 43.000	
	kWh/m^3	9,47 a 10,67	9,72 a 11,94	
Metano, Mín.	% mol	68	85	
Etano, Máx.	% mol	12	12	
Propano, Máx.	% mol	3	6	
Butano E Mais Pesados, Máx.	% mol	1,5	3,0	
Oxigênio, Máx.	% mol	0,8	0,5	
Inertes (N_2+Co_2), Máx.	% mol	18,0	8,0	6,0
Co_2, Máx.	% mol		3,0	
Enxofre Total, Máx.	mg/m^3		70	
Gás Sulfídricos (H_2s), Máx.	mg/m^3	10	13	10
Ponto De Orvalho De Água A 1 Atm, Máx.	°C	-39	-39	-45
Ponto De Orvalho De Hidrocarbonetos A 4,5 Mpa, Máx.	°C	15	15	0

Fonte: Adaptado de ANP, 2008.

Outra característica a ser observado é que o gás natural não possua partículas líquidas ou sólidas visíveis. Em relação à presença dos gases inertes, CO_2 e N_2, suas taxas também são controladas, pois são considerados gases diluentes, ou seja, diminuem a concentração de metano presente no gás, diminuindo consequentemente o seu poder calorífico (BRASIL, 2008; FARAMAWY et al., 2016).

3.1.2 Biogás

O biogás é um biocombustível que pode ser produzido a partir da digestão anaeróbica de resíduos, sendo um recurso renovável que desempenha papel importante para atenuar problemas ambientais. Tipicamente o biogás é composto por: metano, que é o componente majoritário, variando de 35 a 65%, dióxido de carbono (15 a 50%), água (0 a 5%), além de hidrogênio, nitrogênio e oxigênio (RAFIEE,et al, 2021).

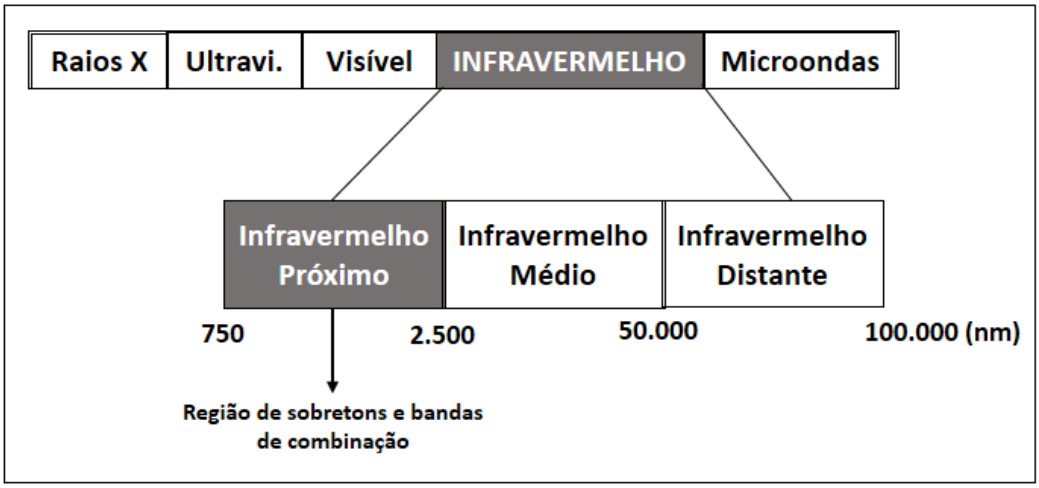
Porém, esse gás só pode ser comercializado após a etapa de purificação, reduzindo, assim, o biogás ao biometano. No Brasil a regulamentação do biometano é realizada através da resolução da ANP nº 685/2017, que estabelece regras para o controle da qualidade e a especificação do biometano oriundo de aterros sanitários e de estações de tratamento de esgoto, e da resolução nº 8/2015 que estabelece regras para controle da qualidade e a especificação do biometano oriundo de produtos e resíduos orgânicos agrossilvopastoris e comerciais (ANP(BIOMETANO), 2022).

3.2 Espectroscopia de infravermelho próximo

A espectroscopia de infravermelho (IR) é fundamentada nas vibrações moleculares. Quando ocorre a passagem de uma certa radiação através da amostra, uma determinada energia é absorvida, gerando como resposta os espectros, a intensidade da absorção vai depender da variação do momento dipolo da molécula e da anarmonicidade da ligação (PASQUINI, 2003; MANLEY e BAETEN, 2018; PASQUINI, 2018).

A região do infravermelho corresponde aos espectros que compreendem os comprimentos de onda de 780 a 100.000 nm. Essa região é dividida em três sub-regiões o infravermelho próximo (NIR), que corresponde a região de 780 – 2500 nm, o infravermelho médio (MIR), de 2500 – 50.000 nm, e o infravermelho distante (FIR), de 50.000 – 100.000 nm (MANLEY e BAETEN, 2018).

Figura 1: Região do infravermelho no espectro eletromagnético.



Fonte: Adaptado de Volmer, 2001.

A região NIR, diferente das outras regiões, possui bandas alargadas, originadas principalmente de sobretons e combinações de modos vibracionais fundamentais. Essas bandas estão associadas aos modos vibracionais de grupos funcionais composto principalmente por átomos relativamente pesados como carbono (C), nitrogênio (N), enxofre (S) e oxigênio (O) ligados a átomos de hidrogênio (H), mas ligações entre átomos mais pesados, como C=O, também pode apresentar espectros, só que de forma menos intensa (PASQUINI, 2003; MANLEY e BAETEN, 2018).

Através do comprimento de onda e da intensidade das absorções, podem ser obtidos as informações qualitativas e quantitativas da amostra. O comprimento de onda de uma banda de absorção vai estar relacionada com a massa dos átomos ligados e envolvidos com cada modo de vibração, e com o a força de ligação. Qualquer alteração detectada em uma dessas duas grandezas alteram o comprimento de onda de absorção (PASQUINI, 2018).

3.2.1 Equipamentos utilizados na espectroscopia NIR

Os equipamentos utilizados na espectroscopia NIR podem ser classificados de acordo com a tecnologia utilizada para a seleção de comprimento de onda, podendo ser um equipamento de bancada ou portátil (PASQUINI, 2003).

Os espectrofotômetros baseados na transformada de Fourier (FT), são considerados um dos instrumentos que obtêm os melhores resultados em termos de

precisão, relação sinal-ruído e alta velocidade de varredura. Por não utilizar fendas de entrada e saída para limitar a intensidade de radiação que chega ao detector, esse instrumento possui um alto rendimento de radiação. Comparado com outras opções é considerado um equipamento de alto custo (PASQUINI, 2003).

Nos equipamentos baseados em filtros óticos acústicos ajustáveis (AOTF), a seleção do sinal é obtido através da radiofrequência que alteram o índice de refração de um cristal, comumente utilizado o TeO_2. Possui velocidade de varredura rápida e até 2.000 comprimentos de onda podem ser selecionados por segundo (PASQUINI, 2003).

Um outro tipo de equipamento, é os dispersivos que são baseados em redes de difração, foi através desse tipo de instrumentação que foi possível consolidar a espectroscopia NIR como uma ferramenta analítica. É utilizada em amostras cuja a medição é realizada por transmitância, muito utilizada no controle da produção de açúcar e álcool. Esse tipo de instrumentação possui algumas desvantagens, como, possuir velocidade de varredura lenta e a falta de precisão do comprimento de onda, dificultando a manutenção do modelos multivariados (PASQUINI, 2003).

As ferramentas dispersivas podem ser melhoradas na relação sinal-ruído através da utilização da transformada de Hadamard, que pode aumenta a velocidade de varredura, cerca de 10 segundos por scan, e simplificar a obtenção dos espectros (PASQUINI, 2003).

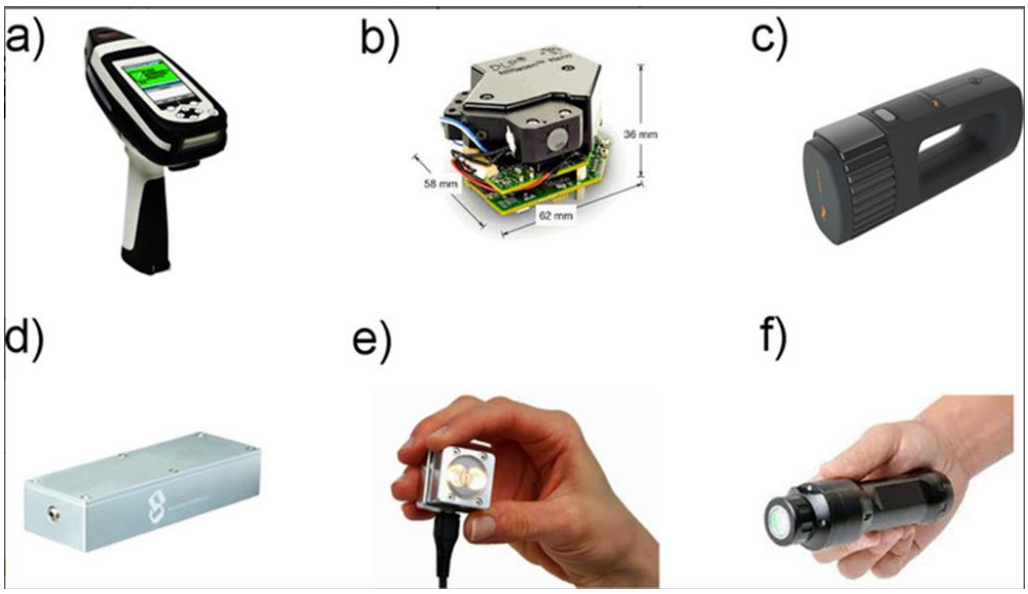
Com o desenvolvimento e avanço de novas tecnologias os equipamentos miniaturizados estão ganhando um amplo espaço nos últimos anos, devido ao seu baixo custo e diminuição do tamanho do equipamento, além de que podem ser utilizados como instrumentação portátil para análise em campo (PASQUINI, 2003; PU et al., 2021; BEĆ et al., 2021).

Devido a vantagem da ampla aplicabilidade desse equipamento quando comparado a espectroscopia convencional, que se limita ao uso laboratorial, o número de trabalhos publicados utilizando esse tipo de equipamento está crescendo rapidamente ultimamente. Sendo uma área ativa atualmente na química analítica (PU et al., 2021; BEĆ et al., 2021).

Alguns trabalhos na literatura demonstram o bom desempenho dos equipamentos miniaturizados, mostrando a eficácia e a confiabilidade dos resultados obtidos por esse tipo de equipamento. Alguns trabalhos podem ser citados, como o de Pu et al (2021) que realizaram uma revisão que se concentra nos desenvolvimentos recentes de dispositivos NIR portáteis e nas suas aplicações no campo de laticínios. Foi utilizado um equipamento miniaturizado para prever o frescor dos ovos, classificando como velhos ou frescos (CRUZ-TIRADO et al., 2021). Por fim, um outro exemplo, que foi utilizado um espectrômetro micro-infravermelho próximo (micro-NIR) conectado a um smartphone para avaliar os atributos do sabor do chá preto (WANG et al., 2021).

Na **figura 2** é apresentada alguns tipos de espectrômetros NIR miniaturizado populares nos artigos presentes na literatura. Destacando o módulo de avaliação DLP NIRscan **(2(b))**, utilizado como equipamento no presente trabalho.

Figura 2: Imagem de modelos de espectrômetros NIR miniaturizados. a) MicroPHAZIR (Thermo Fisher Scientific); b) Módulo de avaliação DLP NIRscan Nano (EVM; Texas Instruments); c) NeoSpectra (Sistemas Si-Ware); d) nanoFTIR NIR (Tecnologia SouthNest); e) Sensor NIRONE S (Motores Espectrais); e f) MicroNIR Pro ES 1700 (VIAVI).



Fonte: BEĆ.et al, 2021.

3.3. Técnicas quimiométricas

A quimiometria é a utilização de ferramentas matemáticas e estatísticas para extrair informações a respeito dos dados analíticos, no caso, dos espectros obtidos. Foi necessário no decorrer de alguns anos desde o descobrimento da técnica, para

poder utilizar a espectroscopia NIR nas primeiras aplicações analíticas. Isso ocorreu devido à complexidade das bandas de absorção obtidas**.** Com o avanço de recursos matemáticos e estatísticos os problemas relacionados com a falta de seletividade da espectroscopia NIR puderam ser minimizados (PASQUINI, 2003; MANLEY e BAETEN, 2018)**.**

Portanto, a quimiometria é importante para extrair informações relevantes dos espectros, visto que além de informações físicas e químicas, também pode conter ruídos, variabilidade, interações entre outras informações que podem prejudicar na interpretação dos espectros.

3.3.1 Calibração Multivariada

O processo de calibração estabelece uma relação entre as medidas instrumentais e os valores correspondentes às propriedades de interesse, através de uma série de operações estatística. Dentre os tipos de calibração, os métodos univariados são os que possuem maior facilidade na aplicação e validação, devido a utilização de uma única variável para cada amostra, porém é necessário que a medida esteja livre de interferentes que possam alterar a relação entre a linearidade da grandeza medida e a propriedade estudada (BRAGA e POPPI, 2004; GOODARZI et al., 2015).

No caso de calibração multivariada é possível relacionar à medida, duas ou mais respostas instrumentais, além de ser possível realizar as análises mesmo na presença de interferentes. Por isso é um alternativa viável quando não é possível a utilização de métodos univariados. Os modelos de calibração multivariada são a Regressão Linear Múltipla (MLR), Regressão Pelos Mínimos Quadrados Parciais (PLS) e Regressão por Componentes Principais (PCR) (BRAGA e POPPI, 2004).

3.3.1.1 Regressão Linear Múltipla

A Regressão Linear Múltipla (MLR) é um método estatístico que é baseado na modelagem entre duas ou mais variáveis interpretativas, que são independentes, e uma variável resposta, que é dependente. É um método mais direto de estimar os coeficientes de regressão. Os valores do coeficiente são obtidos como demonstrado na **Equação 1,** onde $\boldsymbol{X}$ é a variável instrumental, $\boldsymbol{y}$ é a variável química ou física e $\boldsymbol{b}$ o coeficiente de regressão (KUMAR, et al., 2014).

$$\boldsymbol{b} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \quad \textbf{1}$$

Mas o método precisa cumprir alguns requisitos matemáticos como: o número de amostras tem que ser maior ou igual ao número de preditores e as variáveis devem apresentar independência linear entre si (KUMAR, et al., 2014; ROY, 2016).

3.3.2 Métodos de seleção de variáveis

A espectroscopia possui um número de variáveis espectrais elevado, esse fato pode prejudicar o modelo estatístico e o cálculo de predição, tornando a predição pouco confiável. Portanto, antes da utilização do modelo, é encontrado o número mínimo de variáveis necessárias para explicar as propriedades da amostra (CHEN et al., 2017).

Com a utilização da seleção de variáveis é possível simplificar o modelo e melhorar a precisão da predição sem prejudicar a exatidão (CHEN et al., 2017).

3.3.2.1 Algoritmo das projeções sucessivas (SPA)

O Algoritmo das Projeções sucessivas (SPA) tem por objetivo selecionar variáveis para serem aplicadas no modelo de Regressão Linear Múltipla (MLR), afim de obter um modelo estatisticamente estável (SOARES et al., 2013).

O SPA é dividido em três etapas. Na primeira etapa, é onde se inicia a geração das cadeias de variáveis a partir do critério de baixa colinearidade entre si dos conjuntos de variáveis. Esses conjuntos de dados são projetados por operações aplicadas nas colunas da matriz **X** da amostra selecionadas para a calibração. Nesta etapa o SPA tem semelhanças com os demais algoritmos clássicos, esse tipo de seleção de variáveis não tem por objetivo apenas a maximização do determinante $\boldsymbol{X^T X}$**,** mas também visando a relação entre as variáveis x e as propriedade y de interesse, que é analisado na segunda etapa (SOARES et al., 2013).

É na segunda etapa que serão analisadas os subconjuntos de variáveis retiradas das cadeias criadas na primeira etapa. Esses subconjuntos são obtidos tomando as primeiras variáveis L de cada cadeia, é através do Erro Médio Quadrático

para o Conjunto de Validação (RMSEV), que irá estabelecer o subconjunto mais adequado, quanto menor o erro, melhor é o subconjunto (SOARES et al., 2013).

Na terceira e última etapa, onde as variáveis que não contém informações são eliminadas, melhorando o modelo. É criado um índice de relevância para cada variável, que faça parte do subconjunto selecionado no final da segunda etapa. O valor do índice é encontrado multiplicando o desvio padrão da variável pelo valor absoluto do seu coeficiente de regressão. As variáveis então são classificadas e finalmente obtido o menor número de variáveis, de forma que o RMSEV não seja significativamente maior que o menor valor observado de acordo com um teste F, com nível de significância de 25% (SOARES et al., 2013).

3.3.3 LASSO

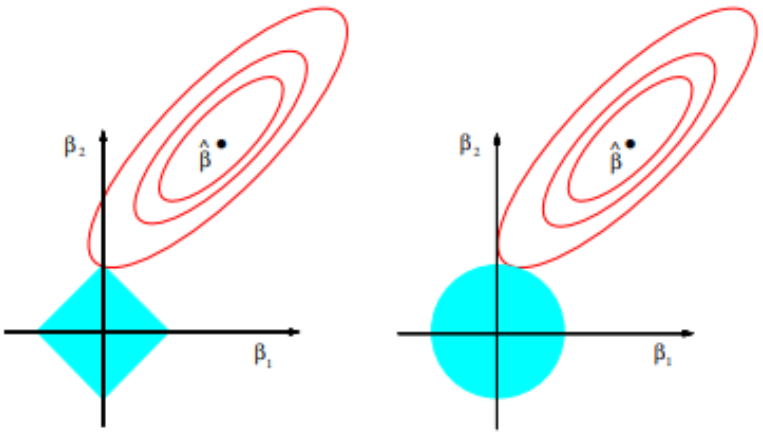
Considerando uma situação utilizando o método de regressão usual, onde temos os dados (x',y), i = 1,2,3,...,n, no qual x'=(x1,x2,...,xp) são as variáveis regressoras ou independentes e y é a variável resposta ou dependente da i-ésima observação. Ao utilizar regressão por mínimos quadrados ordinários é possível minimizar a soma dos quadrados dos resíduos do modelo. Porém há dois fatores que podem reduzir a precisão da predição desse método, como: alta dimensionalidade (p>>n) e covariáveis muito correlacionadas (multicolinariedade) (TIBSHIRANI, 1996).

Diante disso, um método proposto por Robert Tibshirani (1996) pode ser aplicado para contornar esses fatores, o LASSO. Isso ocorre pois o método tem por objetivo minimizar a soma dos quadrados residuais do modelo utilizando um parâmetro de ajuste, considerando que a soma dos valores absolutos dos coeficientes sejam menores que esse parâmetro, através de uma penalização L1. Com isso, alguns coeficientes são zerados e os não nulos podem ser considerados com maior relevância para construção do modelo (TIBSHIRANI, 1996).

Para demonstrar como os parâmetros são nulos na penalização L1, comparando graficamente o que acontece com os coeficientes utilizando as penalizações L1 e L2. Na **figura 3** é apresentado dois gráficos que representam o conjunto de solução para uma regressão com duas covariáveis, aplicando o LASSO e a regressão RIDGE. Analisando a região de L1, é possível observar uma forma quadrada enquanto para L2 é uma forma arredondada. As elipses apresentadas

correspondem aos contornos que apresentam mesmo valor da soma de quadrados dos resíduos. A solução é atingida quando o contorno da elipse encontra o primeiro ponto do conjunto de solução. A penalização L1 possui arestas em sua solução, fazendo com que alguns dos seus coeficientes sejam zerados quando a solução é atingida, o que não ocorre com a penalização L2, já que não possui essas arestas, dificilmente seu conjunto de solução atingirá coeficientes nulos. Por isso, o Lasso possui a propriedade de criar esparsividade dentro do modelo, podendo ser encontrado vários coeficientes nulos em um conjunto de solução (HASTIE et al.,2015; ALCÂNTARA JUNIOR, 2021; TIBSHIRANI 2011).

Figura 3: Gráfico dos contornos da soma de quadrados dos resíduos para a regressão ridge e o lasso, e o conjunto solução para λ com 2 covariáveis.



Fonte: HASTIE et al, 2015.

Na regressão LASSO, a intensidade da penalização é controlada pelo parâmetro de ajuste λ. Por tanto é necessário que esse parâmetro seja estimado corretamente, sendo o método de validação cruzada mais utilizado para estimá-lo (HASTIE et al., 2015; ALCÂNTARA JUNIOR, 2021).

Inicialmente o método divide o conjunto de dados em grupos $k > 1$, de forma aleatória. No qual um desses grupos é fixado como conjunto de teste, para validar o desempenho do modelo, e os grupos $k > 1$ restantes são utilizado como conjunto de treinamento, de modo que esse procedimento se repita k vezes até que todos os

grupos sejam utilizados como grupo de teste uma vez (HASTIE et al., 2015; ALCÂNTARA JUNIOR, 2021).

Posteriormente, o modelo é construído usando o conjunto de treinamento para diferentes valores de λ. Com isso o conjunto de teste é usado no modelo ajustado para previsão e obtenção dos erros quadráticos médios para cada valor de λ (HASTIE et al., 2015; ALCÂNTARA JUNIOR, 2021).

O processo é repetido para todos os grupos do conjunto de dados, com isso para cada valor de λ dentro do intervalo teremos *n* estimativas do erro quadrático médio. Então é calculada a média do erro para cada valor de λ, a fim de escolher o que apresente a menor média dos erros quadráticos médio, no qual será o valor que maximiza a performance do modelo. (HASTIE et al., 2015; ALCÂNTARA JUNIOR, 2021).

4. METODOLOGIA

4.1 Padrões e amostras

Foram utilizados padrões (% mol mol^{-1}) de metano (99,9%), etano (99,0%), propano (99,5%), nitrogênio (99,9%) e três misturas desses gases, com composição certificada, onde seus teores e composição eram compatíveis com os que são encontrados em amostras de gás natural, obtidos da empresa Linde Gases Ltda.

Os padrões obtidos foram utilizados para preparação de 86 misturas gasosas seguindo um planejamento experimental Brereton. As concentrações de gás variam de 47,5 a 100 % mol mol^{-1} para o metano, de 0 a 50 % mol mol^{-1} para etano e 0 a 31,5 % mol mol^{-1} para o propano. Considerando a faixa de concentração dos analitos presentes nas amostras de gás natural e do biogás.

Para etapa de predição dos modelos de calibração foram utilizadas 4 amostras comerciais de gás natural obtidos em postos de combustíveis no município de João Pessoa e 1 amostra de biogás coletada em um biodigestor localizado no Shopping de Camaragibe no munício de Camaragibe (Pernambuco).

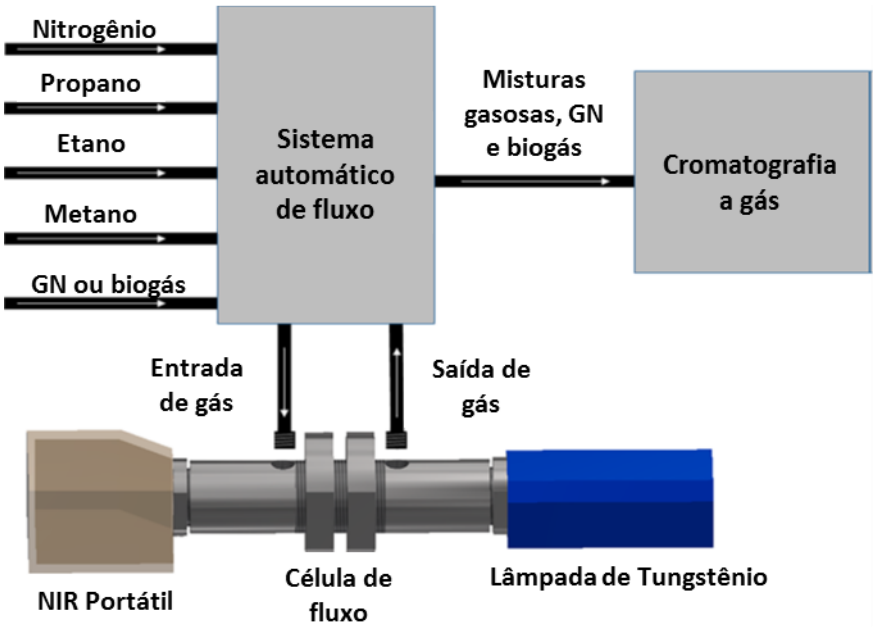
4.1.1 Equipamentos

Na preparação das misturas gasosas foi utilizado um sistema automático de preparação de misturas proposto por Dantas (2015).

O registro dos espectros foi realizado empregando um espectrofotômetro portátil, nanoNIR (Texas Instruments), com faixa de comprimento de onda de 900-1700 nm e resolução digital de 228 pontos ao longo dessa faixa e como fonte de radiação utilizou-se uma lâmpada de Tungstênio da Ocean Optics. Na **figura 4** está disposto o sistema de análise dos gases (BARBOSA et al., 2021).

As análises de referência foram realizadas em um cromatógrafo a gás modelo GC-2014 (Shimadzu), seguindo a ASTM D 1945 na resolução N°16 da ANP.

Figura 4: Diagrama esquemático do sistema de análise das misturas preparadas, gás natural e biogás.



Fonte: Adaptado de Barbosa et al, 2021.

4.2 Análise no espectrômetro portátil

Inicialmente é realizada a etapa de limpeza do sistema, posteriormente foram adicionados ao sistema os componentes gasosos da mistura. As concentrações molares dos gases foram convertidas em pressões parciais e foi adicionado nitrogênio até o sistema atingir uma pressão de até 1,550 bar (BARBOSA et al., 2021).

Em seguida foi realizada a homogeneização da mistura utilizando uma bomba de diafragma. Os espectros foram registrados em 5 segundos, no espectrofotômetro e em seguida realizada a análise cromatográfica (BARBOSA et al., 2021).

As mesmas etapas foram aplicadas para as amostras certificadas, comerciais e de biogás. Todas as pressões foram controladas utilizando um manômetro digital com precisão de ± 0.001 bar do sistema automático e com controle da temperatura ambiente em 23 ± 1 °C (BARBOSA et al., 2021).

4.3 Cromatografia gasosa

As análises de referência foram realizadas utilizando um cromatógrafo gasoso que interliga-se ao sistema automático por meio de um injetor automático de amostras. As injeções foram realizadas no modo Split (1:100), a uma temperatura de 240°C e utilizando uma válvula de amostragem com um volume de 25 microlitros. A pressão do gás hélio, empregado como gás de arraste, foi de 58,3 KPa com uma vazão de 1,42 mL min^{-1}, velocidade linear foi de 27,3 cm s^{-1} e fluxo de purga de 3,0 mL min^{-1}. As análises foram realizadas no modo isotérmico com a temperatura da coluna em 90°C. Foi empregado uma coluna capilar GS-GASPRO de 30 m com diâmetro interno de 0,32 mm e um detector de ionização em chama (FID) com temperatura regulada em 250°C. O tempo da corrida cromatográfica foi em média de 10 minutos (BARBOSA et al., 2021).

4.4 Métodos quimiométricos

O conjunto de dados em estudos possuem 86 misturas gasosas padrão que foram particionadas em conjunto de calibração, 60 misturas, e conjunto de predição, 26 misturas, utilizando o algoritmo SPXY. Sendo adicionado ao conjunto de calibração as 3 amostras certificadas para incluir a variabilidade dos outros componentes não modelados e que fazem parte da composição da amostra e ao conjunto de predição foram adicionadas 4 amostras de gás natural e 1 amostra de biogás (BARBOSA et al., 2021).

Em Barbosa et al.(2021), pode ser visto que o pré-processamento aplicado que proporcionou melhor desempenho foi o da segunda derivada com suavização Savitzky-Golay com janela de 21 pontos e polinômio de segunda ordem. Desta forma, o mesmo foi aplicado para o presente estudo.

Os modelos LASSO e SPA-MLR foram executados na interface do Matlab R2015 e Matlab R2010 (Mathworks, EUA). Os desempenhos dos modelos construídos foram avaliados em termos de raiz do erro quadrático médio de predição e coeficientes de correlação de previsão (RMSEP e rPred, respectivamente).

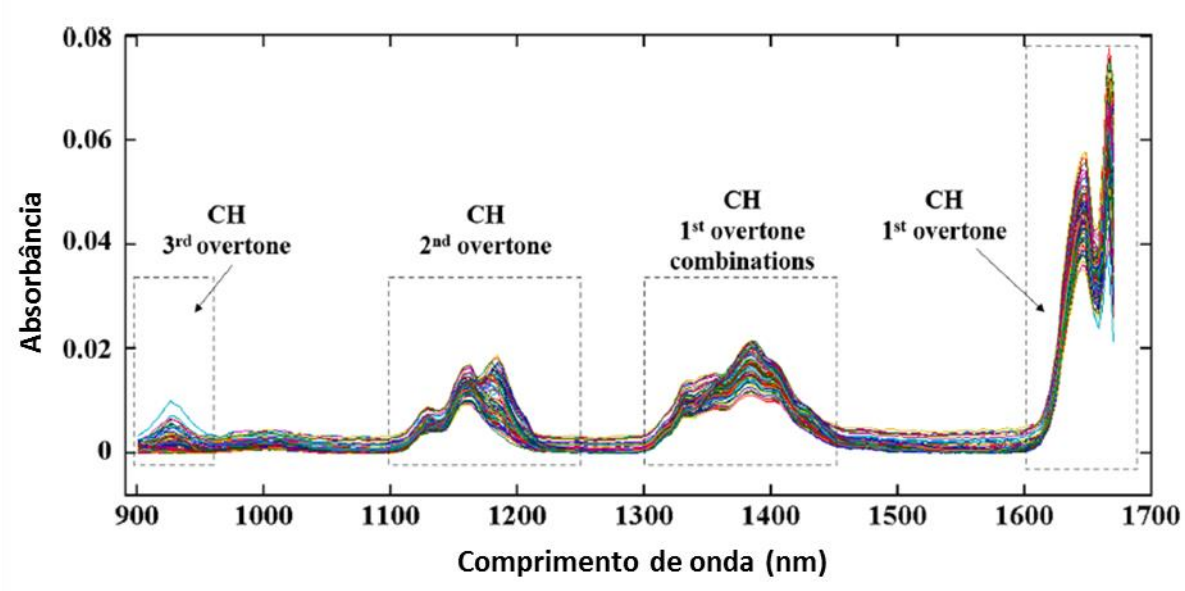
Um teste F, a um nível de 95% de confiança, foi realizado para avaliar se houve diferenças estatisticamente significativas entre as métricas estudadas.

$$\mathbf{F} = (RMSEP_1)^2/RMSEP_2)^2 \qquad \mathbf{2}$$

5. RESULTADOS

Na **figura 5** são apresentados os espectros obtidos no equipamento NIR miniaturizado.

Figura 5: Espectros NIR das misturas gasosas preparadas, gás natural e biogás do equipamento miniaturizado NIR.



Fonte: Adaptado de Barbosa et al, 2021.

Nos espectros é possivel observar a existência das bandas de combinações e de sobretom de CH. Onde estão situadas entre 850 – 950 nm, o terceiro sobretom de CH, entre 1120 – 1220 nm o segundo sobretom de CH e 1300 – 1450 nm, o primeiro sobretom de combinações vibracionais dos estiramentos e das deformações referentes às ligações CH + CH e das CH + CC, ambas com a mesma intensidade. Por fim, entre 1650 –1800 nm, relacionada ao primeiro sobretom de CH (considerando CH, CH_2 e CH_3) com maior intensidade.

A região de trabalho foi limitada entre os comprimentos de onda de 900 e 1670 nm, visto que acima de 1670 nm o detector apresentou saturação. Com relação as pequenas variações da linha de base, elas podem estar relacionadas as mudanças da intensidade da fonte de luz, temperatura e umidade. Por tanto, para contornar esses problemas foram aplicadas diferentes pré-processamentos aos espectros. Sendo a segunda derivada com suavização Savitzky-Golay usando janelas de 21 pontos e polinômio de 2º ordem escolhida como o melhor pré-processamento.

5.1 Avaliação dos modelos de calibração

Na **tabela 2** são apresentados os valores dos coeficientes de correlação (r_{cv}) e da raiz do erro quadrático médio de validação cruzada (RMSECV).

Tabela 2: Resultados de validação dos modelos**,** SPA-MLR, LASSO e PLS, obtidos para a determinação de metano, etano e propano. Entre parênteses estão os números das variáveis selecionadas e latentes para cada modelo.

	Parâmetros	**SPA-MLR**	**LASSO**	**PLS**
Metano (47,5-100%mol mol^{-1})	r_{cv}	0,95	0,95	0,95
	RMSECV	4,55 (3)	5,02 (5)	4,61 (3)
Etano (0-50.0% mol mol^{-1})	r_{cv}	0,97	0,95	0,96
	RMSECV	3,18 (6)	4,39 (7)	3,67 (4)
Propano (0-31,5% mol mol^{-1})	r_{cv}	0,97	0,95	0,96
	RMSECV	1,96 (5)	2,57 (8)	2,20 (4)

Fonte: Próprio Autor, 2022.

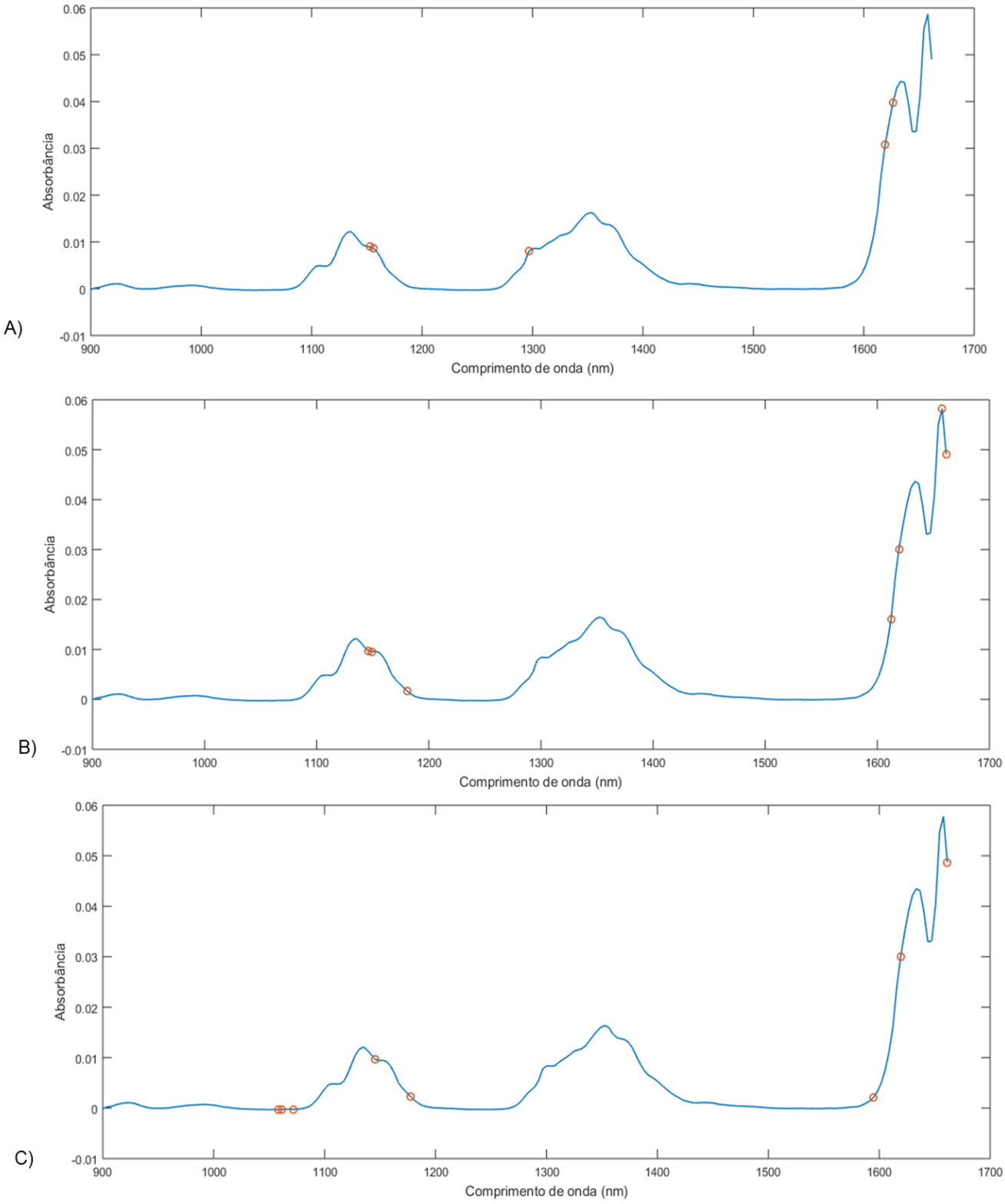
No modelo SPA-MLR foram selecionadas 3 variáveis para o metano, 6 para o etano e 5 para o propano. No modelo LASSO foram selecionados 5, 7 e 8 variáveis para o metano, etano e propano, respectivamente. Já o modelo PLS utilizou 3 variáveis latentes para o metano, 4 para o etano e 4 para o propano.

Ao analisar os valores de coeficiente de correlação dos modelos, todos apresentaram resultados acima de 0,95. Os detalhes de cada modelo serão analisados nos tópicos a seguir.

5.2 Método de seleção de variáveis

Na **figura 6** são apresentadas a variáveis selecionadas pelo modelo LASSO para os analitos metano, etano e propano, respectivamente.

Figura 6: Espectros com as variáveis selecionadas pelo modelo LASSO, onde corresponde: A) metano, B) etano e C) propano.



Fonte: Próprio autor, 2022.

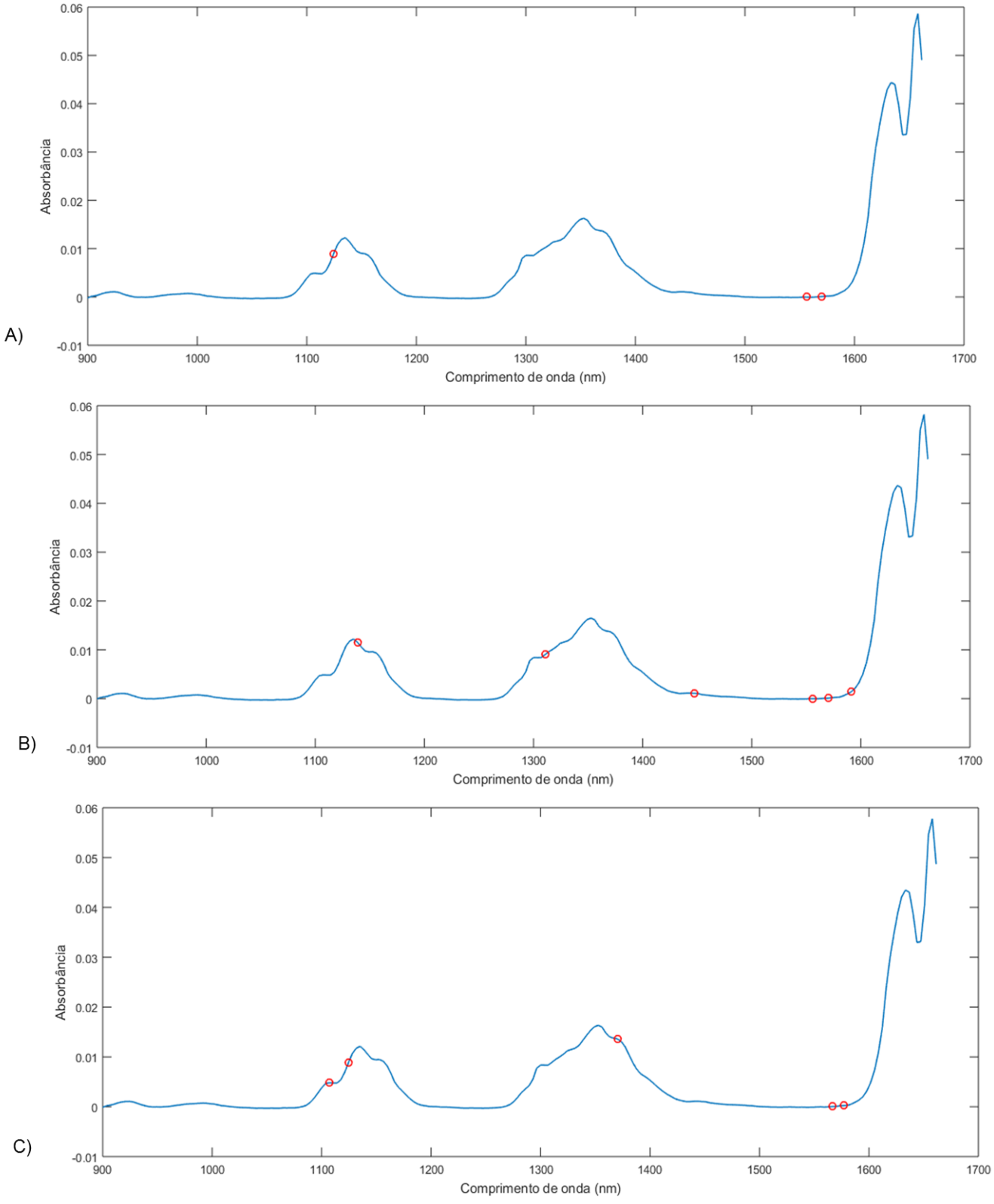
No caso do metano, foram selecionadas pelo LASSO cinco variáveis. Nas quais duas estão situadas no segundo sobretom de CH que corresponde aos comprimentos de onda de 1152 e 1157 nm, uma no início do primeiro sobretom de combinação de CH (1300 nm) e as duas restantes (1617 e 1624 nm) no início do primeiro sobretom de CH. Podendo ser concluído que todas as variáveis selecionadas pelo modelo para o metano estão localizadas em regiões expressivas do espectro.

Para o etano, foram selecionadas sete variáveis, onde todas estavam presentes em regiões de absorção significativas. As variáveis em 1146 e 1150 nm estão localizadas na região do segundo sobretom de CH e no final da mesma região é possível observar uma variável em 1181 nm. As variáveis em 1612, 1618, 1656 e 1660 nm estão situadas no primeiro sobretom de CH.

Já para o propano, o LASSO selecionou oito variáveis, nas quais quatro estão localizadas em regiões informativas. Na região de segundo sobretom de CH encontram-se as variáveis selecionadas em 1145 e 1178 nm. Na região do primeiro sobretom de CH são observadas duas variáveis selecionadas em 1618 e 1659 nm. As variáveis selecionadas em 1059, 1062, 1072 e 1595 nm estão localizadas em regiões não expressivas do espectro.

Na **figura 7** estão apresentadas os espectros com as variáveis selecionadas pelo modelo SPA-MLR para o metano, o etano e o propano.

Figura 7: Espectros com as variáveis selecionadas pelo SPA, onde corresponde: A) metano, B) etano e C) propano.



Fonte: Próprio autor, 2022.

Ao analisar o espectro do metano é possível observar que foram selecionadas pelo SPA três variáveis. No início da região do segundo sobretom de CH foi selecionada uma variável em 1124 nm e considerada a única relevante no espectro, visto que as outras duas, em 1556 e 1569 nm, estão situadas fora da região informativa.

Para o etano, é observado uma variável (1158 nm) na região do segundo sobretom de CH. Na região do primeiro sobretom de combinações de CH foram selecionadas duas variáveis (1311 e 1447 nm). Uma variável foi selecionada no início da região do primeiro sobretom de CH (1592 nm). As demais variáveis (1555 e 1570 nm) que o SPA selecionou não estão em regiões expressivas do espectro.

Enquanto para o propano foram selecionadas cinco variáveis. As duas primeiras, 1107 e 1123 nm, no começo da região do segundo sobretom de CH e uma variável, 1370 nm, na região do primeiro sobretom de combinações de CH. As outras duas, 1566 e 1577 nm, estão situadas em regiões sem caráter espectral relevante.

Ambos os modelos, LASSO e SPA, selecionaram variáveis relevantes. Sendo o LASSO, o algoritmo que mais selecionou variáveis em regiões com maior relação sinal/ruído para o metano, etano e propano.

5.3 Desempenho dos modelos na predição

Os desempenhos dos modelos LASSO, SPA-MLR e PLS na etapa de predição estão expostos na **tabela 3**.

Tabela 3: Resultados de predição dos modelos**,** SPA-MLR, LASSO e PLS, obtidos na determinação de metano, etano e propano. Entre parênteses estão os valores das variáveis selecionadas e latentes para cada modelo.

	Parâmetros	**SPA-MLR**	**LASSO**	**PLS**
Metano (47,5-100%mol mol^{-1})	r_{pred}	0,97	0,96	0,96
	RMSEP	3,81 (3)	4,45 (5)	3,74 (3)
Etano (0-50.0% mol mol^{-1})	r_{pred}	0,92	0,95	0,95
	RMSEP	4,46 (6)	3,98 (7)	3,71 (4)
Propano (0-31,5% mol mol^{-1})	r_{pred}	0,97	0,97	0,98
	RMSEP	1,11 (5)	1,46 (8)	1,38 (4)

Fonte: Própria do autor, 2022.

O modelo o LASSO apresentou um bom desempenho na predição. Os valores do coeficiente de correlação (r_{pred}) obteveram valores maiores que 0,95, demonstrando uma alta correlaçao entre os valores preditos e os de referência. No caso dos valores de RMSEP, todos ficaram abaixo de 10% da faixa experimental de trabalho de cada um dos analitos.

Foi realizado um teste F a um nivel de 95% de confiança para avaliar se havia diferenças estatisticamente significativas entre os modelos. Os valores obtidos estão expostos na **tabela 4**.

Tabela 4: Resultados do teste F a um nível de 95% de confiança obtidos na comparação entre o SPA-MLR, LASSO e PLS.

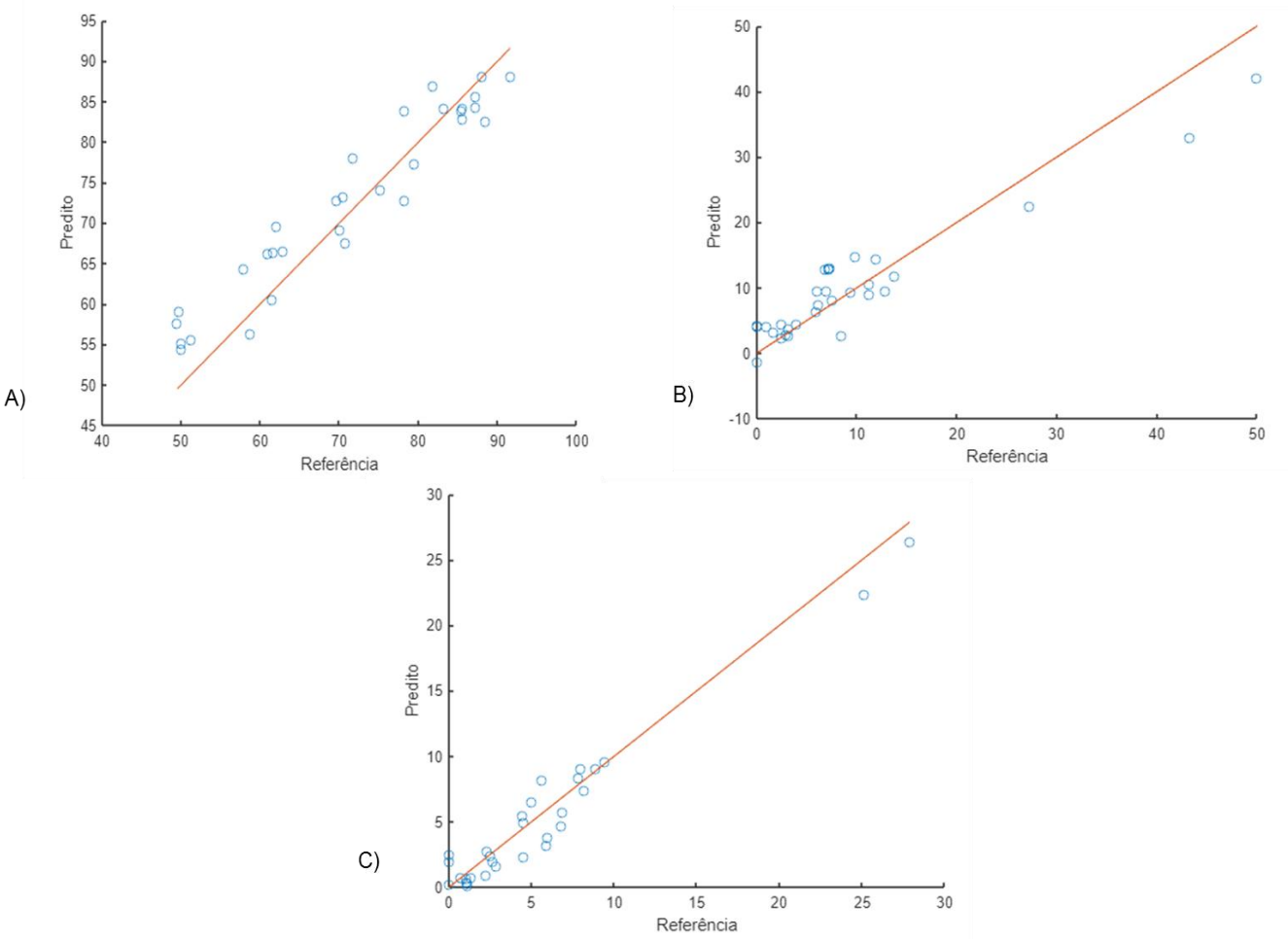
Modelos	**Metano**	**Etano**	**Propano**	**F crítico**
	F calculado	**F calculado**	**F calculado**	
SPA-MLR e LASSO	1,36	1,25	1,73	**1,82**
SPA-MLR e PLS	1,03	1,44	1,55	
LASSO e PLS	1,41	1,15	1,12	

Fonte: Próprio autor, 2022.

Observa-se que em todas os casos não houve diferenças estaticamente significativas entre as métricas estudadas, já que todos os valores de F calculado ficaram abaixo do F crítico. Logo, não há diferença mna precisão dos modelos.

Na **figura 8** estão presentes os gráficos de valores preditos pelos valores de referência do modelo Lasso para o metano, etano e propano.

Figura 8: Gráficos do valor predito pelo valor de referência para o modelo LASSO, onde corresponde: A) metano, B) etano e C) propano.



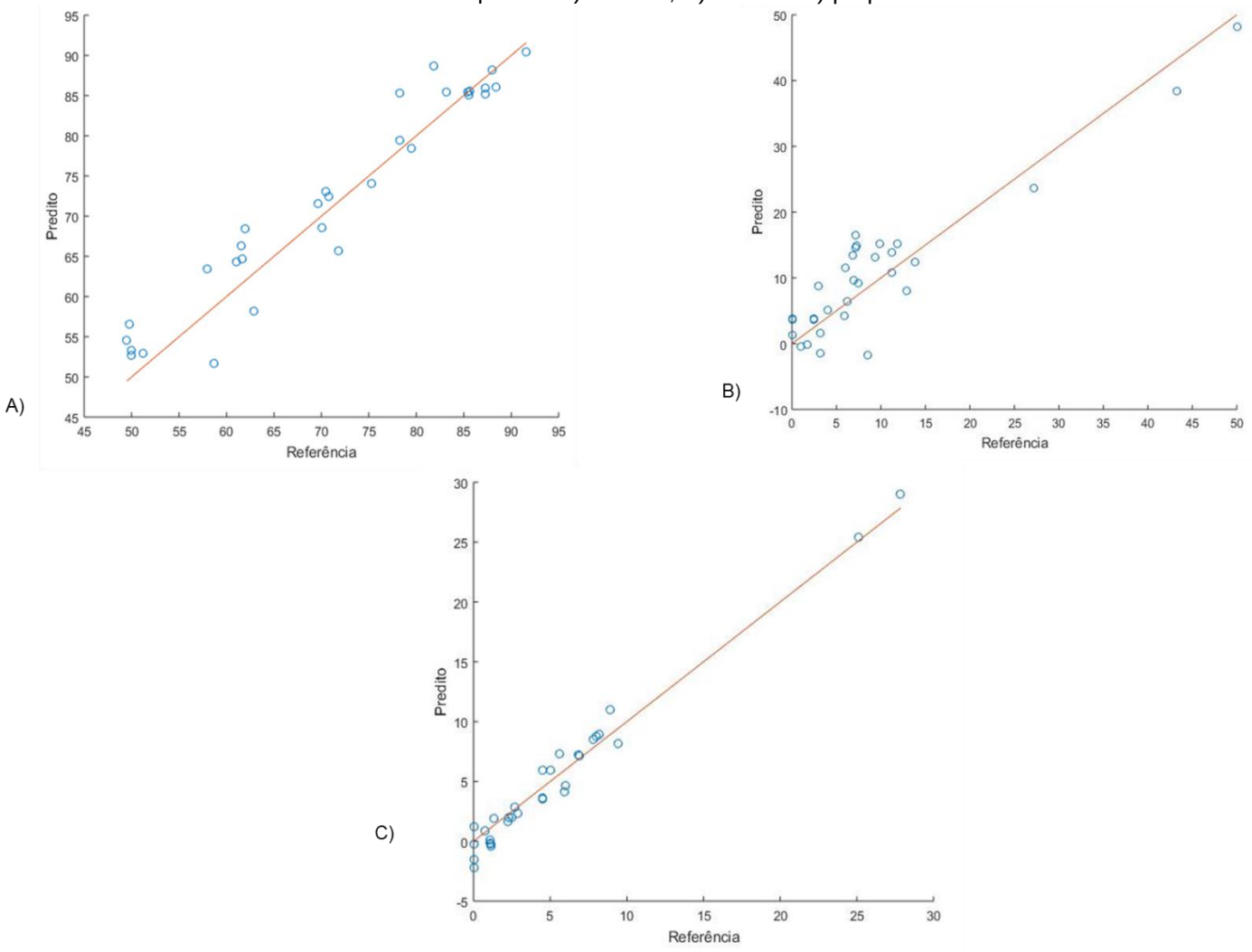
Fonte: Próprio autor, 2022.

No gráfico de predição do metano é possivel observar que as amostras estão dispostas aleatoriamente por toda bissetriz e sem a presença de anomalias nas amostras. No caso do etano, observa-se a presença de duas amostras mais afastadas das outras amostras e da bissetriz, porém as demais estão situadas próximas da linha de regresão. Já para o propano as amostras estão bem dispostas na bissetriz com duas amostras presentes no limite da faixa experimental de trabalho para o analito.

Portanto, os gráficos demonstram que o modelo consegue explicar bem a variação das amostras para o metano, etano e propano.

Na **figura 9** estão presentes os gráficos de valores preditos pelos valores de referência do modelo SPA-MLR para o metano, etano e propano.

Figura 9: Gráficos do valor predito pelo valor de referência para o modelo SPA-MLR, onde corresponde: A) metano, B) etano e C) propano.



Fonte: Próprio autor, 2022.

No gráfico de predição do metano, ocorreu o mesmo do gráfico do LASSO, as amostras estão dispostas aleatoriamente por toda bissetriz e sem a presença de anomalias nas amostras.

Já para o etano, observa-se a presença de duas amostras mais afastadas das outras, análogo ao que ocorreu no gráfico do LASSO, porém no SPA as amostras estão menos dipersas em torno da bissetriz.

Já para o propano as amostras estão bem dispostas na bissetriz, apenas com duas amostras presentes no limite da faixa experimental de trabalho para o analito, mas próximas da linha de regressão.

Com isso, pode-se concluir a partir da análise dos gráficos do modelo SPA-MLR que o mesmo consegue explicar de forma satisfatória a variação das amostras, para os analitos em estudo.

6. CONCLUSÃO

Nesse trabalho foi avaliado uma metodologia de regressão, o LASSO, para determinação de metano, etano e propano em amostras de gás natural e biogás. Os parâmetros obtidos pelo LASSO, RMSEP e r_{pred}, foram comparados com os obtidos utilizando os modelos MLR com a variáveis selecionadas pelo Algoritmo das Projeções Sucessivas e usando os espectros completos pelo método PLS. Sendo os espectros obtidos usando um equipamento NIR miniaturizado com uma faixa espectral de 900 a 1670 nm.

Considerando primeiramente o caso do analito metano, o LASSO obteve o valor de RMSEP de 4,45% mol mol^{-1} e r_{pred} de 0,96, sendo similar ao modelo SPA-MLR, que obteve um valor de RMSEP de 3,81% mol mol^{-1} e r_{pred} de 0,97, e ao PLS, com RMSEP de 3,74% mol mol^{-1} e r_{pred} de 0,96.

Para o etano, o modelo LASSO apresentou um resultado similar ao PLS e superior ao SPA-MLR, com o valor de RMSEP de 3,98% mol mol^{-1} e r_{pred} de 0,95, enquanto que para o modelo PLS o RMSEP foi de 3,71% mol mol^{-1} e r_{pred} de 0,95, e para o SPA-MLR o RMSEP foi de 4,46% mol mol^{-1} e r_{pred} de 0,92.

Considerando o analito propano, observa-se que os valores do LASSO, não divergiram de modo significativo dos valores do PLS e SPA-MLR. Onde os valores de RMSEP de 1,46% mol mol^{-1} e r_{pred} de 0,97 correspondem ao LASSO e os valores de RMSEP de 1,15 e 1,38% mol mol^{-1} correspondem ao SPA-MLR e ao PLS, respectivamente, e com valores de r_{pred} de 0,97 para o SPA-MLR e de 0,98 para o PLS.

A partir da análise dos gráficos de valores preditos pelos valores de referência conclui-se que o modelo LASSO consegue explicar bem a variação das amostras para os três analitos.

O teste F mostrou que os modelos não apresentaram diferenças estatiscamente significativas entre as metodologias estudadas. Mostrando que o modelo LASSO pode ser utilizado de forma satisfatória para a predição dos analitos metano, etano e propano.

Portanto, o estudo antingiu seu objetivo inicial que é demonstrar que o LASSO consegue realizar a predição dos analitos em amostras de gás natural e biogás de modo satisfatório, da mesma forma que os modelos PLS e SPA-MLR.

REFERÊNCIAS

AHETO, J. M. K., DUAH, H. O., AGBADI, P., & NAKUA, E. K..**A predictive model, and predictors of under-five child malaria prevalence in Ghana: How do LASSO, Ridge and Elastic net regression approaches compare? *Preventive*** *Medicine Reports, 23,* 101475, 2021.

ALCÂNTARA JUNIOR, G.P.**Avaliação do lasso e métodos alternativos em modelos de regressão logística**. 2021. 138 f. Dissertação (Mestrado em Estatística) - Universidade Federal de São Carlos, São Carlos, 2021.Disponível em:https://repositorio.ufscar.br/bitstream/handle/ufscar/14052/dissertação.pdf?sequence=1&isAllowed=y. Acesso em: 28 de setembro de 2022

ANP- Agência Nacional de Petróleo, Gás Natural e Biocombustíveis. Disponível em: https://www.gov.br/anp/pt-br/assuntos/producao-de-derivados-de-petroleo-e-processamento-de-gas-natural/processamento-de-gas-natural/gas-natural. Acessado em: 19 de março de 2022.

ANP(Biometano) - Agência Nacional de Petróleo, Gás Natural e Biocombustíveis. Disponível em: https://www.gov.br/anp/pt-br/assuntos/producao-e-fornecimento-de-biocombustiveis/biometano. Acessado em: 15 de abril de 2022.

BARBOSA, M. F., SANTOS, J. R. B., SILVA, A. N., SOARES, S. F. C., & ARAUJO, M. C. U.. **A cheap handheld NIR spectrometric system for automatic determination of methane, ethane, and propane in natural gas and biogas***. Microchemical Journal, 170, 106752,* 2021*.*

BEĆ, Krzysztof B.; GRABSKA, Justyna; HUCK, Christian W.. **Principles and applications of miniaturized near-infrared (NIR) spectrometers**. *Chemistry–A European Journal,* v. 27, n. 5, p. 1514-1532, 2021.

BP - Statistical Review of World Energy 2022. Disponível em: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2022-full-report.pdf Acessado em: 25 de setembro de 2022

BRAGA, J.W.B; POPPI, R.J. **Validação de modelos de calibração multivariada: uma aplicação na determinação de Pureza polimórfica de carbamazepina por espectroscopia no infravermelho próximo.** *Quim. Nova,* 27 1004-1011, 2004.

BRASIL. Resolução ANP N° 16, DE 17.06.2008 – DOU 18.06.2018. Disponível em: https://atosoficiais.com.br/anp/resolucao-n-16-2008?origin=instituicao&q=16/2008. Acessado em: 20 de março de 2022.

CAMPOS, A. F.; SILVAA, N. F.; PEREIRA, M. G.; FREITAS, M. A. V. **A review of Brazilian natural gas industry: Challenges and strategies.** *Renewable and Sustainable Energy Reviews* 75 1207-1216, 2017.

CHEN, H.; CHEN, T.; ZHANG, Z.; LIU, G. **Variable Selection Using Adaptive Band Clustering and Physarum Network**. *Algorithms 10*, 73, 2017.

CHEN, J., YU; J., AI, B.; SONG, M.; & HOU, W**.. Determinants of global natural gas consumption and import–export flows.** *Energy Economics*. v. 83, p. 588-602, 2019.

CRUZ-TIRADO, J. P.; DA SILVA MEDEIROS, M. L.; BARBIN, D. F.. **On-line monitoring of egg freshness using a portable NIR spectrometer in tandem with machine learning**. *Journal of Food Engineering*, v. 306, p. 110643, 2021.

DANTAS, H. V.; BARBOSA, M. F.; MOREIRA, P. N.; GALVÃO, R. K.; ARAÚJO, M. C. **An automatic system for accurate preparation of gas mixtures**. Microchemical Journal, v. 119, p. 123–127, 2015.

FARAMAWY, S.; ZAKI, T.; SAKR, A. A.-E. **Natural gas origin, composition, and processing: A review**. *J Natural Gas Sci and Eng* 34, 34-54, 2016.

GOODARZI, M.; SHARMA, S.; RAMON, H.; SAEYS, W**. Multivariate calibration of NIR spectroscopic sensors for continuous glucose monitoring**. *TRAC - Trend Anal Chem* 67, 147-158, 2015.

HASTIE, T.; TIBSHIRANI, R.; WAINWRIGHT, M. **Statistical learning with sparsity: The lasso and generalizations**. Seattle: Chapman and Hall Book, 2015.

IBP - Instituto brasileiro de petróleo e gás. Disponível em: https://www.ibp.org.br/observatorio-do-setor/snapshots/maiores-consumidores-de-gas-natural-em-2020/. Acessado em: 25 de setembro de 2022.

IEA – Internacional Energy Agency. Disponível em: https://www.iea.org/topics/naturalgas. Acessado em: 20 de março de 2022.

KUCHA, CHRISTOPHER T.; NGADI, MICHAEL O. **Rapid assessment of pork freshness using miniaturized NIR spectroscopy.** *Journal of Food Measurement and Characterization*, v. 14, n. 2, p. 1105-1115, 2020.

KUMAR, N.; BANSAL, A.; SARMA, G. S.; RAWAL, R. K. **Chemometrics tools used in analytical chemistry: An overview**. *Talanta* 123, p.186-199, 2014.

LI, LINA; GUO, SHANGXING. **A Wavelength Selection Model Based on Successive Projections Algorithm for pH Detection of Water by VIS-NIR Spectroscopy**. *In: Journal of Physics: Conference Series. IOP Publishing*, p. 012002, 2021.

LUAN, X., LIU, J., & LIU, F**.. Multilevel LASSO-based NIR temperature-correction modeling for viscosity measurement of bisphenol-A***. ISA Transactions.107,* p. 206-213, 2020.

MANLEY, M.; & BAETEN, V.. **Spectroscopic Technique: Near Infrared (NIR) Spectroscopy.** *Modern Techniques for Food Authentication*, p. 51–102, 2018.

MORAIS, P. A. DE O., D. M., MADARI, B. E., DE OLIVEIRA, A. E. **. Predicting silicon, aluminum, and iron oxides contents in soil using computer vision and infrared.** *Microchemical Journal,* v. 170, p. 106669, 2021.

PASQUINI, C. **Near infrared spectroscopy: A mature analytical technique with new perspectives – A review**. *Analytica Chimica Acta*, 1026, p. 8–36, 2018.

PASQUINI, C.. **Near Infrared Spectroscopy: fundamentals, practical aspects and analytical applications**. *Journal of the Brazilian Chemical Society*, 14(2), p. 198–219, 2003.

PU, Y., PÉREZ-MARÍN, D., O'SHEA, N., GARRIDO-VARO, A. **Recent advances in portable and handheld NIR spectrometers and applications in milk, cheese and dairy powders**. *Foods*, v. 10, n. 10, p. 2377, 2021.

.RAFIEE, A., KHALILPOUR, K. R., PREST, J., SKRYABIN, I. **Biogas as an energy vector**. *Biomass and Bioenergy*, v. 144, p. 105935, 2021.

ROY, K.; AMBURE, P. **The "double cross-validation" software tool for MLR QSAR model development**. *Chem Intel Lab Sys* 159, p. 108-126, 2016.

SHAFIEE, S., LIED, L. M., BURUD, I., DIESETH, J. A., ALSHEIKH, M., & LILLEMO, M.. **Sequential forward selection and support vector regression in comparison to LASSO regression for spring wheat yield prediction based on UAV imagery.** *Computers and Electronics in Agriculture, 183,* 106036, 2021.

SOARES, S. F. C.; GOMES, A. A.; ARAUJO, M. C. U.; FILHO, A. R. G., GALVÃO, R. K. H.. **The successive projections algorithm.** *TrAC Trends in Analytical Chemistry,* 42 p, 84–98, 2013.

TIBSHIRANI, R**. Regression Shrinkage and Selection via the LASSO**. *Royal Statistics Society*, v. 58, n. 1, p. 267–288, 1996.

TIBSHIRANI, R.. **Regression shrinkage and selection via the lasso: a retrospective.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3),* p. *273–282,* 2011*.*

VOLMER,M. **Infrared spectroscopy in clinical chemistry, using chemometric calibration techniques.** *Nederland: Proefschriift Groningen*, 2001.

WANG, Y.J., LI, T. H., LI, L. Q., NING, J. M., & ZHANG, Z. Z. . **Evaluating taste-related attributes of black tea by micro-NIRS**. *Journal of Food Engineering*, v. 290, p. 110181, 2021.

WANG, YJ. et al. **NIR hyperspectral imaging coupled with chemometrics for nondestructive assessment of phosphorus and potassium contents in tea leaves.** *Infrared Physics & Technology*, v. 108, p. 103365, 2020.