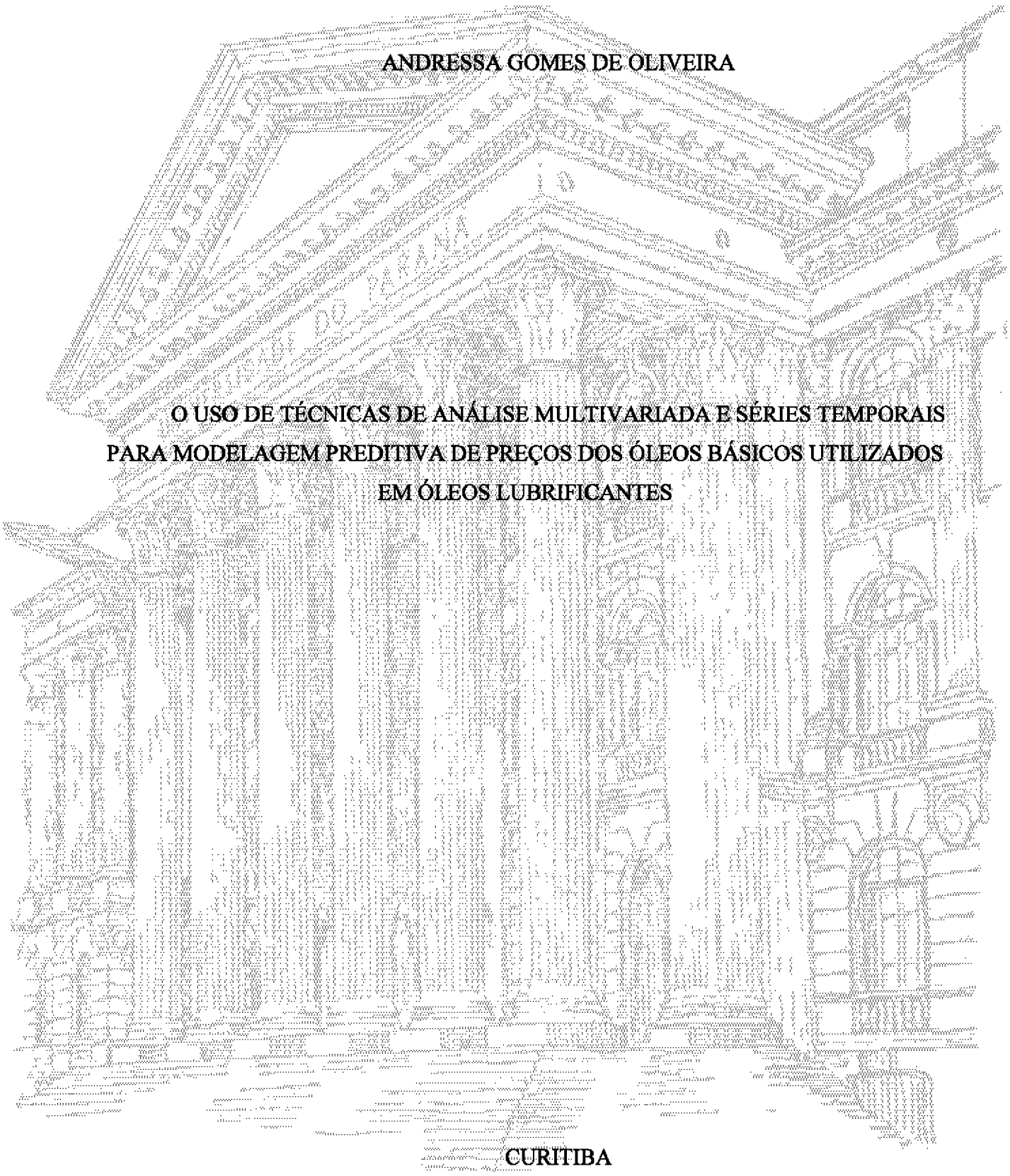
UNIVERSIDADE FEDERAL DO PARANÁ

ANDRESSA GOMES DE OLIVEIRA



O USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA E SÉRIES TEMPORAIS PARA MODELAGEM PREDITIVA DE PREÇOS DOS ÓLEOS BÁSICOS UTILIZADOS EM ÓLEOS LUBRIFICANTES

CURITIBA

2025

ANDRESSA GOMES DE OLIVEIRA

O USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA E SÉRIES TEMPORAIS PARA MODELAGEM PREDITIVA DE PREÇOS DOS ÓLEOS BÁSICOS UTILIZADOS EM ÓLEOS LUBRIFICANTES

Dissertação apresentada ao curso de Pós-Graduação em Métodos Numéricos em Engenharia, Área de Concentração em Programação Matemática, Linha de pesquisa em Métodos Estatísticos dos Setores de Ciências Exatas e de Tecnologia da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Métodos Numéricos em Engenharia.

Orientador: Prof. Dr. Cassius Tadeu Scarpin
Coorientador: Dr. Alexandre Cancian Bajotto

CURITIBA

2025

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

Oliveira, Andressa Gomes de
O uso de técnicas de análise multivariada e séries temporais para modelagem preditiva de preços dos óleos básicos utilizados em óleos lubrificantes / Andressa Gomes de Oliveira. – Curitiba, 2025.
1 recurso on-line : PDF.

Dissertação (Mestrado) - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Métodos Numéricos em Engenharia.

Orientador: Cassius Tadeu Scarpin
Coorientador: Alexandre Cancian Bajotto

1. Óleo – Indústria. 2. Óleos lubrificantes. 3. Análise Multivariada. 4. Análise de séries temporais. 5. Análise de regressão. I. Universidade Federal do Paraná. II. Programa de Pós-Graduação em Métodos Numéricos em Engenharia. III. Scarpin, Cassius Tadeu. IV. Bajotto, Alexandre Cancian. V. Título.

Bibliotecário: Elias Barbosa da Silva CRB-9/1894

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação MÉTODOS NUMÉRICOS EM ENGENHARIA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **ANDRESSA GOMES DE OLIVEIRA**, intitulada: **O USO DE TÉCNICAS DE ANÁLISE MULTIVARIADA E SÉRIES TEMPORAIS PARA MODELAGEM PREDITIVA DE PREÇOS DOS ÓLEOS BÁSICOS UTILIZADOS EM ÓLEOS LUBRIFICANTES**, sob orientação do Prof. Dr. CASSIUS TADEU SCARPIN, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 10 de Dezembro de 2025.

Assinatura Eletrônica
17/12/2025 21:34:06.0
CASSIUS TADEU SCARPIN
Presidente da Banca Examinadora

Assinatura Eletrônica
07/01/2026 20:11:38.0
WYRLLEN EVERSON DE SOUZA
Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica
18/12/2025 08:50:12.0
ANSELMO CHAVES NETO
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
18/12/2025 16:25:47.0
CIBELE MARIA RUSSO NOVELLI
Avaliador Externo (INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO, UNIVERSIDADE DE SÃO PAULO)

Dedico este trabalho à minha mãe e ao meu noivo, que nunca mediram esforços para me apoiar e incentivar na busca pelos meus objetivos. Vocês me fazem perceber o cuidado de Deus para comigo. Amo vocês!

AGRADECIMENTOS

Encontrar pessoas com quem contar, sorrir e compartilhar bons momentos é uma dádiva.

Sou infinitamente grata a Deus e a Nossa Senhora por todas as vezes em que minha fé n'Eles foi o meu consolo e meu lugar de paz; pelas ocasiões em que, nos pequenos detalhes, me mostraram a importância do dom da perseverança e ainda me concederam a graça de conhecer pessoas incríveis.

Diante disso, agradeço à minha mãe, Maria de Lourdes, por todo apoio durante essa trajetória e por todas as vezes em que me ensinou a lutar pelos meus objetivos de maneira honesta e sendo gentil com as pessoas ao meu redor. Ao meu pai, José Everardo (*in memoriam*), por ter me mostrado que vale a pena se doar pelos outros.

Aos meus irmãos, que me apoiaram e me compreenderam nos meus momentos de estudo e fazem de tudo para que eu consiga alcançar meus objetivos.

Ao meu melhor amigo, meu amor e noivo, Izaquiel Celestino, por todas as vezes em que seu abraço foi o meu abrigo, seus conselhos foram meu incentivo, seus puxões de orelha foram suficientes para me motivar a melhorar e por acreditar na minha capacidade, mesmo quando eu duvidava disso. Você é uma das pessoas mais incríveis que conheço, e que sorte a minha ter você em minha vida.

À professora Sissy e aos professores Sérgio e Carlos Matioli, por terem me apresentado o Programa de Pós-Graduação em Métodos Numéricos (PPGMNE). Vocês acreditaram em mim, e sou profundamente grata por isso.

Ao meu orientador, Cassius Tadeu Scarpin, por, mesmo sem me conhecer, ter acreditado e confiado em mim, não apenas para o mestrado, mas também em projetos de pesquisa. Você é um profissional excelente. Obrigada por tudo!

Ao meu colega de projeto de pesquisa, Alexandre Cancian Bajotto, pelas orientações e pela disponibilidade em me auxiliar em todos os momentos em que precisei, e não foram poucos, rsrs. Que Deus e Nossa Senhora abençoem grandemente sua vida e sua família.

Aos colegas que, mesmo de forma remota, foram essenciais durante esse percurso. Em especial, agradeço à Juliana Casellas pela generosidade em sempre auxiliar quando precisei. Aos colegas do projeto de pesquisa, em especial à Rúbia, ao Júlio e ao Edison, pela ajuda constante. Gratidão por tudo!

Que a tua vida não seja uma vida estéril — Sê útil — deixa rastro.

(São Josemaría Escrivá)

Ao brilhar de um relâmpago nascemos e ainda dura seu fulgor quando morremos.

Tão curto é o viver.

Para que serve um ser humano?

Para tocar fogo de esperança nesse mundo!

E o que eu quero se não que arda?

E o que eu quero se não que arda?

Não deixar que nenhuma pessoa passe por você sem que ela saia um pouco melhor do que ela entrou.

Que você seja esse girassol que brota nessa rua de desatenção dessa vida vulgarmente cotidiana.

Entenda que o mundo não é muito dado às coincidências, as coisas têm um propósito, um motivo.

Que essa seja a grande oportunidade da sua vida para ajudar outras dezenas, centenas de pessoas.

Aos corações que amam, há olhares de esperança nesse mundo.

(Ítalo Marsili, Girassol)

RESUMO

A precificação de Óleos Básicos, principais insumos utilizados na formulação de lubrificantes, é influenciada por múltiplos fatores econômicos, financeiros e setoriais, tornando sua previsão um desafio relevante para a indústria. Este estudo propõe uma abordagem inédita que integra técnicas de Análise Multivariada e de Séries Temporais para prever os preços desses insumos e identificar os fatores que mais influenciam suas variações. O banco de dados inicial é composto por 227 variáveis relacionadas à oferta e demanda de petróleo, indicadores macroeconômicos, commodities, moedas, índices financeiros e ações de empresas do setor, sendo os dados dispostos de janeiro de 2010 a julho de 2023. Foram analisadas as séries de preços mensais de 31 Óleos Básicos dos Grupos I, II e III. A partir das técnicas de Análise Multivariada empregadas, foi possível formar, a partir do banco original, dois subconjuntos estruturados de variáveis, um com 135 e outro com 71 variáveis, preservando apenas aquelas com maior relevância estatística ou comportamento semelhante. Assim, todos os testes foram realizados considerando as três bases: (a) 227 variáveis, (b) 135 variáveis e (c) 71 variáveis. Essa redução de dimensionalidade visou especificamente aprimorar a precisão da previsão, sendo que essas bases de dados variáveis demonstraram bom desempenho preditivo para Óleos Básicos e horizontes específicos, validando a eficácia da seleção de variáveis. Inicialmente, aplicaram-se Análise de Componentes Principais (ACP) e Análise Fatorial (AF) para reduzir a dimensionalidade e identificar estruturas latentes no conjunto de dados. Em seguida, técnicas de Reconhecimento de Padrões, como a Análise de Agrupamentos (*Cluster Analysis*) e a Análise Discriminante, foram utilizadas para validar a separação entre grupos homogêneos de variáveis. Com base nos resultados Multivariados, desenvolveu-se um modelo de Regressão Linear Múltipla para prever os preços dos Óleos Básicos. Para incorporar a dinâmica temporal das variáveis independentes, foram projetados valores futuros via modelos ARIMA e utilizados como entradas na Regressão, permitindo avaliar a capacidade preditiva para horizontes de 3, 6 e 12 meses. A acurácia dos modelos foi avaliada pelo MAE, MAPE e MSE. Os resultados demonstraram desempenho robusto, especialmente no horizonte de seis meses, no qual a maioria dos modelos apresentou MAPE inferior a 5%. Em resumo, é possível afirmar que a integração de técnicas de Análise Multivariada, agrupamento e modelos de Séries Temporais constitui uma abordagem eficaz para previsão de preços de Óleos Básicos e para a identificação das variáveis com maior influência na precificação.

Palavras-chave: Óleos Básicos; Análise Multivariada; Séries Temporais; Análise de Componentes Principais e Fatorial; Regressão Múltipla.

ABSTRACT

The pricing of Base Oils, the main inputs used in the formulation of lubricants, is influenced by multiple economic, financial, and sectoral factors, making their forecasting a relevant challenge for the industry. This study proposes an innovative approach that integrates Multivariate Analysis techniques and Time Series models to predict the prices of these inputs and identify the factors that most strongly influence their variability. The initial dataset comprises 227 variables related to global petroleum supply and demand, macroeconomic indicators, commodities, exchange rates, financial indices, and energy-related company stocks, with data spanning from January 2010 to July 2023. Monthly price series for 31 Base Oils from Groups I, II, and III were analyzed. Based on the multivariate techniques employed, two structured subsets of variables were derived from the original dataset, containing 135 and 71 variables, respectively. These subsets retained only the most relevant variables or those exhibiting similar statistical behavior. Thus, all tests were conducted across three datasets: (a) 227 variables, (b) 135 variables, and (c) 71 variables. This dimensionality reduction aimed to enhance forecasting accuracy, and these variable datasets demonstrated strong predictive performance for specific Base Oils and forecast horizons, validating the effectiveness of the variable selection process. Principal Component Analysis (PCA) and Factor Analysis (FA) were first applied to reduce dimensionality and identify latent structures within the dataset. Subsequently, pattern-recognition techniques, such as Cluster Analysis and Discriminant Analysis, were used to validate the separation of homogeneous groups of variables. Based on the multivariate results, a Multiple Linear Regression model was developed to forecast Base Oil prices. To incorporate the temporal dynamics of the independent variables, future values were projected using ARIMA models and inserted into the regression equation, enabling the evaluation of predictive performance for 3-, 6-, and 12-month horizons. Model accuracy was assessed using MAE, MAPE, and MSE. Results showed robust performance, especially for the six-month horizon, in which most models achieved MAPE values below 5%. In summary, the integration of Multivariate Analysis, clustering techniques, and Time Series models provides an effective approach for forecasting Base Oil prices and identifying the variables that exert the greatest influence on their pricing behavior.

Keywords: Base Oils; Mutivariate Analysis; Time Series; Principal Component and Factor Analysis e Fatorial; Mutiple Regression.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Óleos lubrificantes são misturas complexas de hidrocarbonetos saturados e aromáticos, obtidos a partir do refino do petróleo, com a função de reduzir o atrito e promover maior eficiência energética em motores (AL-ZAHRANI; PUTRA, 2013; FREITAS *et al.*, 2003).

Segundo Lima (2016), os óleos lubrificantes consistem em dois materiais: óleo base e aditivos químicos. O óleo base refere-se aos Óleos Básicos, que são os principais componentes dos lubrificantes acabados, ou seja, prontos para serem utilizados. São compostos por uma mistura de moléculas com 18 a 40 átomos de carbono e podem ser classificados como parafínicos, naftênicos ou aromáticos, podendo apresentar pequenas porcentagens de enxofre, nitrogênio ou oxigênio (LIMA, 2016). A classificação desses óleos, segundo a *American Petroleum Institute* (API), é dividida em cinco grupos, de acordo com o processo de obtenção, teor de saturados, teor de enxofre e índice de viscosidade.

Os óleos do Grupo I são minerais obtidos por refino convencional, com maior teor de compostos aromáticos e índice de viscosidade entre 80 e 120, enquanto os do Grupo II, também minerais, passam por processos adicionais, como hidrogenação, apresentando maior pureza e menor teor de enxofre. O Grupo III compreende óleos minerais altamente refinados, frequentemente denominados "sintéticos", com índice de viscosidade acima de 120. Já o Grupo IV inclui óleos sintéticos verdadeiros, como as polialfaolefinas (PAO), caracterizados pelo alto desempenho em uma ampla faixa de temperaturas. Por fim, o Grupo V abrange Óleos Básicos especiais, como ésteres e óleos vegetais modificados, frequentemente utilizados como aditivos em formulações específicas.

De acordo com Ayub (2019), o crescimento das tecnologias digitais aplicadas aos processos de fabricação tem levado as organizações a modificar sua eficiência em processos e desempenho. Assim, tendo em vista os custos operacionais na produção dos Óleos Básicos, o uso de técnicas estatísticas multivariadas que auxiliem na previsão de preços torna-se fundamental para aprimorar a organização e o planejamento da produção (DAVE, *et. al.*, 2003). Tais técnicas consistem no tratamento de dados correspondentes às medidas de várias variáveis simultaneamente.

Neste estudo, foi utilizado um banco de dados formado por 227 de diferentes aplicações como: (a) oferta, demanda, estoques de petróleo e derivados, e capacidade de refino em diferentes continentes, (b) commodities minerais e de energia derivados do petróleo, (c) índices Dow Jones setoriais, (d) ações de empresas de mundiais de energia, (e) paridade entre

moedas mundiais e o dólar, (f) índices de bolsa globais e (g) quantidade de moeda em circulação como indicador de liquidez mundial. A variável dependente analisada é o preço de 31 Óleos Básicos fornecidos pela empresa A, cujo nome será omitido por questões de confidencialidade. Dentre os 31 Óleos Básicos, 15 pertencem ao Grupo 1, 10 ao Grupo 2 e os demais pertencem ao Grupo 3.

Entre esses produtos, um Óleo Básico específico apresenta a série histórica mais longa, com registros mensais de janeiro de 2010 à fevereiro de 2024. Os demais 30 Óleos Básicos contam com séries históricas mensais que abrangem o período de janeiro de 2015 à junho de 2024. No entanto, para esta pesquisa, a análise com dados reais de todas os Óleos Básicos foi restrita aos dados reais até julho de 2023, para que se alinhasse com a disponibilidade das variáveis independentes.

A partir desses dados, foi gerado um modelo de regressão linear múltipla, levando em consideração os resultados obtidos nas técnicas de Análise Multivariada utilizadas. Para avaliar a precisão do modelo, utilizou-se a projeção de valores futuros com a aplicação do modelo ARIMA. Esses valores projetados foram, então, incorporados na equação de Regressão Múltipla para verificar a consistência e a capacidade preditiva do modelo nos períodos futuros.

Dentre as técnicas estatísticas aplicadas estão a Análise Fatorial, Análise das Componentes Principais, Análise de Agrupamentos (*Cluster Analysis*) e Análise Discriminante. Essas abordagens permitem identificar padrões, reduzir a dimensionalidade dos dados e agrupar variáveis com características semelhantes (JOHNSON; WICHERN, 1998). Para a verificação da precisão dos modelos considerou-se as métricas de erros como MAE (*Mean Absolute Error*), MAPE (*Mean Absolute Percentage Error*) e MSE (*Mean Squared Error*).

As abordagens multivariadas podem viabilizar a obtenção de resultados consistente, tanto de fabricação quanto de compra, permitindo previsões para decisões futuras, que visam proporcionar uma vantagem competitiva viável e diferenciada para o negócio (AYUB, 2019).

Embora a análise tenha sido aplicada especificamente aos Óleos Básicos dos Grupos 1, 2 e 3, a metodologia proposta é flexível e pode ser adaptada para outras categorias de produtos ou diferentes mercados, bastando adequar as variáveis independentes de acordo com as particularidades químicas, econômicas ou mercadológicas analisadas.

1.1 OBJETIVOS

Para responder à pergunta da pesquisa e alcançar os resultados desejados, foram elaborados os objetivos que servirão como base para o estudo. Assim, o objetivo geral e os objetivos específicos são apresentados a seguir:

1.1.1 Objetivo geral

Desenvolver um modelo de Regressão Múltipla baseado em técnicas de Análise Estatística Multivariada e Séries Temporais, com o objetivo de prever os preços de Óleos Lubrificantes Básicos, e analisar como as variáveis independentes utilizadas influenciam esses preços.

1.1.2 Objetivos específicos

Os objetivos específicos do trabalho são:

a) Tratar os dados para garantir resultados consistentes.
b) Projetar valores futuros a partir dos dados originais organizados em série temporal, utilizando um modelo ARIMA.
c) Aplicar métodos estatísticos multivariados, como análise das Componentes Principais e Análise Fatorial, para gerar uma equação de Regressão Múltipla com base nos resultados obtidos.
d) Validar e ajustar os resultados da análise de *Cluster* e verificar a precisão dos grupos formados, utilizando análise discriminante.
e) Identificar, a partir do banco de dados, quais variáveis têm impacto positivo (preço sobe) e negativo (preço desce).

1.2 IMPORTÂNCIA DO TRABALHO

O petróleo cru, fonte primária para a fabricação de lubrificantes, é amplamente reconhecido como uma *commodity* essencial devido à sua influência no crescimento econômico global (Jha *et al.*, 2024). As mudanças drásticas nos preços do petróleo nos últimos anos,

caracterizadas por aumentos e quedas substanciais, afetam não apenas a economia global, mas também representam desafios para a indústria de lubrificantes.

Segundo a Petrobras (2010), o preço dos Óleos Básicos, principais componentes dos lubrificantes, é influenciado pelo preço do petróleo. No entanto, suas respostas às variações no preço do petróleo são bem mais lentas do que as observadas nos combustíveis. Isso indica que, embora o petróleo seja uma matéria-prima chave na formulação de Óleos Básicos, outros fatores, como condições de mercado, flutuações financeiras, dinâmica da oferta e demanda, e eventos externos, como mudanças políticas ou econômicas podem afetar a precificação de lubrificantes de forma diferente (JHA *et al.*, 2024).

Com o aumento no volume de dados compostos por medições simultâneas de diversas variáveis, tornou-se necessário o uso de técnicas específicas para o seu tratamento (JOHNSON e WICHERN, 1998). Nesse contexto, a análise de dados desempenha um papel fundamental para as organizações, auxiliando na geração de melhorias de processos e custos (LIBES, SHIN e WOO, 2015).

De acordo com Johnson e Wichern (1998), a análise de dados frequentemente envolve a remoção ou adição de variáveis, e as abordagens multivariadas permitem realizar análises preditivas de forma mais robusta.

Com isso, a importância deste estudo, destaca-se por apresentar diversas técnicas de Análise Multivariada aplicadas a um conjunto de dados composto por variáveis de mercados financeiros, monetários, de petróleo, de metais, entre outros. O objetivo principal é desenvolver uma equação de Regressão Múltipla como modelo preditivo e avaliar o impacto dessas variáveis no preço de alguns Óleos Básicos, que são amplamente utilizados na fabricação de lubrificantes.

Além disso, as técnicas utilizadas auxiliaram na redução do volume de dados, preservando as informações essenciais das variáveis, como nas Análises das Componentes Principais e Fatorial. Adicionalmente, as análises de cluster e discriminante foram aplicadas para identificar e agrupar variáveis semelhantes, permitindo a adição ou remoção de variáveis no banco de dados conforme necessário.

Por fim, este trabalho contribui para auxiliar as organizações na tomada de decisões, fornecendo informações importantes sobre os preços futuros dos Óleos Básicos. Além disso, oferece vantagens competitivas ao destacar quais variáveis exercem maior influência na precificação desses óleos. A originalidade e relevância deste estudo estão na integração de diversas técnicas de Análise Multivariada no mercado de Óleos Básicos, servindo de base para estudos futuros voltados à previsão de preços.

1.3 LIMITAÇÃO DO TRABALHO

Apesar das contribuições deste estudo, é importante considerar algumas limitações. Primeiramente, o trabalho utiliza exclusivamente dados quantitativos, contendo variáveis financeiras e de mercado. Isso significa que aspectos qualitativos, como eventos políticos, mudanças regulatórias, percepção do mercado e período atípicos como o da pandemia da COVID-19, apesar de o banco de dados analisado abranger esse período, foram desconsiderados. Esses fatores, embora difíceis de quantificar, podem exercer influência significativa sobre os preços dos Óleos Básicos.

Outra limitação está relacionada ao uso do modelo ARIMA para projeções. Embora seja eficiente para séries temporais, o ARIMA baseia-se em pressupostos de linearidade e estacionaridade nos dados, características que nem sempre refletem a dinâmica complexa dos mercados financeiros e de *commodities*. Contudo, vale ressaltar que este é um dos modelos clássicos mais utilizados e que foram realizados tratamentos específicos para adequar o banco de dados à aplicação do ARIMA, como o teste de ADF (*Augmented Dickey-Fuller test*) para verificar a estacionariedade da série e a análise do critério CAIC (*Consistent Akaike Information Criterion*) para seleção do modelo mais adequado.

Adicionalmente, o estudo foi focado exclusivamente nos Óleos Básicos do Grupo 1, 2 e 3. Como os outros grupos de Óleos Básicos possuem características distintas, tanto na composição quanto no comportamento de mercado, as conclusões deste trabalho podem não ser diretamente aplicáveis a esses outros grupos. No entanto, os tratamentos e as técnicas empregadas neste estudo podem ser reaplicados.

1.4 ESTRUTURA DO TRABALHO

A estrutura do trabalho está organizada em cinco capítulos, dispostos de forma a facilitar o entendimento da pesquisa e dos procedimentos adotados.

No capítulo 1, apresenta-se uma breve análise do problema, com destaque para os objetivos, a justificativa, as limitações e a estrutura do trabalho.

O capítulo 2 aborda a descrição do problema, mostrando os impasses e o que se deseja solucionar.

O capítulo 3 o referencial teórico, trazendo uma revisão de estudos relacionados ao tema e os conceitos fundamentais sobre Óleos Básicos. Também são apresentadas as definições das

principais técnicas de Análise Multivariada, e os fundamentos de Séries Temporais, com ênfase no modelo de previsão ARIMA.

No capítulo 4, são detalhados os materiais e métodos utilizados na estruturação e tratamento do banco de dados. Este capítulo inclui informações sobre as variáveis analisadas, o período do estudo, o pré-tratamento aplicado e o procedimento adotado para cada técnica multivariada, como Análise das Componentes Principais, Análise Fatorial, Análise de *Cluster* e Análise Discriminante.

O capítulo 5 apresenta os resultados obtidos na pesquisa, acompanhados de uma discussão mais detalhada sobre as análises e técnicas empregadas.

Por fim, o capítulo 6 traz as considerações finais, sintetizando os principais resultados da pesquisa, bem como sugestões e recomendações para estudos futuros.

2 DESCRIÇÃO DO PROBLEMA

O mercado de Óleos Básicos é um componente vital da indústria de lubrificantes, representando entre 70% e 90% de sua composição (SAINI *et al.*, 2020). Apesar da sua importância, ele tem sido pouco explorado em estudos acadêmicos, especialmente no que diz respeito à influência de variáveis externas, como os preços do petróleo bruto, sobre suas flutuações.

Um relatório da Maximize Market Research (MMR) de outubro de 2024 aponta que a receita de Óleos Básicos deve crescer a uma taxa composta anual (CAGR) de 5% entre 2024 e 2030. Embora o valor desses produtos esteja diretamente ligado ao preço do petróleo bruto, os Óleos Básicos apresentam menor volatilidade em comparação com outros derivados do petróleo, com variações que não são imediatamente refletidas no mercado, conforme observado pela Kline & Company.

Diversos estudos sobre o mercado de petróleo indicam que os aumentos nos preços do petróleo podem impactar negativamente o desempenho das ações, com implicações significativas para a economia global e os mercados financeiros (JONES e KAUL, 1996; SADORSKY, 1999; WANG e WANG, 2016). De acordo com a DYM Resource (2023), esse impacto não se limita aos mercados financeiros, mas também afeta diretamente a cadeia de suprimentos de lubrificantes, como os setores automotivo, metalúrgico e industrial. Como o custo do Óleo Básico representa uma parte significativa do custo final de produção dos lubrificantes, ele influencia sua competitividade e disponibilidade para as indústrias.

Outros estudos apontam que os preços do petróleo e seus derivados também são afetados por diversas variáveis macroeconômicas, como atividade econômica (HUANG *et al.*, 2021), oferta e demanda (SHAH e KIRUTHIGA, 2020) e índices do mercado financeiro (DU *et al.*, 2023). Desta forma, o desenvolvimento de modelos preditivos que considerem essas variáveis é estratégico para estimar os preços futuros dos Óleos Básicos, auxiliando na formulação de contratos mais precisos e reduzindo a exposição a oscilações inesperadas no mercado. No entanto, a literatura especializada ainda carece de estudos aprofundados sobre os fatores que impactam diretamente o preço dos Óleos Básicos.

Diante da relevância do Óleo Básico na indústria de lubrificantes e de sua menor volatilidade em comparação a outros derivados do petróleo, as empresas do setor frequentemente firmam contratos de longo prazo para garantir estabilidade no fornecimento e previsibilidade financeira. Segundo Chiam e Ahuja (1997), contratos desse tipo são necessários para mitigar riscos e assegurar margens de lucro estáveis ao longo do período de vigência,

permitindo que empresas planejem sua produção e estrutura de custos com maior segurança. Entretanto, a eficácia desses acordos depende da capacidade de compreender e antecipar os fatores que influenciam o preço do Óleo Básico. Embora os mercados de futuros de petróleo para prazos mais longos apresentem menor liquidez, esses horizontes estendidos são frequentemente considerados por formuladores de políticas e agentes do mercado devido à sua importância na gestão de riscos e no planejamento estratégico (ALQUIST; KILIAN, 2005).

Diante desse contexto, o problema geral pode ser dividido em dois aspectos: o primeiro diz respeito à identificação das variáveis externas que influenciam o preço dos Óleos Básicos minerais, considerando fatores como os índices do mercado financeiro, a atividade econômica, as dinâmicas da cadeia de suprimentos, dentre outros. O segundo aspecto está relacionado à modelagem dessas influências, buscando, por meio de técnicas de Estatística Multivariada, compreender como essas variáveis afetam a formação de preços dos Óleos Básicos ao longo do tempo, dada sua menor volatilidade em comparação a outros derivados do petróleo. Para isso, foi criado um banco de dados com 227 variáveis, e as análises foram conduzidas tanto para identificar as mais relevantes na determinação do preço dos diferentes grupos Óleo Básico quanto para excluir aquelas que não contribuem significativamente para a precisão do modelo preditivo.

As variáveis foram selecionadas com base em estudos que identificaram fatores que afetam o preço do petróleo. Embora não existam pesquisas específicas sobre o preço dos Óleos Básicos, as inter-relações conhecidas entre o petróleo e seus derivados serviram como base para essa seleção. Foram analisados 31 Óleos Básicos da empresa A, cujo nome será omitido por questões de confidencialidade.

3 REFERENCIAL TEÓRICO

De modo geral, este capítulo tem como objetivo apresentar os fundamentos teóricos que sustentam a pesquisa. Inicialmente, são discutidos os principais aspectos do mercado de Óleos Lubrificantes Básicos e estudos relacionados ao mercado de petróleo. Em seguida, abordam-se conceitos de Estatística Multivariada, incluindo Análise de Agrupamentos e Regressão Múltipla. Por fim, exploram-se os principais métodos de séries temporais, como a metodologia Box-Jenkins, com foco na modelagem e previsão de preços.

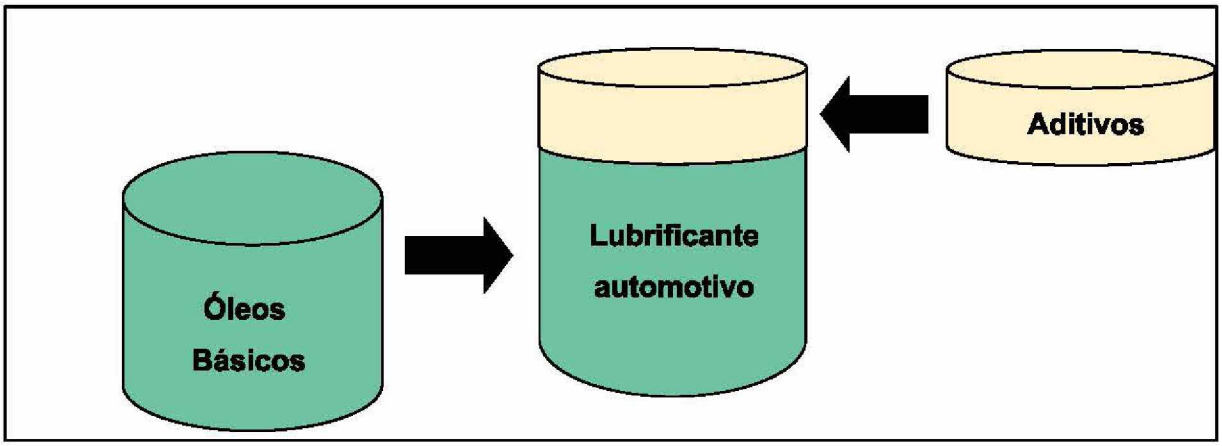
3.1 Óleo Lubrificante

Óleos Lubrificantes são substâncias derivadas do petróleo utilizadas para reduzir o atrito, o desgaste e controlar a temperatura entre superfícies em movimento, aumentando a eficiência e a durabilidade de máquinas e motores. Esses óleos são obtidos pelo processo de refino do petróleo, que gera diversos subprodutos, dentre os quais se destacam os Óleos Básicos (CAVALCANTI, 2014). Sua composição é formada principalmente por hidrocarbonetos parafínicos, naftênicos e aromáticos, que influenciam suas propriedades e desempenho como lubrificantes.

Segundo Almeida (2011), os lubrificantes possuem uma característica em comum: são todos formados por um óleo lubrificante básico, que pode receber aditivos. Além de sua função na lubrificação, esses óleos também desempenham um papel importante na eficiência energética das máquinas, influenciando diretamente o consumo de combustível e o desempenho geral (YANG, 2016).

A formulação de óleos lubrificantes para diferentes aplicações exige a combinação adequada de Óleos Básicos, obtidos por meio do processo de refino. Tipicamente, os lubrificantes são compostos por até 95% de Óleo Básico e o restante de aditivos, sendo o Óleo Básico o principal componente de sua formulação (FIGURA 1) (SAINI *et al.*, 2020). A adição de aditivos tem como objetivo atender a requisitos específicos de desempenho, como proteção contra oxidação, corrosão e desgaste.

FIGURA 1 - COMPOSIÇÃO TÍPICA DO ÓLEO LUBRIFICANTE



FONTE: Adaptado de Lima (2016).

3.1.1 Óleos Lubrificantes Básicos

O Óleo Lubrificante Básico é a principal matéria-prima na formulação de lubrificantes. Esses Óleos podem ser obtidos a partir do refino do petróleo ou por meio de processos sintéticos, sendo classificados, portanto, em Óleos Básicos minerais ou sintéticos. A escolha entre esses tipos depende das exigências da aplicação e das propriedades desejadas no lubrificante final (ALMEIDA, 2011).

Os Óleos Básicos minerais, derivados diretamente do petróleo, são amplamente utilizados na indústria devido ao seu custo reduzido, versatilidade e maior disponibilidade. Além disso, apresentam a vantagem de serem mais facilmente recicláveis, tornando-se uma opção viável para diversas aplicações. No entanto, possuem limitações em relação à estabilidade térmica e à resistência à oxidação quando comparados aos óleos sintéticos. No Brasil, a grande maioria dos Óleos Lubrificantes Básicos consumidos é de origem mineral, reforçando sua importância no mercado nacional (LIMA, 2016).

Por outro lado, os Óleos Básicos sintéticos, geralmente produzidos através de reações químicas a partir de produtos extraídos do petróleo, apresentam vantagens significativas, como maior estabilidade térmica, menor volatilidade e melhor desempenho em temperaturas extremas, o que os torna ideais para aplicações de alto desempenho, como motores de alta rotação e sistemas operando sob condições severas. Contudo, apesar desses benefícios, seu custo mais elevado limita sua utilização em larga escala, sendo geralmente empregados em aplicações específicas que exigem características superiores de lubrificação e resistência à degradação (ALMEIDA, 2011).

Independentemente da origem, os Óleos Lubrificantes Básicos são considerados matérias-primas nobres, representando apenas uma pequena fração do petróleo. O avanço da tecnologia de refino tem permitido melhorias na qualidade dos Óleos Básicos minerais, reduzindo sua presença de impurezas e ampliando sua eficiência em determinadas aplicações.

De acordo com Almeida (2011), os Óleos Básicos lubrificantes são classificados em naftênicos e parafínicos. Os de origem naftênica possuem baixo ponto de fluidez, menor índice de viscosidade e alto poder de solvência, sendo indicados para lubrificação em baixas temperaturas. Já os parafínicos apresentam elevado índice de viscosidade, alto ponto de fluidez e menor poder de solvência, sendo recomendados para lubrificantes de motores a combustão, sistemas hidráulicos e engrenagens operando em condições severas. A TABELA 1 apresenta as principais características dos Óleos Básicos parafínicos e naftênicos.

TABELA 1 – COMPARAÇÃO DAS CARACTERÍSTICAS DOS ÓLEOS BÁSICOS MINERAIS PARAFÍNICOS E NAFTÊNICOS.

Propriedades	**Parafínicos**	**Naftênicos**
Índice de viscosidade	Alto	Baixo
Ponto de fluidez	Alto	Baixo
Volatilidade	Baixa	Alto
Resistencia a oxidação	Boa	Média
Resíduo de carbono	Alto	Baixo

FONTE: Belmiro e Carreteiro (2006).

3.1.2 Classificação dos Óleos Lubrificantes Básicos

A necessidade de padronização na indústria de lubrificantes levou à adoção de um sistema de classificação de Óleos Básicos, desenvolvido pela API (*American Petroleum Institute*) nos Estados Unidos e pela ATIEL (*Association Technique de L'Industrie Européenne des Lubrifiants*) na Europa. Esse sistema organiza os óleos em cinco grupos (I a V), com base em três parâmetros principais: teor de saturados, teor de enxofre e índice de viscosidade (IV). A padronização permite que refinarias ao redor do mundo produzam óleos com propriedades consistentes, adequados a diversas aplicações industriais e automotivas (REVISTA LUBES EM FOCO, 2010).

Essa classificação auxilia na compreensão das propriedades e do desempenho dos óleos na formulação de lubrificantes. A TABELA 2 apresenta os critérios estabelecidos pela Agência

Nacional do Petróleo (ANP, 2022) com base nos parâmetros já citados, auxiliando na seleção do produto mais adequado a cada aplicação.

TABELA 2 – CLASSIFICAÇÃO DOS ÓLEOS BÁSICOS

Categoria	**Índice de Viscosidade**	**Saturados(%)**	**Enxofre(%)**
Grupo I	80 a 120	< 90	> 0,03
Grupo II	80 a 120	≥ 90	≤ 0,03
Grupo III	≥ 120	≥ 90	≤ 0,03
Grupo IV	Polialfaolefinas (PAOs)		
Grupo V	Demais óleos		

Fonte: Resolução ANP 911 (2022).

Diante dessas características, os óleos parafínicos apresentam propriedades técnicas compatíveis com aplicações industriais relevantes e ampla disponibilidade de mercado.

3.2 VARIÁVEIS QUE IMPACTAM O PREÇO DO PETRÓLEO

Pesquisas sobre os fatores que impactam o preço do petróleo têm sido amplamente exploradas. Como os Óleos Básicos minerais são derivados do petróleo, identificar as variáveis que influenciam seu preço permite avaliar se esses mesmos fatores afetam seus derivados. Para isso, a seguir, são apresentadas algumas pesquisas relevantes sobre o tema.

3.2.1 Trabalhos Correlatos

Vo *et al.* (2019) analisaram a relação entre os preços do petróleo bruto e das *commodities* agrícolas no período de 2000 a 2018, utilizando um modelo vetorial autorregressivo estrutural (*SVAR*). O estudo revelou que os impactos do petróleo sobre os preços agrícolas variam conforme o tipo de choque considerado, destacando a influência dos choques de demanda agregada e das dinâmicas do mercado de energia na explicação da volatilidade das *commodities*. Em particular, os autores identificaram que choques de demanda agregada, ao refletirem mudanças na atividade econômica global, impactam positivamente tanto o mercado de petróleo quanto o de *commodities* agrícolas. Além disso, a expansão da produção de biocombustiveis fortalece a correlação entre os mercados de energia e agrícolas.

Yang *et al.* (2022) investigaram a relação não linear entre o preço do petróleo bruto (*West Texas Intermediate* – WTI), a taxa de câmbio do dólar americano e a produção de petróleo WTI, por meio da estimativa de um modelo autorregressivo de transição suave com variáveis exógenas. Foi utilizado dados semanais de 1983 a 2021, sendo identificado que a variação do preço do WTI influencia a produção de forma não linear, com mudanças ocorrendo em diferentes limiares conforme a variável analisada. Constatou-se ainda que, diante de variações no Índice do Dólar Americano, a relação entre a taxa de variação do preço do WTI e sua capacidade de produção pode ser inversa no curto prazo, mas tende a se alinhar na mesma direção após um período de defasagem.

Kayalar *et al.* (2017) examinaram a relação entre os preços do petróleo bruto, os índices do mercado de ações e as taxas de câmbio em diferentes economias, levando em consideração sua classificação como mercados emergentes ou desenvolvidos, bem como sua condição de importadores ou exportadores de petróleo. No estudo, foi utilizado modelos *copula*, ARIMA e GARCH para analisar a estrutura de dependência entre essas variáveis ao longo do tempo, incluindo os impactos da crise financeira de 2008 e as variações de dependência em janelas de 1 a 30 dias. Os resultados indicaram que os índices de mercado dos países exportadores de petróleo apresentam maior dependência dos preços do petróleo, enquanto mercados emergentes importadores são menos sensíveis a essas oscilações.

Huang *et al.* (2022) avaliaram a previsão dos preços do petróleo a partir de variáveis macroeconômicas e financeiras. O estudo analisou fatores como tendências de mercado, política monetária, especulação e condições econômicas globais. Para aprimorar a precisão das previsões, os autores utilizaram variáveis exógenas em intervalos mensais. Entre as variáveis analisadas, destacam-se os índices S&P 500 (INX) e *Dow Jones Industrial Average* (DJI), que afetam fluxos de caixa e taxas de juros reais; os preços futuros do ouro da COMEX e do cobre da LME, utilizados como indicadores econômicos; e o *spread* entre os preços do *West Texas Intermediate* (WTI) e Brent, que mede a influência tecnológica. Também foram consideradas variáveis como a taxa dos fundos federais, o índice real do dólar americano, a busca pelo termo "preço do petróleo" no *Google Trends* e a relação líquida de posições longas não comerciais. Os resultados sugerem que a inclusão dessas variáveis melhora significativamente a previsão dos preços do petróleo, capturando com maior precisão as variações do mercado.

Outros estudos também exploraram variáveis determinantes da precificação do petróleo. Gunarto *et al.* (2020) destacaram a importância de variáveis macroeconômicas e financeiras, como gastos militares, oferta e demanda no mercado, PIB, atividade no mercado de capitais e taxa de câmbio. De forma semelhante, Safari e Davallou (2022) identificaram, além da oferta e

demanda, que fatores como taxa de câmbio do dólar americano, mudanças políticas e desastres naturais influenciam na determinação dos preços do petróleo.

Além disso, Zhao *et al.* (2021) desenvolveram um modelo de previsão em tempo real utilizando dados dos preços de fechamento dos contratos futuros de WTI, Brent e da *Shanghai International Energy Exchange* (INE), bolsa chinesa que negocia contratos futuros de petróleo. Analisou-se diferentes frequências temporais: mensal para WTI, semanal para Brent e diária para INE para avaliar a eficácia do modelo proposto. Os autores aplicaram uma abordagem baseada na utilização dos valores transformados para uma base logarítmica, permitindo melhor análise das variações nos preços do petróleo.

Desta forma, os estudos revisados indicam que os preços do petróleo são influenciados por uma ampla gama de variáveis, incluindo fatores macroeconômicos, condições financeiras, eventos geopolíticos e características do próprio mercado de petróleo. Essa complexidade justifica a crescente adoção de modelos preditivos mais robustos ou que incluam modelos estatísticos combinados com outras técnicas, para melhorar a previsão e a gestão do risco no setor industrial.

3.3 ANÁLISE DE DADOS

Segundo Ayub (2016), a análise de dados consiste em examinar os dados com o objetivo de identificar padrões ocultos, correlações inesperadas e informações úteis que auxiliem na tomada de decisões ou no desenvolvimento de soluções mais eficazes. Contudo, um dos maiores desafios é a obtenção de modelos adequados utilizando grandes volumes de dados (HASHEM et. al. 2015).

As informações geradas por inúmeros processos organizacionais são constantemente registradas em arquivos de dados, como aqueles relacionados a preços, compras, vendas e produção. No entanto, nem todos esses dados são imediatamente utilizáveis ou confiáveis. Nesse cenário, a análise de dados atua na verificação, tratamento e transformação dessas informações para que possam ser utilizadas de forma adequada na tomada de decisões.

Conforme apontam Shao *et al.* (2015), a análise de dados permite identificar padrões, tendências, ineficiências e riscos com base em registros passados, informações em tempo real e projeções. No contexto atual, onde os dados são gerados em grande volume, variedade, velocidade e com diferentes níveis de veracidade, torna-se necessário aplicar técnicas de análise capazes de lidar com essas dimensões, viabilizando o processamento, a interpretação e o uso adequado das informações na tomada de decisão.

Desse modo, é necessário adotar abordagens estruturadas para extrair valor dos dados disponíveis. Assim, com base nas definições apresentadas por Gartner (2014) e aprofundadas por Shao *et al* (2015), a análise de dados pode ser dividida em três métodos principais: análise descritiva, análise preditiva e análise prescritiva. A análise descritiva busca identificar o que aconteceu ou está acontecendo. Envolve a apresentação de dados coletados por meio de tabelas, gráficos, relatórios e painéis, permitindo visualizar padrões e tendências com base em informações históricas ou em tempo real. Já a análise preditiva está voltada para estimar o que é provável que ocorra. Utiliza técnicas como mineração de dados, modelos estatísticos e simulações para prever comportamentos futuros a partir de diferentes cenários e variáveis. Por fim, a análise prescritiva tem como foco indicar quais ações devem ser tomadas para atingir determinados objetivos, avaliando as consequências de diferentes alternativas de decisão. Assim, esse tipo de análise faz uso de simulações e procedimentos de otimização para propor soluções com base nas previsões geradas (SHAO *et al*, 2014). Essas três abordagens oferecem diferentes níveis de suporte à decisão e podem ser aplicadas de forma complementar na interpretação e no aproveitamento dos dados.

3.4 MÉTODOS ESTATÍSTICOS

A estatística pode ser compreendida como a ciência voltada à coleta, organização e interpretação de dados, sendo frequentemente utilizada para explorar correlações, relações causais e padrões entre variáveis (PÉBAY *et al.*, 2011). Sua aplicação tem se tornado cada vez mais necessária diante do crescimento no volume e na complexidade dos dados.

Ayub (2019) afirma que as técnicas estatísticas tradicionais não são as mais adequadas para lidar com grandes volumes de dados. Assim, conforme discutido por Pébay *et al.* (2011), para tratar conjuntos de dados em larga escala, o processo estatístico pode ser estruturado em quatro etapas: aprender o modelo (*Learn*), derivar estatísticas a partir do modelo (*Derive*), avaliar observações com base no modelo (*Assess*) e testar hipóteses (*Test*).

Nesse contexto, os métodos estatísticos permitem desde descrições iniciais até análises mais aprofundadas, seja com uma ou muitas variáveis, incluindo técnicas como estatísticas descritivas, Análise das Componentes Principais (PCA), Análise de Agrupamento (*Cluster Analysis*) e Regressão Linear múltipla. Sendo assim, destaca-se a importância de utilizar técnicas multivariadas para lidar com múltiplas variáveis simultaneamente, visto que tais

técnicas se mostram mais eficientes, ao oferecer diferentes métodos que auxiliam na interpretação e análise dos dados.

3.5 ANÁLISE MULTIVARIADA

A aplicação de métodos estatísticos no monitoramento de processos industriais tem sido historicamente fundamentada no Controle Estatístico de Processos (CEP), com gráficos como os de Shewhart, CUSUM e EWMA sendo comumente empregados para detectar variações com causas atribuíveis. Todavia, com o avanço das tecnologias de automação e aquisição de dados, evidencia-se que esses métodos não são mais suficientes para lidar com a complexidade dos processos industriais modernos, nos quais centenas de variáveis de processo são monitoradas simultaneamente em tempo real.

Segundo MacGregor e Kourti (1995), os métodos tradicionais de Controle Estatístico de Processos (CEP) tratam cada variável de forma isolada, ignorando as interdependências e correlações existentes entre elas, o que dificulta a interpretação e o diagnóstico de falhas ou desvios no processo. Essa limitação torna-se ainda mais crítica ao se considerar que apenas alguns eventos subjacentes são, de fato, responsáveis pelas variações observadas nas medições, as quais, na realidade, refletem simultaneamente esses mesmos eventos. Ou seja, quando as relações entre variáveis não são identificadas, efeitos ocultos dificultam a compreensão do fenômeno com base nas variáveis analisadas (VICINI, 2005).

Além disso, os métodos estatísticos voltados à análise de variáveis costumam ser classificados em dois grupos: o univariado, que considera variáveis de forma isolada, e o multivariado, que as analisa de maneira conjunta. O primeiro apresenta limitações importantes, especialmente quando o fenômeno em estudo depende de múltiplas variáveis. Nesses casos, não basta conhecer informações estatísticas isoladas, mas é necessário também conhecer a totalidade destas informações fornecida pelo conjunto das variáveis (BRAULIO, 2005).

Nesse contexto, MacGregor e Kourti (1995) destacam que os métodos multivariados são capazes de tratar os dados de forma integrada. Essas abordagens permitem identificar padrões direcionais de variação, reduzir os efeitos de ruído e aumentar a sensibilidade do monitoramento, possibilitando a medição, interpretação e previsão do grau de relacionamento das variáveis (AYUB, 2019). Em outras palavras, nessas técnicas é realizado um cruzamento entre variáveis dependentes e independente, ou ainda um cruzamento de dados envolvendo informações de várias questões de ordem dependente.

De acordo com Johnson e Wichern (1998), a Análise Multivariada pode ser aplicada em diferentes contextos, tais como:

- Redução ou simplificação de dados;
- Identificação de agrupamentos e padrões de distribuição;
- Investigação da dependência entre variáveis;
- Predição de valores futuros;
- Realização de testes de hipóteses, entre outras possibilidades.

Além disso, conforme destacado por Hair Jr. *et al.* (2009), com o aumento significativo da quantidade de dados disponíveis nas organizações, impulsionado pelo avanço da automação, de sensores e das tecnologias de conectividade, o desafio deixou de ser a obtenção de dados e passou a ser a transformação desses dados em conhecimento útil para a tomada de decisão. Desse modo, Chaves Neto (2025) afirma que a Análise Multivariada emprega técnicas que consideram todas as variáveis simultaneamente, exigindo que o objetivo da pesquisa seja previamente definido para que a técnica mais adequada possa ser selecionada.

Segundo Kendall (1980), a Análise Multivariada compreende diversas técnicas, que podem ser agrupadas em dois grandes conjuntos: i) as técnicas de avaliação da interdependência, que estudam as relações entre conjuntos de variáveis, como análise de agrupamento, Componentes Principais, Correlações Canônicas e Análise Fatorial; e ii) as técnicas de avaliação da dependência, voltadas à análise da influência de uma ou mais variáveis sobre as demais, como a Regressão Múltipla e a análise discriminante. Logo, os principais objetivos da Análise Multivariada segundo Pla (1986) são:

- Transformar variáveis interdependentes em um conjunto independente ou de menor dimensão.
- Agrupar observações ou variáveis com base em características semelhantes, permitindo identificar padrões ou estruturas ocultas nos dados.
- Estudar a relação entre variáveis - de independência total à colinearidade ou outras formas funcionais.

Segundo Gouvêa, Prearo e Romeiro (2012) ao utilizar técnicas estatísticas multivariadas, é necessário considerar algumas premissas que, ao desconsiderá-las, podem influenciar de forma negativa os resultadas da análise. Desta forma, as pressuposições das técnicas multivariadas podem ser entendidas como sendo todos os cuidados referentes a

obtenção e a avaliação dos dados, de modo a atender as exigências estatísticas necessárias de cada técnica (HAESBAERT, 2016). Tais premissas podem ser consideradas obrigatórias, embora, dependendo do objetivo da análise, algumas delas possam ser flexibilizadas.

3.5.1 Multicolinearidade dos dados

Conforme a definição exposta por Hair Jr. *et al.* (2009), a multicolinearidade representa o grau em que o efeito de uma variável pode ser previsto ou explicado pelas demais variáveis da análise. Desta forma, a alta multicolinearidade, também denominada colinearidade exata, pode reduzir a habilidade de identificar os efeitos individuais. Em outras palavras, quando há variáveis inter-relacionadas o processo de verificação dos seus efeitos torna-se mais complexo, uma vez que se torna mais difícil verificar o efeito de cada uma delas.

De acordo com Gujarati (2011), é importante que, em um modelo de análise de dados, sejam incluídas variáveis que não sejam funções lineares exatas de uma ou mais variáveis do modelo. No entanto, o autor ressalta que, na prática, ao coletar dados, torna-se quase impossível encontrar duas ou mais variáveis que não apresentem algum grau de correlação entre si. Assim, em determinados casos, essa condição pode ser flexibilizada. Um estudo aprofundado sobre multicolinearidade pode ser obtido em Haesbaert (2016).

3.5.2 *Outliers*

Em análise de dados, seja univariada ou multivariada, um dos processos importantes é a detecção de observações atípicas ou extremas (*outliers*). Como exemplo, Hair Jr. *et al.* (2009) cita o caso em que se deseja determinar a renda média familiar de um grupo com 20 indivíduos, cujos valores variam entre R$ 20.000,00 e R$ 100.000,00 por ano, resultando em uma média de R$ 45.000,00. No entanto, ao incluir um 21° indivíduo com renda de R$ 1 milhão, a média sobe para R$ 90.000,00. Nesse cenário, os autores reforçam que casos atípicos devem ser analisados com cuidado, pois podem influenciar de forma significativa os resultados da análise.

Mingoti (2005), recomenda que no caso de valores atípicos nos dados, o ideal é a exclusão da variável analisada. Entretanto, tais observações podem ser tanto benéficas quanto problemáticas, sendo necessário que o pesquisador avalie criteriosamente a conveniência de mantê-las ou eliminá-las do estudo (HAIR JR. *et al.*, 2009).

3.5.3 Ausência de erros correlacionados

Compreende-se por erros correlacionados, segundo Kendall e Buckland (1971), a correlação entre amostras observadas no tempo ou no espaço. Quando há correlação entre os erros, não é possível afirmar que eles são independentes, comprometendo a validade das inferências nos níveis de significância utilizados para prever a variável dependente (GOUVÊA, PREARO e ROMEIRO, 2012).

Conforme aponta Haesbaert (2016), a correlação entre os erros ocorre quando fatores não incluídos no modelo afetam os resultados. Por isso, é desejável que os erros de previsão sejam independentes entre si, ou seja, que o erro associado a uma observação não influencie o erro de outra. Assim, segundo Hair Jr. *et al.* (2009), quando os grupos são analisados separadamente, os efeitos permanecem constantes dentro de cada grupo, não impactando, portanto, a estimação da relação. Em contrapartida, quando combinadas as observações entre dois ou mais grupos, os resultados tendem a ser viesados. Para esses casos, de acordo com os autores, alguns testes podem ser utilizados para verificar a correlação entre os erros, sendo eles: teste de Breush-Godfrey, teste M de Durbin, teste de Geary (ou teste das carreiras) e teste de Durbin-Watson, o mais utilizado.

3.5.4 Homoscedasticidade

A homoscedasticidade refere-se à suposição de que as variáveis apresentam níveis iguais de variância ao longo do domínio, ou seja, que as variâncias sejam homogêneas no conjunto de variáveis (HAESBAERT, 2016). De acordo com Hair Jr. *et al.* (2009), essa suposição é desejável porque a variância da variável dependente não deve se concentrar apenas em um intervalo restrito dos valores das variáveis independentes. Para que a relação entre as variáveis seja completamente capturada, a dispersão dos valores da variável dependente deve ser relativamente constante ao longo de todo o conjunto das variáveis preditoras. Quando essa variância é desigual, tem-se a violação da homoscedasticidade, caracterizando uma relação heteroscedástica.

A homoscedasticidade pode ser verificada, inicialmente, por meio da análise gráfica dos resíduos. Isso pode ser feito ao se comparar os erros com os valores reais e previstos, observando a distribuição dos pontos no gráfico. Quando esses pontos estão distribuídos de forma aleatória, sem apresentar um padrão definido, compreende-se que há homoscedasticidade

(GOUVÊA, PREARO e ROMEIRO, 2012). Um teste que avalia a dependência das variáveis conjuntamente é o teste de M de Box multivariado (HAIR JR *et al.*, 2009).

3.5.5 Linearidade

Em técnicas multivariadas, tais como Regressão Múltipla, regressão logística, Análise Fatorial e modelagem de equações estruturais, é necessário avaliar a linearidade. Essa necessidade decorre do fato de a correlação quantificar somente a relação linear entre as variáveis, não identificando os efeitos não lineares (HAESBAERT, 2016).

Conforme Gouvêa, Prearo e Romeiro (2012), uma maneira prática utilizada para verificar essa premissa é por meio da análise visual de diagramas de dispersão (*scatterplots*), que permitem observar a existência de padrões ou curvaturas na relação entre as variáveis. Essa verificação pode ser complementada com análises de correlação e testes estatísticos mais complexos. De forma geral, os modelos lineares assumem que a variável dependente apresenta uma variação constante em resposta a mudanças na variável independente, produzindo uma relação que se ajusta a uma linha reta. Essa suposição está implícita em técnicas baseadas em medidas correlacionais de associação (HAIR JR. *et al.*, 2009). No entanto, autores como Huberty (1994) e Eisenbeis (1997) reconhecem que essa exigência pode ser flexibilizada em situações nas quais a normalidade multivariada é atendida e a amostra possui tamanho suficientemente grande.

3.5.6 Normalidade

Denomina-se Normalidade ou Gaussianidade o grau em que a distribuição de probabilidade dos dados se aproxima de uma distribuição Normal. No contexto da Estatística Multivariada, utiliza-se o termo normalidade multivariada para se referir à combinação de duas ou mais variáveis com distribuição conjunta Normal, uma generalização da normalidade univariada, ou seja, $N_p(\mu, \Sigma)$, onde se lê Normal p-variada com vetor de médias μ e martriz de covariância Σ. Essa é considerada uma das condições mais fundamentais para a aplicação de diversas técnicas multivariadas (HAIR JR. *et al.*, 2009). Quando a variação em relação à distribuição normal é suficientemente grande, os testes estatísticos baseados nessa premissa, como as estatísticas F e t, tornam-se inválidos. Johnson e Wichern (1998) alertam, no entanto, que em dados empíricos é raro encontrar variáveis que apresentem normalidade multivariada exata.

3.5.7 Padronização

A padronização refere-se ao processo no qual a variável é transformada em uma nova variável com uma média de 0 e um desvio padrão igual a 1 (HAIR JR. *et al.*, 2009). Segundo Regazzi (2000) é conveniente realizar essa padronização quando as escalas das unidades de medidas das características observadas não forem as mesmas. No caso da Análise Fatorial, a padronização não é uma exigência, porém pode ser aplicada previamente ao processamento da técnica, (HAIR JR. *et al.*, 2009).

3.5.8 Algumas definições de Análise Multivariada

Segundo Johnson e Wichern (1998), na Análise Multivariada considera-se um conjunto de p características observáveis de n indivíduos de uma população. Os valores dessas observações são atribuídos a cada item ou unidade amostral distinta, sendo representados pela notação X_{jk} que indica o valor da k-ésima variável observada no j-ésimo indivíduo (TABELA 3).

TABELA 3 – MATRIZ DE AMOSTRAS POR VARIÁVEL

Unidades amostrais ou experimentais	**Variáveis**			
	1	$2\cdots$	$k\cdots$	p
1	X_{11}	$X_{12}\cdots$	$X_{1k}\cdots$	X_{1p}
2	X_{21}	$X_{22}\cdots$	$X_{2k}\cdots$	X_{2p}
⋮	⋮	⋮	⋮	⋮
j	X_{j1}	$X_{j2}\cdots$	$X_{jk}\cdots$	X_{jp}
⋮	⋮	⋮	⋮	⋮
n	X_{n1}	$X_{n2}\cdots$	$X_{nk}\cdots$	X_{np}

FONTE: Adaptado de Johnson e Wichern (1998).

Essas características são representadas pelas variáveis aleatórias $X_1, X_1, X_3, \cdots, X_P$, que podem ser agrupadas no vetor aleatório $\underline{X} = [\underline{x_1}, \underline{x_2}, \underline{x_3}, \cdots, \underline{x_n}]$. A partir das observações empíricas desse vetor para cada indivíduo ou unidade amostral, obtém-se uma matriz de dados

X de ordem $n \times p$, ${}_pX_n$, em que cada linha representa um indivíduo (ou unidade experimental), e cada coluna corresponde a uma variável observada.

$$ {}_pX_n = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \quad (1) $$

Conforme destaca Bráulio (2005), as informações amostrais contidas nas observações multivariadas, representadas por $[\underline{x_1}, \underline{x_2}, \underline{x_3}, \cdots, \underline{x_n}]$, podem ser resumidas por meio de estatísticas descritivas, que servem como base para a inferência estatística. Essas estatísticas descritivas estimam os vários parâmetros, entre os quais destacam-se o vetor médio $\underline{\mu}$, a matriz de covariância Σ, que é estimada pela matriz de correlação amostral S e a matriz de correlação ρ, estimada por R. O vetor médio populacional deve ser estimado pelo vetor amostral $\underline{\overline{X}}$ que é dado por:

$$ \underline{\overline{X}} = \frac{\sum_{i=1}^{n} \underline{x_i}}{n} \quad (2) $$

sendo $\underline{x_i}$, com $i = 1, 2, \cdots, n$ o vetor correspondente as observações amostrais do vetor $\underline{X}$ e n o tamanho da amostra observada. Adicionalmente, outra estatística descritiva citada foi a matriz covariância amostral S, cuja diagonal principal é formada pelas variâncias amostrais das variáveis aleatórias que formam o vetor observado, enquanto os elementos fora da diagonal principal representam as covariâncias amostrais entre elas, conforme apresentado em (3).

$$ S = \frac{\sum_{i=1}^{n} (\underline{x_i} - \underline{\overline{x}})(\underline{x_i} - \underline{\overline{x}})^T}{n-1} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix} \quad (3) $$

onde s_j^2 é a variância amostral da variável aleatória X_j, obtida por:

$$ s_j^2 = \frac{\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2}{n-1} \quad (4) $$

De acordo com Regazzi (2000) é conveniente padronizar as variáveis X_j $(j = 1,2,3,\cdots,p)$ quando as escalas das unidades de medida das características observadas não forem homogêneas (ver Subseção 3.7.1.7). Por fim, tem-se a matriz de correlação R, definida por:

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \quad (5)$$

3.5.9 Técnicas de Análise Multivariada

A Análise Multivariada apresenta diversas técnicas, cada uma adequada a análise e solução de problemas específicos. Segundo Lattin *et al.* (2011), para a aplicação dos métodos multivariados, é necessário considerar a natureza dos diferentes tipos de dados. Assim, o autor destaca três características que orientam as escolhas das técnicas:

- Se a análise busca dependência ou interdependência entre os dados;
- Se o objetivo é explorar ou confirmar informações; e
- Se os dados são métricos ou não métricos.

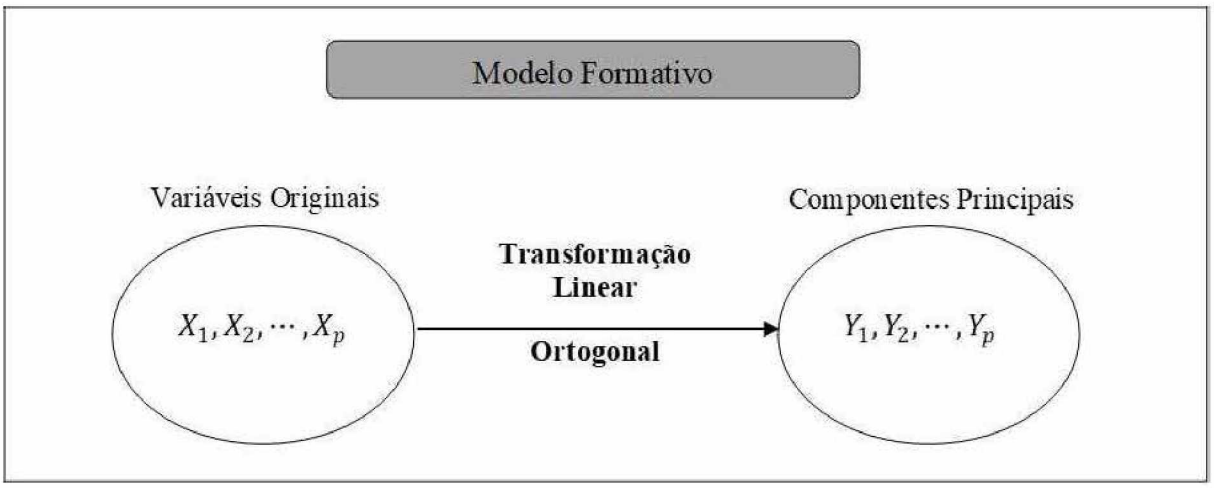
3.5.9.1 Análise das Componentes Principais – PCA

A Análise das Componentes Principais (PCA) é uma técnica Estatística Multivariada que visa obter um novo conjunto de variáveis INDEPENDENTES e reduzir a dimensionalidade de um conjunto de dados, ao mesmo tempo em que preserva o máximo possível da variabilidade presente nas p variáveis originais, ou seja, mantém praticamente o mesmo nível de informação do conjunto original de variáveis, pois se descartam apenas as combinações lineares das variáveis originais que não são significativamente importantes a critério do experimentador. As componentes principais são obtidas por autovalores que determinam as combinações lineares das variáveis originais e os autovalores que indicam a importância, pelo grau de explicação, das combinações lineares obtidas.

Estas Componentes Principais são ordenadas de forma decrescente de importância, por meio das combinações lineares das variáveis originais, de maneira que o primeiro Componentes Principal seja o que possuam a máxima variância, e assim sucessivamente (AYUB, 2019). Por

isso, como afirmado por Bajotto (2025), tem-se um modelo formativo. Este processo é representado na FIGURA 2.

FIGURA 2 - COMPONENTES PRINCIPAIS



Fonte: Adaptado de Marques (2025).

A transformação ortogonal que gera as Componentes Principais pode ser vista como uma operação que gera um novo conjunto de dados, ou um novo conjunto de coordenadas, que são perpendiculares entre si (STEFFEN, 2021).

De acordo com Nascimento e Guardani (2007) os objetivos da Análise das Componentes Principais são:

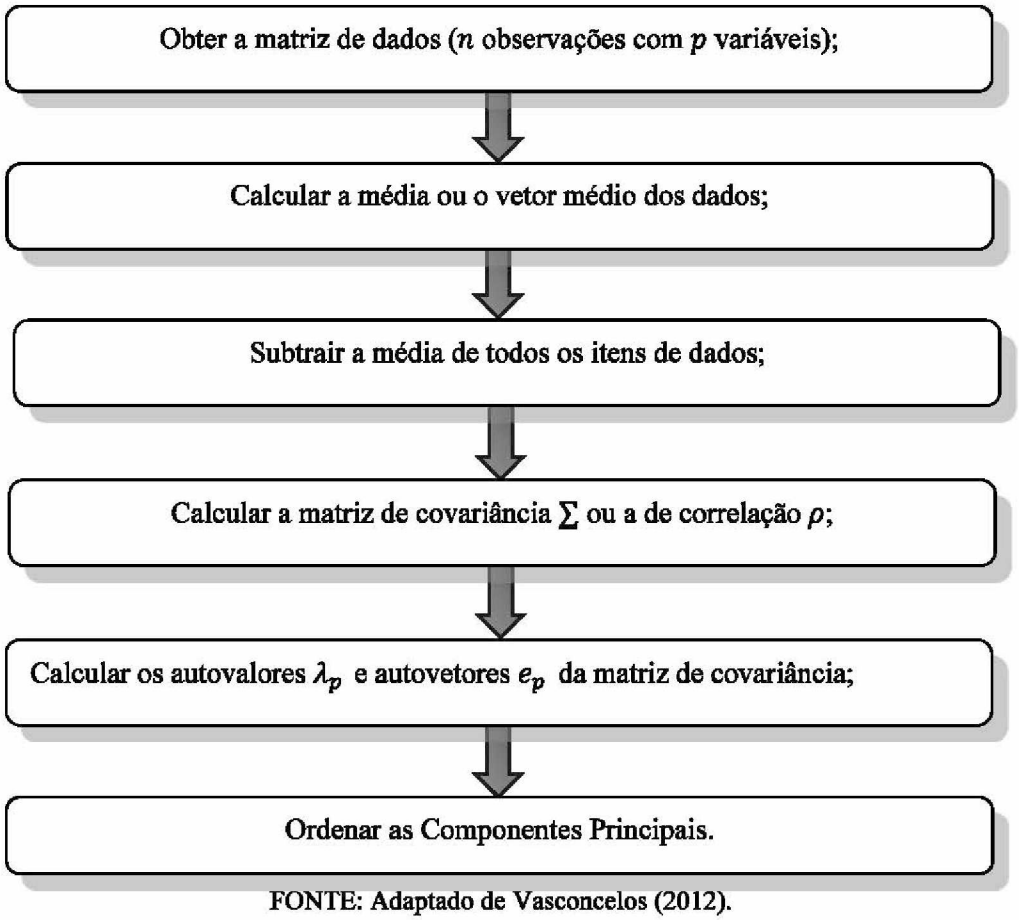
1. Fornecer pela transformação linear ortogonal NOVAS VARIÁVEIS QUE SÃO combinações lineares independentes.
2. Reduzir o número de novas variáveis, que são combinações lineares das variáveis originais, ou as que correspondem aos menores autovalores e por consequência menos informativas. Possibilitando, deste modo, a criação de modelos mais simples para simulação e análise, preservando a maior parte da variabilidade dos dados.
3. Identificar, pela correlação dentro de cada combinação linear, relações entre variáveis originais ou grupos de variáveis originais, e fornecer uma estrutura mais simples para efeito de análise e interpretação nos dados.

Embora seja necessária as p componentes para explicar a variabilidade da covariância do vetor $\underline{X}$ de dimensão p, em muitos casos a grande parte desta variabilidade pode ser explicada por um número k de Componentes Principais. Neste caso, entende-se que sendo $k <$

p, existe tanta informação nos componentes k quanto nas variáveis p, podendo ser substituídas pelos Componentes Principais k (JOHNSON e WICHERN, 1998; AYUB, 2019).

Na Análise das Componentes Principais, os componentes não são correlacionados entre si, o que satisfaz o pressuposto de ausência de multicolinearidade, uma vez que esta é eliminada no processo. Desta forma, Vasconcelos (2012) apresenta os passos necessários para a obtenção das Componentes Principais, os quais estão elencados na FIGURA 3.

FIGURA 3 - PASSOS DO CÁLCULO – PCA



FONTE: Adaptado de Vasconcelos (2012).

Assim, dadas a matriz de covariância S ou a matriz de correlação R, as Componentes Principais são obtidas da decomposição espectral dessas matrizes.

Os autovalores λ_i são determinados pela solução da equação característica

$$|S - \lambda I| = 0 \tag{6}$$

e os autovetores $\underline{e_i}$ são obtidos a partir da equação

$$Se_i = \lambda_i e_i. \quad S_i = 1, 2, \ldots p. \tag{7}$$

Essa equação permite encontrar os autovalores da matriz S, que representa as variâncias explicadas por cada Componentes Principal. Quando a matriz S é de posto completo p, a equação (6) possui autovalores reais e positivos, ordenados de forma decrescente, ou seja, $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$.

Além disso, é importante destacar que, quando os dados são padronizados (com média zero e desvio padrão igual a um), a análise das Componentes Principais é realizada com base na matriz de correlação R, uma vez que a matriz de covariância dos dados padronizados equivale à matriz de correlação das variáveis originais.

Assim, a cada autovalor λ_i está associado um autovetor e_i, o qual define a direção da Componentes Principal correspondente. Esses autovetores formam um conjunto ortonormal, ou seja, são mutuamente ortogonais e de norma unitária, o que garante a eliminação da multicolinearidade entre as Componentes Principais. Desta forma, as Componentes Principais são dadas pelas seguintes combinações lineares:

$$\begin{aligned} Y_1 &= e_{11}X_1 + e_{12}X_2 + e_{13}X_3 + \cdots + e_{1p}X_p \\ Y_2 &= e_{21}X_1 + e_{22}X_2 + e_{23}X_3 + \cdots + e_{2p}X_p \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ Y_i &= e_{i1}X_1 + e_{i2}X_2 + e_{i3}X_3 + \cdots + e_{ip}X_p \end{aligned} \tag{8}$$

Matricialmente, tem-se:

$$Y_i = \begin{bmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_p \end{bmatrix} \begin{bmatrix} e_{11}\ e_{12} \ldots e_{1p} \\ e_{21}\ e_{22} \ldots e_{2p} \\ \ldots\ldots \\ e_{p1}\ e_{p2} \ldots e_{pp} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \ldots\ldots \\ X_p \end{bmatrix} \tag{9}$$

Como visto, a quantidade de informação contida nas $k, k < p$, primeiras Componentes Principais é praticamente equivalente àquela presente nas p variáveis originais, pois essas Componentes estão associadas às direções de maior variabilidade dos dados. Em outras palavras, o número de Componentes Principais Y_i pode chegar até p, se desejado, conforme ilustrado a seguir:

$$Y_i = e_{p1}X_1 + e_{p2}X_2 + e_{p3}X_3 + \cdots + e_{pp}X_p \quad (10)$$

No entanto, embora seja possível utilizar todas as p componentes, quando um número reduzido de Componentes Principais k, com $k < p$, é capaz de reter entre 80% e 90% da variabilidade total do conjunto de dados, pode-se considerar que esses componentes explicam adequadamente a estrutura das variáveis originais, sendo possível utilizar apenas as primeiras componentes (JOHNSON e WICHERN, 1998).

3.5.9.2 Análise Fatorial - AF

A Análise Fatorial é uma abordagem estatística utilizada para investigar as inter-relações entre um grande número de variáveis observadas, com o objetivo de explicá-las por meio de um conjunto reduzido de variáveis latentes (não observáveis diretamente) denominadas fatores (HAIR JR *et al.*, 2009).

De acordo com Morrison (1976), cada variável observada pode ser expressa como uma combinação linear de alguns fatores comuns e de um termo específico, exclusivo daquela variável. Os Fatores comuns explicam as covariâncias entre as variáveis observáveis, enquanto os termos específicos contribuem apenas para suas variâncias individuais.

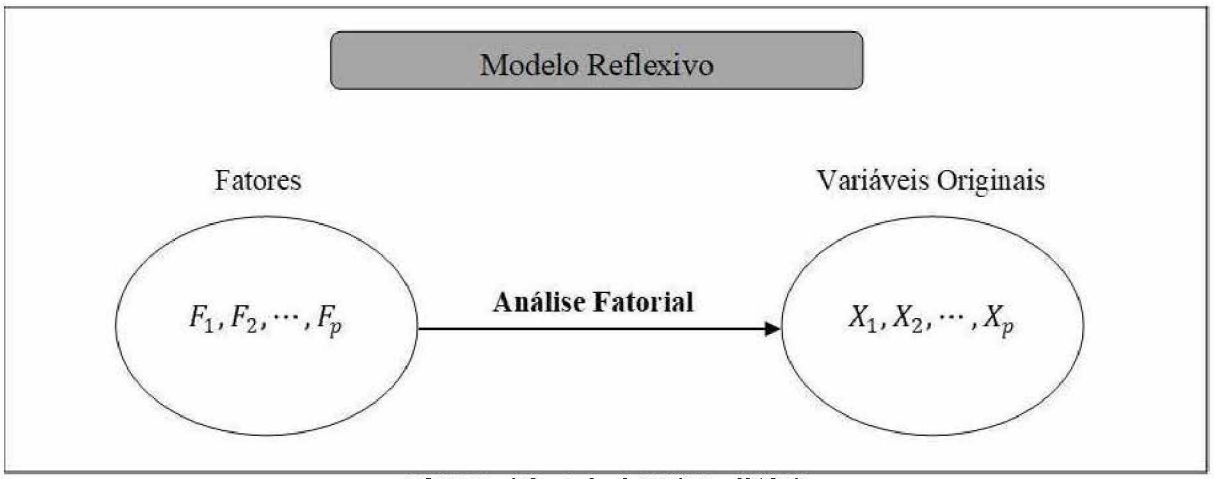
Desta forma, os objetivos da Análise Fatorial, segundo Gontijo e Aguirre (1988), estão elencados a seguir:

- Condensar um grande número de observações em grupos;
- Obter um número reduzido de variáveis em comparação ao original sem grande perda de informação;
- Obter os fatores que reproduzem um padrão separado de relações entre as variáveis;
- Interpretar de forma lógica o padrão de relações entre as variáveis;
- Identificar as variáveis apropriadas para análises posteriores, tais como: Análise posteriores: Regressão, Correlação ou Discriminante, entre outros procedimentos estatísticos.

Kubrusly (1981) destaca que a Análise Fatorial tem como objetivo principal reproduzir, da melhor forma possível, as correlações entre as variáveis originais, diferente da Análise das Componentes Principais, que busca explicar a variância total dos dados. Esta concepção reforça

o caráter reflexivo do modelo fatorial, uma vez que as variáveis latentes são consideradas reflexos de fatores subjacentes, conforme ilustrado na FIGURA 4.

FIGURA 4 - ANÁLISE FATORIAL



FONTE: Adaptado de Bajotto (2025).

Desse modo, seja o vetor $\underline{X}$, composto por variáveis aleatórias, representado como uma função linear de um menor número ($m < p$) de variáveis hipotéticas, denominadas fatores comuns, e considerando a matriz de covariância $\sum$, tem-se que cada variável no modelo fatorial é dada por:

$$\underline{X_p} = \sum_{j=1}^{m} \ell_{ij} F_j + \varepsilon_i, \qquad i = 1,2,\cdots,p \tag{11}$$

onde F_j é o j-ésimo Fator Comum, sendo que o número m de fatores é especificado e ε_i é um resíduo representando fontes de variação que afetam apenas a variável X_p. Os coeficientes ℓ_{ij} são cargas fatoriais (*loadings*) e refletem a importância do j-ésimo fator na composição da i-ésima variável X_p (FACHEL, 1976).

Desta forma, Jonhson e Wichern (1998) apresenta o modelo fatorial, para cada $i = 1,2,\cdots,p$, com $m < p$, como:

$$\begin{aligned} X_1 - \mu_1 &= \ell_{11} F_1 + \ell_{12} F_2 + \ldots + \ell_{1m} F_m + \varepsilon_1 \\ X_2 - \mu_2 &= \ell_{21} F_1 + \ell_{22} F_2 + \ldots + \ell_{2m} F_m + \varepsilon_2 \\ &\ldots\ldots\ldots\ldots \\ X_i - \mu_i &= \ell_{i1} F_1 + \ell_{i2} F_2 + \ldots + \ell_{im} F_m + \varepsilon_i \\ &\ldots\ldots\ldots\ldots \end{aligned} \tag{12}$$

$$X_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p$$

e na forma matricial, tem-se:

$$\underset{px1}{\underline{X}} - \underset{}{\underline{\mu}} = \underset{pxm}{L}\ \underset{mx1}{\underline{F}} + \underset{px1}{\underline{\varepsilon}} \tag{13}$$

onde tem-se para $i = 1, 2, \ldots, p$ a variável representativa do modelo fatorial que é dada por $X_i = \Sigma\, ijF_j + \varepsilon_i, i = 1,2,\cdots,p$.

Segundo Fachel (1976), supõe-se que os p termos de erro ε_i são não correlacionados entre si e também não apresentam correlação com os fatores comuns. Além disso, assume-se que os fatores F_j são ortogonais, ou seja, não correlacionados entre si. A variância de ε_i, denominada Variâncias Específicas, é representada por ψ_i.

Considerando $\underline{F}$ o vetor formado pelos fatores e $\underline{\varepsilon_i}$ o vetor dos resíduos, e assumindo que:

$$E(\underline{F}) = \underset{mx1}{\underline{0}}, V(\underline{F}) = E(\underline{FF'}) = I_m \tag{14}$$

$$E(\underline{\varepsilon}) = \underset{px1}{\underline{0}}, V(\underline{\varepsilon}) = E(\underline{\varepsilon\varepsilon'}) = \underset{pxp}{\Psi} = \begin{bmatrix} \Psi_1 & 0 & 0 & \ldots & 0 \\ 0 & \Psi_2 & 0 & \ldots & 0 \\ 0 & 0 & \Psi_3 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \Psi_p \end{bmatrix} \tag{15}$$

e que $\underline{F}$ e $\underline{\varepsilon}$ são independentes, tem-se $cov(\underline{\varepsilon}, \underline{F}) = E(\underline{\varepsilon}, \underline{F}) = 0$ com $m = p$. Assim, sob estas condições, tem-se o Modelo Fatorial Ortogonal apresentado em (12) e a matriz de covariância de $\underline{X}$, dada por:

$$\Sigma = cov(\underline{X}) = E\left(\underline{X} - \underline{\mu}\right)\left(\underline{X} - \underline{\mu}\right)' = E\left[\underline{L}\underline{F}(\underline{L}\underline{F})' + \underline{\varepsilon}(\underline{L}\underline{F})' + \underline{L}\underline{F}\underline{\varepsilon}' + \underline{\varepsilon}\underline{\varepsilon}'\right]$$

$$\Sigma = LE(\underline{F}\underline{F}')L' + E(\underline{\varepsilon}\underline{F}')L' + LE(\underline{F}\underline{\varepsilon}') + E(\underline{\varepsilon}\underline{\varepsilon}') = \underset{pxp}{LL'} + \underset{pxp}{0} + \underset{pxp}{0} + \underset{pxp}{\Psi}\ \Sigma \tag{16}$$

$$\Sigma = \underset{pxp}{LL'} + \underset{pxp}{\Psi}$$

A partir desta decomposição, a variância da variável X_i, é dada por:

$$V(X_i) = \sum_{j=1}^{m} \ell_{ij}^2 + \psi_i, \qquad i = 1,2,\cdots,p$$

$$V(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2 + \psi_i \tag{17}$$

A matriz produto LL', portanto, tem na diagonal principal as comunalidades representadas por:

$$h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \cdots + \ell_{im}^2, \qquad para\ i = 1,2,\ldots,p \tag{18}$$

sendo ℓ_{ij} a carga fatorial da variável X_i no fator F_j, com p fatores comuns extraídos.

A covariância entre duas variáveis originais e entre a variável original X_i e o Fator F_j tem, respectivamente, as seguintes expressões:

$$cov(X_i X_k) = \ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} + \cdots + \ell_{im}\ell_{km} \tag{19}$$

$$cov(X_i F_j) = \ell_{ij} \tag{20}$$

Segundo Steffen (2021), um bom modelo fatorial deve apresentar comunalidades elevadas para cada variável, o que indica que a maior parte da variância está sendo explicada pelos fatores comuns. Em contrapartida, uma variância específica baixa sugere que a variável tem relação significativa com o fenômeno estudado. Variáveis com baixa comunalidade e alta variância específica contribuem pouco para o modelo e, portanto, podem ser descartadas da análise.

A estimação dos parâmetros do modelo Fatorial Ortogonal, ou seja, $\underset{pxm}{L}$, é pode ser feita usando-se o Método da Máxima Verossimilhança. Esse método exige o conhecimento da distribuição multivariada de probabilidade do vetor observado, daí a maior dificuldade. Porém, a forma usual de estimar esses parâmetros é por meio da solução baseada em Componentes Principais, em que as cargas fatoriais são obtidas pela multiplicação da raiz quadrada dos autovalores pelos autovetores correspondentes (JOHNSON e WICHERN, 1998).

Considerando a solução por componentes principais partindo-se da matriz S ou R que fornece os pares de autovalores/autovetores $(\hat{\lambda}_1, \underline{\hat{e}}_1)$, $(\hat{\lambda}_2, \underline{\hat{e}}_2)$, ... , $(\hat{\lambda}_p, \underline{\hat{e}}_p)$ onde $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p > 0$ tem-se a matriz de carregamentos (pesos, *loads*)

$$\underset{pxm}{\hat{L}} = \left[\sqrt{\hat{\lambda}_1}\,\hat{e}_1 \quad \sqrt{\hat{\lambda}_2}\,\hat{e}_2 \quad \ldots \quad \sqrt{\hat{\lambda}_m}\,\hat{e}_m\right] = \begin{bmatrix} \sqrt{\hat{\lambda}_1}\hat{e}_{11} & \sqrt{\hat{\lambda}_2}\hat{e}_{12} & \ldots & \sqrt{\hat{\lambda}_m}\hat{e}_{1m} \\ \sqrt{\hat{\lambda}_1}\hat{e}_{21} & \sqrt{\hat{\lambda}_2}\hat{e}_{22} & \ldots & \sqrt{\hat{\lambda}_m}\hat{e}_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ \sqrt{\hat{\lambda}_1}\hat{e}_{p1} & \sqrt{\hat{\lambda}_2}\hat{e}_{p2} & \ldots & \sqrt{\hat{\lambda}_m}\hat{e}_{pm} \end{bmatrix}.$$

A matriz das variâncias específicas é: $\underset{pxp}{\hat{\Psi}} = \begin{bmatrix} \hat{\psi}_1 & 0 & \ldots & 0 \\ 0 & \hat{\psi}_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \hat{\psi}_p \end{bmatrix}$ com $\hat{\psi}_{ii} = s_i^2 - \sum_{j=1}^{m} \hat{\ell}_{ij}^2$

e onde as comunalidades estimadas são $\hat{h}_i^2 = \hat{\ell}_{i1}^2 + \hat{\ell}_{i2}^2 + \ldots + \hat{\ell}_{im}^2 = \sum_{j=1}^{m} \hat{\ell}_{ij}^2$ e interpreta-se esses resultados como:

- a contribuição do 1°. fator p/ a variância s_1^2 da v.a. X_i é $\hat{\ell}_{i1}^2$.
- a contribuição do 1°. fator p/ a variância total $s_1^2 + s_2^2 + \ldots + s_p^2 = tr(S)$ é $\sum_{i=1}^{m} \hat{\ell}_{i1}^2$.

3.5.9.2.1 Escores Fatoriais

De acordo com Fachel (1976), é conveniente, além de estimar os parâmetros do modelo fatorial, buscar descrever os fatores em termos das variáveis observadas. Nesse sentido, para corresponder a aproximação dos valores das variáveis não observáveis, estimam-se os valores de cada fator para cada indivíduo, o que se denomina escore fatorial. Esses escores representam, para cada unidade amostral i, os valores dos m fatores estimados com base em suas combinações lineares com as variáveis observadas. Assim, uma observação $\underline{x_i}$, originalmente expressa como um vetor de dimensão p, pode ser representada por um vetor reduzido de Escores Fatoriais $\hat{f}_i = [\hat{f}_1, \hat{f}_2, \ldots, \hat{f}_m]$, com m < p.

Além disso, é possível calcular um escore fatorial agregado (ESC_I), que resume os fatores estimados em um único valor por meio de uma média ponderada dos Escores Fatoriais, cujos pesos são definidos pelos autovalores λ_j associados a cada fator comum extraído. Assim, os Escores Fatoriais para as n observações da Matriz de Dados de ordem $n \times p$ têm por expressão:

$$ESC_I = \frac{\sum_{j=1}^{m} \lambda_j f_{ij}}{\sum_{j=1}^{m} \lambda_j} \; i = 1, 2, \ldots, n \qquad (21)$$

Devido à complexidade da estimação dos Escores Fatoriais, Mingoti (2005), destaca os Métodos dos Mínimos Quadrados (MMQ) e de Regressão para determinar os Escores Fatoriais. Para este estudo, será abordado apenas o MMQ.

3.5.9.2.2 Método dos Mínimos Quadrados

Bartlett foi o primeiro a sugerir o uso do Métodos dos Mínimos Quadrados Ponderados para estimar os Escores Fatoriais (JOHNSON e WICHERN, 1998). No entanto, a aplicação desse método requer que o vetor médio $\underline{\mu}$, a matriz de cargas fatoriais L e a matriz de variâncias específicas Ψ sejam conhecidos no modelo $\underline{X} - \underline{\mu} = L\underline{F} + \underline{\varepsilon}$. A partir disso, escolhe-se as estimativas $\underline{\hat{f}_i}$ de $\underline{f}$ para minimizar a soma de quadrados dos erros, ponderada pela recíproca da variância de cada variável, ou seja:

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\psi_i} = \underline{\varepsilon}'\psi^{-1}\underline{\varepsilon} = \left(\underline{X} - \underline{\mu} - L\underline{F}\right)' \psi^{-1} \left(\underline{X} - \underline{\mu} - L\underline{F}\right) \tag{22}$$

e, ao utilizar as estimativas $\hat{L}, \hat{\Psi}$ de $\mu = \underline{\overline{x}}$, como verdadeiros, obtém-se o j-ésimo escore fatorial:

$$\underline{\hat{f}_j} = (\hat{L}\,\hat{\psi}^{-1}\hat{L})^{-1}\hat{L}'\hat{\psi}^{-1}\left(\underline{x}_j - \underline{\overline{x}}\right)\ j = 1, 2, \ldots\ldots, n \tag{23}$$

Se a análise é feita a partir da matriz de correlação R, tem-se:

$$\underline{\hat{f}_j} = (\hat{L}_z'\hat{\psi}_z^{-1}\hat{L}_z)^{-1}\hat{L}_z'\hat{\psi}_z^{-1}\underline{z}_j\ j = 1, 2, \ldots\ldots, n \tag{24}$$

Mas, quando se usa Componentes Principais para estimar as cargas fatoriais é usual estimar os Escores Fatoriais usando os Mínimos Quadrados Ordinários (MQO). Desta forma, as Variâncias Específicas ψ_1 são consideradas como iguais ou como aproximadamente iguais e os Escores Fatoriais estimados são:

$$\underline{\hat{f}_j} = (\hat{L}'\hat{L})^{-1}\hat{L}'\left(\underline{x}_j - \underline{\overline{x}}\right)\ j = 1, 2, \ldots\ldots, n \tag{25}$$

E, ainda, se a análise é feita a partir de matriz de correlação R tem-se:

$$\hat{\underline{f}}_j = (\hat{L}_z'\hat{L}_z)^{-1}\hat{L}_z'\underline{z}_j \quad j = 1, 2, \ldots\ldots, n \tag{26}$$

3.5.9.2.3 Estimação do Número de Fatores

De acordo com Steffen (2021), uma das formas de estimar o número adequado de fatores a serem retidos na Análise Fatorial é por meio do critério da raiz latente, também conhecido como critério de Kaiser. Esse critério sugere a retenção apenas dos fatores com autovalores superiores a 1, considerados estatisticamente significativos. No entanto, nem todo autovalor superior a um tem necessariamente significado interpretativo claro. Da mesma forma, alguns autovalores ligeiramente inferiores a 1 podem ainda representar uma porção relevante da variabilidade total e, portanto, podem ser considerados na análise. Assim, a determinação do número de fatores deve também considerar a proporção da variância explicada por cada fator, que é dada por.

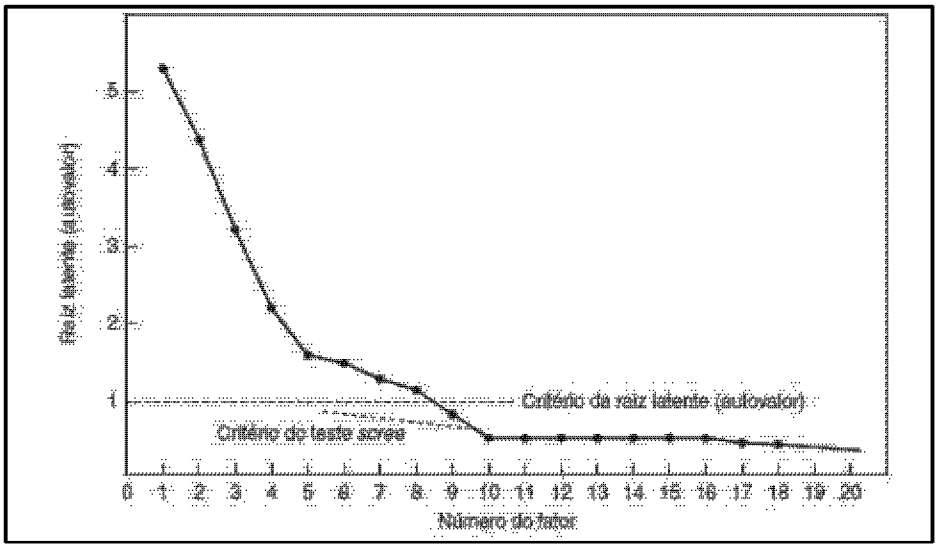
$$\frac{\hat{\lambda}_j}{s_{11}+s_{22}+\cdots+s_{pp}}, \text{ para a análise a partir de } S \tag{27}$$

$$\frac{\hat{\lambda}_j}{p}, \text{ para a análise a partir de } R \tag{28}$$

Além disso, o teste Scree Plot é utilizado para identificar o número ótimo de fatores a serem extraídos, antes que a variância específica comece a dominar a estrutura de variância comum. Esse critério é aplicado por meio de um gráfico dos autovalores (raízes latentes) em função do número de fatores, em ordem decrescente de extração. A curvatura da linha resultante permite visualizar um ponto de inflexão, o chamado "cotovelo", que indica o número ideal de fatores a ser retido (HAIR JR. *et al.*, 2009).

A FIGURA 5 ilustra um exemplo em que, segundo o critério de Kaiser, seriam selecionados oito fatores, enquanto pelo Scree Plot, a escolha seriam dez fatores.

FIGURA 5 - GRÁFICO SCREE PLOT



Fonte: Hair JR *et al.* (2009).

3.5.9.2.4 Critério Varimax

O Critério Varimax é um método de rotação ortogonal que tem como objetivo facilitar a interpretação dos fatores extraídos. A rotação redistribui as cargas fatoriais de modo que cada variável apresente cargas elevadas em apenas um ou poucos fatores, e cargas próximas de zero nos demais (HAIR JR. *et al.*, 2009).

Kaiser definiu o critério Varimax para cada fator j como sendo:

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}(\tilde{e}_{ij}^{*2})^2 - \frac{1}{p^2}\left(\sum_{i=1}^{p}\tilde{e}_{ij}^{*2}\right)^2\right], \quad i = 1,2,\cdots,p \text{ e } j = 1,2,\cdots,m. \tag{29}$$

Define-se o Critério Varimax $\tilde{e}_{ij}^{*2} = \frac{\tilde{e}_{ij}^{2}}{\hat{h}_{ij}^{2}}$, em que:

$\tilde{e}_{ij}^{2}$ é a carga fatorial da variável i no fator j;

$\tilde{e}_{ij}^{*2}$ representa a proporção da comunalidade de X_i explicada pelo fator j;

$\hat{h}_{ij}^{2}$ é a comunalidade total da variável i pelo fator j;

Assim, de acordo com Chaves Neto (2025), para se obter a rotação é necessário aplicar uma transformação T a ser determinada, ou seja, $\underline{\hat{f}}_j^* = T'\underline{\hat{f}}_j$, $j = 1,2,\cdots,n$, em que a carga fatorial $\tilde{e}_{ij}^{*}$ são multiplicados por $\hat{h}_{ij}$ de modo que as comunalidades originais sejam preservadas.

3.5.9.2.5 Adequação da Análise Fatorial

Como discutido na Subseção 3.7.1, algumas técnicas de Análise Multivariada requerem a verificação de certos pressupostos para serem adequadamente aplicadas. No caso da Análise Fatorial, é necessário assumir a normalidade dos dados, para facilitar a interpretação e a determinação da importância dos fatores. Além disso, é importante que exista multicolinearidade, ou seja, correlação significativa entre as variáveis, para que os fatores possam ser identificados. Para avaliar a adequação dos dados para a Análise Fatorial, dois testes são comumente utilizados: o teste de Kaiser-Meyer-Olkin (KMO) e o teste de Esfericidade de Bartlett.

3.5.9.2.5.1 Teste de Kaiser-Mayer-Olkin (KMO)

De acordo com Steffen (2021) critério de Kaiser-Meyer-Olkin (KMO) identifica se o modelo de Análise Fatorial será adequadamente ajustado aos dados. Assim, considerando a inversa da matriz de correlação amostral, $[R_{(p\,x\,p)}]^{-1}$ esta medida é dada por:

$$KMO = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} q_{ij}^2} \tag{30}$$

onde:

- r_{ij}^2: é o quadrado do elemento pertencente a i-ésima linha e j-ésima coluna da matriz de correlação amostral $R_{(pxp)}$, sendo que i, j = 1, 2, ..., p.
- q_{ij}^2: é o quadrado do elemento pertencente a i-ésima linha e j-ésima coluna da matriz $Q=DR^{-1}D$. onde

$$D = \left[\sqrt{diag(R_{(p\,x\,p)}^{-1})}\right]^{-1} \tag{31}$$

O teste KMO varia de 0 a 1 e tem como objetivo avaliar a adequação do modelo adequação da amostra de dados quanto ao grau de correlação parcial entre as variáveis (STEFFEN, 2021). Para valores de KMO abaixo de 0,6 é desaconselhável ajustar um Modelo de Análise Fatorial para os dados (CHAVES NETO, 2025). Sharma (1996) classifica o KMO conforme o quadro a seguir:

TABELA 3 – CLASSIFICAÇÃO DA ESTATÍSTICA KMO

KMO	Classificação
≥ 0,9	Ótimo
de 0,8 a 0,9	Bom
de 0,7 a 0,8	Razoável
de 0,6 a 0,7	Baixo
< 0,6	Inadequado

FONTE: Sharma (1996).

3.5.9.2.5.2 Esfericidade de Bartlett

De acordo com Hair Jr. *et al.* (2009), o teste de esfericidade de Bartlett é uma técnica estatística utilizada para verificar a presença de correlações significativas em uma matriz de correlação, essencial para a aplicação da Análise Fatorial. Este teste avalia a hipótese nula, H_o, de que a matriz de correlação populacional, ρ, do vetor aleatório observado é uma matriz identidade $I_{(p \times p)}$, ou seja, que não há correlação entre as variáveis analisadas. A hipótese alternativa, H_a, assume que a matriz de correlação é diferente da identidade, indicando a presença de correlações significativas.

Quando o teste de Bartlett apresenta um resultado estatisticamente significativo, rejeita-se a hipótese nula, confirmando que existem correlações suficientes para prosseguir com a Análise Fatorial. Assim, seja:

$$H_o: \rho = I_{(p \times p)} \times H_a: \rho \neq I_{(p \times p)} \tag{32}$$

A estatística de Bartlett é definida por

$$T = -\left[n - \frac{1}{6}(2p + 11)\right] \sum_{i=1}^{p} \ln \widehat{\lambda_i} \tag{33}$$

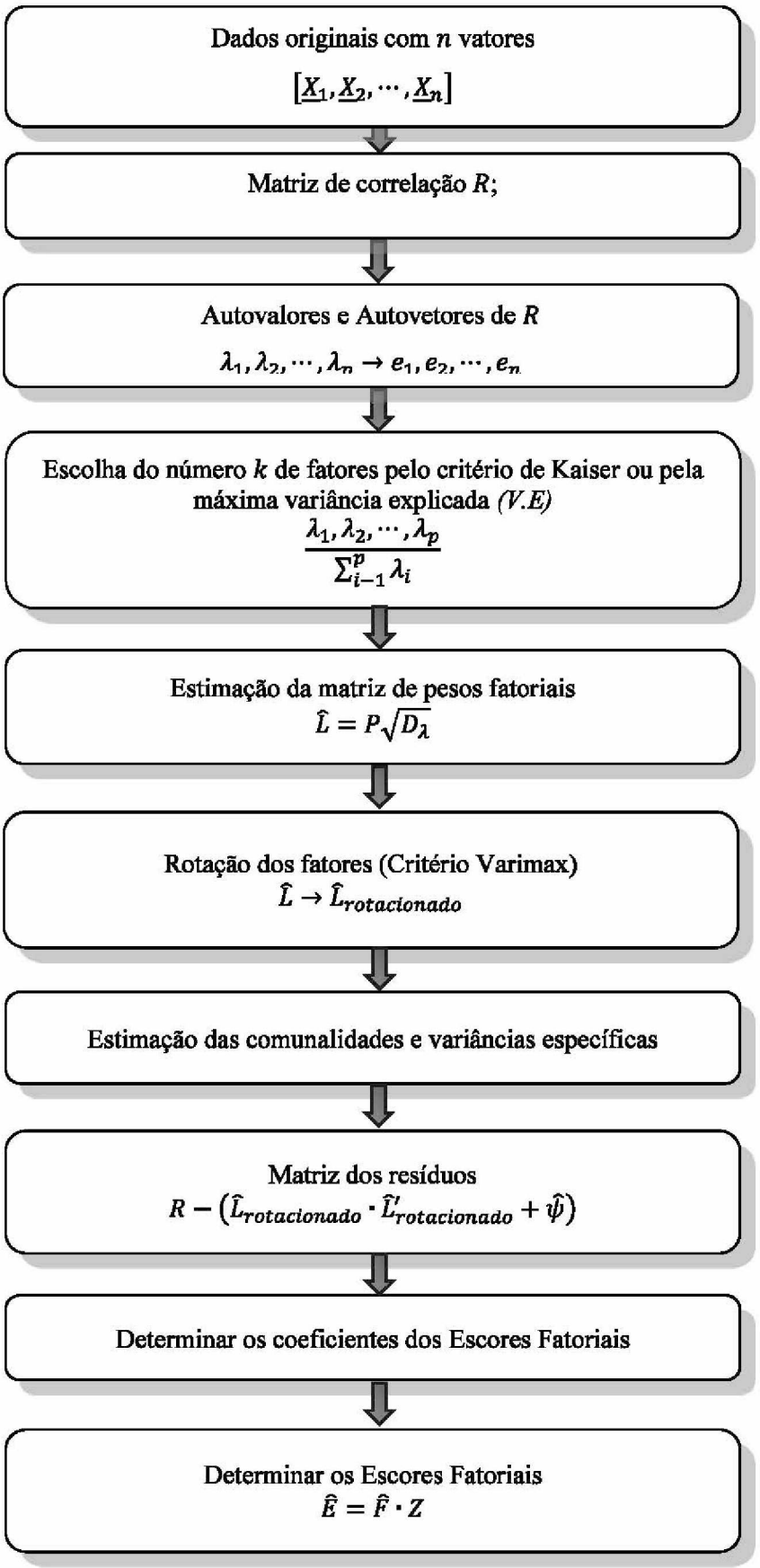
onde:

- n: é o tamanho da amostra.
- p: é o número de variáveis utilizadas na matriz de correlação amostral R, ou seja, a ordem da matriz.
- $\widehat{\lambda_i}$: é o i-ésimo autovalor da matriz de correlação amostral $R_{(p \times p)}$.

Então, sob H_0 e um tamanho de amostra n grande a estatística T tem aproximadamente uma distribuição Qui-quadrado com $\upsilon = \frac{1}{2}p(p-1)$ graus de liberdade, ou seja, $T \sim \chi^2_{\upsilon}$. Desta forma, para que o Modelo Fatorial seja aceitável, é necessário que o teste de Bartlett rejeite a hipótese de nulidade H_0, pois, se isso não ocorrer, não haverá correlações entre as variáveis originais a ser modelada pela Análise Fatorial (CHAVES NETO, 2025).

Desta forma, Steffen (2021) apresenta os passos necessários para a realização de uma Análise Fatorial, os quais estão elencados na FIGURA 6.

FIGURA 6 - PASSOS DO CÁLCULO – AF



Fonte: Steffen (2021).

3.5.9.3 Análise de Agrupamento (*Cluster Analysis*)

A Análise de Agrupamento é uma técnica multivariada que busca a formação de grupos homogêneos de objetos ou variáveis. Estes grupos são formados calculando-se as distâncias entre os itens, representados por vetores compostos pelas suas características, construindo-se uma matriz de distâncias e juntando os itens em grupos de acordo com suas proximidades.

De acordo Crivisqui (1993) os chamados Métodos de Agrupamento, ou *Cluster Analysis*, ou ainda Métodos de Classificação Automática, são métodos estatísticos destinados a dividir em subconjuntos um conjunto de dados observados. Aplicar estes métodos significa definir nesse conjunto as classes em que se distribuem os elementos do conjunto. Por isso, este método se diferencia do Reconhecimento de Padrões e Classificação (Análise Discriminante), em que o número de grupos é previamente conhecido e o objetivo é alocar novas observações nesses grupos, o agrupamento é mais primitivo e exploratório, pois não pressupõe conhecimento prévio sobre o número de grupos ou a estrutura dos dados. O agrupamento é feito com base na similaridade ou distância (BRAULIO, 2005).

Segundo Guerreiro (2021), devido a falta de conhecimento prévio do domínio, é difícil escolher um número apropriado de grupos (*clusters*), pois parte das técnicas de aprendizado de máquina utilizadas nestes casos são não supervisionadas, já que os dados não possuem rótulos. Isso se torna ainda mais complexo quando os dados têm muitas dimensões, tamanho e quando os grupos se diferem em forma. Para lidar com essa complexidade, é possível combinar técnicas de Análise Multivariada, como a Análise das Componentes Principais (PCA), para reduzir a dimensionalidade dos dados, preservando suas características essenciais, antes de aplicar métodos de agrupamento.

3.5.9.3.1 Medidas de similaridade e dissimilaridade

A Clusterização consiste em formar grupos de objetos sendo cada grupo composto por elementos mais similares entre si do que em relação aos elementos de grupos distintos. Esse processo de similaridade é geralmente medido usando distâncias, dependendo da natureza dos dados e do objetivo do agrupamento.

Por outro lado, quando o foco do agrupamento é em variáveis, como características ou atributos, os grupos são frequentemente formados com base em coeficientes de correlação ou

outras medidas de associação. Essas medidas avaliam a similaridade entre as variáveis, considerando tanto a direção quanto a força das suas relações.

Assim, quanto maior o valor do índice de similaridade (ou menor o valor do índice de dissimilaridade), mais semelhantes são os objetos, enquanto valores menores de similaridade (ou maiores de dissimilaridade) indicam maior diferença entre os elementos.

Existem vários índices de similaridades, sendo que a sua principal medida é o coeficiente de correlação. No entanto, para a análise de distância entre pontos tem-se a Distância Euclidiana, Distância de Manhattan ou Distância de Mahalanobis.

Para este estudo, optou-se pela Distância Euclidiana, que, para duas observações multivariadas de dimensão p, $\underline{x}' = [x_1, x_2, \ldots, x_p]$ e $\underline{y}' = [y_1, y_2, \ldots, y_p]$ é dada por:

$$d(\underline{x}, \underline{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} = \sqrt{(\underline{x} - \underline{y})'(\underline{x} - \underline{y})} = \|\underline{x} - \underline{y}\| \qquad (34)$$

A distância estatística entre as mesmas observações é da forma:

$$d(\underline{x}, \underline{y}) = \sqrt{(\underline{x} - \underline{y})' A^{-1} (\underline{x} - \underline{y})} \qquad (35)$$

onde A^{-1} é tal que $d(x,y) \geq 0$. Assim $(\underline{x} - \underline{y})' A^{-1} (\underline{x} - \underline{y})$ é uma forma quadrática e as entradas de A^{-1} são variâncias e covariâncias amostrais. Contudo, sem conhecimento dos grupos distintos, estas quantidades não podem ser calculadas e, então, a Distância Euclidiana é preferível na Análise de Agrupamento (BAJOTTO, 2025).

Segundo Bajotto (2025) os algoritmos de formação dos grupos, partem da matriz de dados de ordem $n \times p$, n vetores (itens) de dimensão p, e formam a matriz de distâncias de ordem $n \times n$ calculando as distâncias entre os n itens (vetores) usando uma medida de distância. E, então, juntam-se os itens (vetores) com menor distância obtendo uma nova matriz de distâncias de ordem inferior.

3.5.9.3.2 Método de Agrupamento

Figueiredo *et al.* (2019) abordam alguns métodos para a divisão dos *cluster*, sendo classificados como abordagens particionais, hierárquicas, sobrepostas e baseadas em grafos. No entanto, para este estudo, optou-se por utilizar métodos particionais, especificamente o

algoritmo *K-Means*, devido à sua simplicidade e eficiência para conjuntos de dados de alta dimensionalidade. Para a determinação do número ideal de clusters, foi utilizado o método do cotovelo (*elbow method*) em conjunto com a distância euclidiana como métrica de similaridade, além do índice de silhueta (*Silhouette Score*) para avaliar a qualidade dos agrupamentos gerados. Tais métodos são descritos a seguir:

3.5.9.3.3 Agrupamento particional

De acordo Figueiredo *et al.* (2019), o agrupamento particional organiza os dados em grupos com base em critérios de adequação, que afetam diretamente a estrutura dos clusters. Após a escolha de uma métrica adequada, o processo de particionamento se torna um problema de otimização, focado na minimização das distâncias ou maximização da correlação entre os padrões, visando otimizar a densidade no espaço dimensional.

O método de partição mais proeminente e popular é o algorítimo *K-Means*. Esse método é aplicado quando se deseja formar um número definido k de grupos. Sharma (1996) define o método como a sequência das seguintes etapas para a construção dos grupos.

1. Definir o número de grupos iniciais, definindo-se os centroides de cada grupo.
2. Designar cada observação/elemento para o centroide mais próximo de acordo com a Distância Euclidiana.
3. Recalcula-se o novo centroide do grupo que recebeu uma nova observação/elemento e o novo centroide para o grupo que perdeu aquela observação/elemento.
4. Repete-se o processo até que não há mais nenhuma recolocação de observações/elementos.

A seguir, tem-se um pseudocódigo com os passos necessários deste método, conforme apresentado por Guerreiro (2021).

Pseudocódigo 1: K-Means.

```
1  início
2      parâmetros de entrada: dados, quantidade de centroides e critério de parada
3      geram-se os centroides aleatoriamente dentro do espaço dos dados
4      alocam-se os dados com menor distância a cada centroide formando os grupos
5      para cada grupo formado faça
6          recalcula-se o centro geométrico de cada grupo, gerando um novo
           centroide
7          aloca-se os dados a cada centroide formando os grupos
8          verifica-se critério de parada
9      fim
10     retorna-se os centroides e grupos formados
   fim
```

Fonte: Guerreiro (2021).

3.5.9.3.4 Métrica de Agrupamento

Um dos processos mais importantes da Análise de Agrupamento é a definição de métricas de avaliação adequadas ao problema (GUERREIRO, 2021). Assim, o Índice de Silhueta (SI) (*Silhueta Score*) é uma métrica amplamente utilizada para avaliar a qualidade de agrupamentos. Esse índice varia de -1 a 1 e mede o quão bem uma observação está alocada em seu grupo, comparando a similaridade média com as outras observações do mesmo grupo e a similaridade média com as observações do grupo mais próximo. Valores próximos de -1 indicam que a observação pode ter sido incorretamente alocada, enquanto valores próximos de zero sugerem que o ponto é quase igualmente semelhante ao seu próprio grupo e a outros grupos (GUERREIRO, 2021; FIGUEIREDO *et al*, 2019).

Seja $\underline{X} = [\underline{x_1}, \underline{x_2}, \underline{x_3}, \cdots, \underline{x_n}]$ um conjunto de dados com n amostras. Suponha que as amostras em $\underline{X}$ têm rótulos rígidos que os marcam como representantes de k clusters sem sobreposição, ou $K = \{c_1, c_2, \cdots, c_k\}$ representando os centroides. O algoritmo de agrupamento busca encontrar a partição ideal $P = \{P_1, P_2, \cdots, P_m\}$, posicionando iterativamente os k centroides (ZHAO; XU; FRÄNTI, 2009; GUERREIRO, 2021). Assim, considere a_i dado por:

$$a_i = \frac{1}{n_k} \sum_{x_j \in C_k} dist(x_i, x_j) \tag{36}$$

e b_i dado por:

$$b_i = min_{h \in \{1,\cdots,K\},\ h \neq k} \left(\frac{1}{n_h} \sum_{x_j \in C_k} dist(x_i, x_j) \right) \tag{37}$$

Então a equação (34) apresenta a fórmula para o cálculo do SI:

$$SI = \frac{1}{n} \sum_{i=1}^{n} \frac{b_i - a_i}{max(b_i, a_i)} \tag{38}$$

O SI é um índice de maximização, $SI \in [-1, 1]$, em que a é a distância média de um ponto x_i aos demais pertencentes ao mesmo grupo c_k, b é a mínima distância média de um ponto x_i pertencente a algum outro grupo c_h (JOSÉ-GARCÍA; GÓMEZ-FLORES, 2016)(FIGUEIREDO *et al.*, 2019; GUERREIRO, 2021).

3.5.9.3.5 Curva do Cotovelo (*Elbow Curve*)

Um dos maiores desafios da Análise de Cluster consiste na estimação do número de grupos. Desta forma, para identificar os k cluster, um dos métodos mais conhecidos é o Elbow. Este método é semelhante ao apresentado na Subseção 3.5.9.2.3.

São testadas uma quantidade e depois plotada em um gráfico no qual se torna possível avaliar o ponto em que o aumento deste número traz um benefício pequeno para métricas (GUERREIRO, 2021). Este ponto apresentado no gráfico se assemelha a um cotovelo e, conforme destacado por Guerreiro (2021), significa que os resultados do teste, no final da execução, em algum momento o ganho marginal cairá drasticamente, o que resulta um ângulo no gráfico. Em outras palavras, é quando o ponto onde a redução da inércia (soma das distâncias quadráticas intra-*cluster*) deixa de ser significativa à medida que se aumenta o número de grupos,

3.5.9.4 Análise Discriminante

Fisher foi um dos pioneiros no estudo da Análise Discriminante, propondo esse método como um dos critérios mais confiáveis para a classificação de novas espécies de vegetais (MAROCO, 2003).

Johnson e Wichern (1998) apresentam a discriminação e a classificação como técnicas multivariadas que visam separar conjuntos de observações e alocar novos elementos em grupos previamente definidos. A análise discriminante, conforme proposta por R.A. Fisher, busca descrever características diferenciais de observações a partir de variáveis preditoras, identificando funções discriminantes que maximizam a separação entre grupos. Já os métodos de classificação definem regras para alocar novas observações a classes conhecidas, estabelecendo critérios para a atribuição de objetos com base em suas características observadas.

Embora os objetivos de discriminação e classificação possam se sobrepor, a discriminação concentra-se na identificação de características que melhor separam os grupos, enquanto a classificação busca criar regras para a alocação precisa de novos elementos.

Desta forma, os objetivos principais dessa técnica, de acordo com Johnson e Wintcher (1998), são:

- Descrever as características diferenciais de objetos de diversas populações, seja graficamente (em até três dimensões) ou algebricamente, de forma que os conjuntos sejam separados o máximo possível;
- Classificar objetos em duas ou mais classes rotuladas, desenvolvendo regras que permitam a alocação ideal de novos objetos com base em características prévias, buscando encontrar uma regra que possa ser usada na alocação ótima de um novoobjeto (observação) nas classes consideradas

3.5.9.4.1 Função Discriminante Linear de Fisher para Duas Populações

A ideia de Fisher foi transformar as observações multivariadas $\underline{X}'s$ em observações univariadas $\underline{Y}'s$ tal que os $\underline{Y}'s$ das populações π_1 e π_2 sejam separados tanto quanto possível (BRAULIO, 2005).

Assim, seja μ_{1y} a média dos $\underline{Y}'s$ obtidos dos $\underline{X}'s$ pertencentes a π_1 (população 1) e μ_{2y} a média dos $\underline{Y}'s$ obtidos dos $\underline{X}'s$ pertencentes a π_2 (população 2), então Fisher selecionou a

combinação linear que maximiza a distância quadrática entre μ_{1y} e μ_{2y} relativamente à variabilidade dos $\underline{Y}'s$ (ALVES, 2005). Logo, tem-se:

$$\underline{\mu}_{1y} = E(\underline{X}|\,\pi_1) = \text{valor esperado de uma observação multivariada } \pi_1. \tag{39}$$

$$\underline{\mu}_{2y} = E(\underline{X}|\,\pi_2) = \text{valor esperado de uma observação multivariada } \pi_2. \tag{40}$$

e considerando a matriz de covariância

$$\Sigma = E(\underline{X}|\,\pi_1)E(\underline{X}|\,\pi_2)', \qquad i = 1,2. \tag{41}$$

e

$$\mu_{1y} = E(\underline{Y}|\,\pi_1) = E(\underline{c}'\underline{X}|\,\pi_1) = \underline{c}'E(\underline{X}|\,\pi_1) = \underline{c}'\underline{\mu}_1 \tag{42}$$

$$\mu_{2y} = E(\underline{Y}|\,\pi_2) = E(\underline{c}'\underline{X}|\,\pi_2) = \underline{c}'E(\underline{X}|\,\pi_2) = \underline{c}'\underline{\mu}_2 \tag{43}$$

onde $\underline{Y}$ é dado pela combinação linear:

$$\underline{Y}_{1x1} = \underline{c}'_{1xp}\underline{X}_{px1} \tag{44}$$

A variância de $\underline{Y}$ é dada por:

$$V(\underline{Y}) = \sigma_y^2 = V(\underline{c}'\underline{X}) = \underline{c}'V(\underline{X})\underline{c} \tag{45}$$

Deste modo, ao calcular a derivada da razão entre o quadrado da distância entre as médias e a variância de $\underline{Y}$, obtém-se, segundo Fisher, a melhor combinação linear, sendo dada por:

$$\frac{(\mu_{1y} - \mu_{1y})^2}{\sigma_y^2} = \frac{\left(\underline{c}'\underline{\mu}_1 - \underline{c}'\underline{\mu}_2\right)^2}{\underline{c}'\Sigma\underline{c}} = \frac{\underline{c}'\left(\underline{\mu}_1 - \underline{\mu}_2\right)\left(\underline{\mu}_1 - \underline{\mu}_2\right)\underline{c}}{\underline{c}'\Sigma\underline{c}} \tag{46}$$

Assim, sendo $\underline{\delta} = \underline{\mu}_1 - \underline{\mu}_2$, então em (46), tem-se:

$$\frac{(\underline{c}'\delta)^2}{\underline{c}'\Sigma\,\underline{c}} \tag{47}$$

Dado $\underline{Y} = \underline{c}'\underline{X}$, a expressão (43) é maximizada por:

$$\underline{c}' = k\Sigma^{-1}\underline{\delta} = k\Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2), \text{ para todo } k \neq 0. \tag{48}$$

Já para $k = 1$, tem-se, então, a Função Discriminante Linear de Fisher, dada por:

$$\underline{c} = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \quad \text{e} \quad \underline{Y} = \underline{c}'\underline{X} = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\,\underline{X}. \tag{45}$$

Considerando $\underline{X}_0$ as medidas de um novo item a ser classificado, então em (45), obtém-se:

$$\underline{Y}_0 = (\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\,\underline{X}_0. \tag{49}$$

Seja o ponto médio dado por:

$$m = \frac{1}{2}(\mu_{1y} + \mu_{2y}), \tag{50}$$

então, substituindo (38), (39) e (45) na equação acima, obtém-se:

$$m = \frac{1}{2}\left[(\underline{\mu}_1 - \underline{\mu}_2)'\Sigma^{-1}\,(\underline{\mu}_1 - \underline{\mu}_2)\right] \tag{51}$$

Se $\underline{X}_0$ pertence a π_1, se espera que $\underline{Y}_0$ seja igual ou maior do que o ponto médio. Por outro lado, se $\underline{X}_0$ pertence a π_2, o valor esperado de $\underline{Y}_0$ será menor que o ponto médio (ALVES, 2005). Em outras palavras, tem-se:

$$E(\underline{Y}_0 | \pi_1) - m \geq 0 \tag{52}$$

e

$$E(\underline{Y}_0 | \pi_2) - m < 0 \tag{53}$$

Logo, a regra de classificação é:

$$\text{alocar } \underline{x}_0 \text{ em } \pi_1 \text{ se } \underline{x}_0 - \text{m} \geq 0$$
$$\text{alocar } \underline{x}_0 \text{ em } \pi_2 \text{ se } \underline{x}_0 - \text{m} < 0$$

Conforme apontado por Chaves Neto (2025), geralmente os parâmetros π_1, π_2 e $\sum$ são desconhecidos. Considerando n_1 observações da variável aleatória multivariada $\underline{X}_1$ de dimensão p, representando uma amostra aleatória da população π_1, e n_2 observações da variável aleatória multivariada $\underline{X}_2$ de dimensão p, que constituem uma amostra aleatória da população π_2, os resultados amostrais correspondentes podem ser descritos como:

$$\bar{\underline{x}}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \underline{x}_{i1}\,; \qquad S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} \left(\underline{x}_{i1} - \bar{\underline{x}}_1\right)\left(\underline{x}_{i1} - \bar{\underline{x}}_1\right)' \tag{54}$$

$$\bar{\underline{x}}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} \underline{x}_{i2}\,; \qquad S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2} \left(\underline{x}_{i2} - \bar{\underline{x}}_2\right)\left(\underline{x}_{i2} - \bar{\underline{x}}_2\right)' \tag{55}$$

E, uma vez que se assume que as populações sejam assemelhadas é natural considerar a variância como a mesma e daí estima-se a matriz de covariância comum $\sum$ pela matriz de covariância amostral conjunta (CHAVES NETO, 2025),

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)} \tag{56}$$

que é um estimador não-viciado daquele parâmetro $\sum$.

O método utilizado para duas populações pode ser estendido para diversas populações. Assim, a matriz conjunta para este caso é:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + \cdots + (n_g - 1)S_g}{n_1 + n_2 + \cdots + n_g - g} \tag{57}$$

3.5.9.4.2 Avaliação de Funções de Reconhecimento e Classificação

Segundo Johnson e WICHERN (1998), a probabilidade condicional de reconhecer um objeto $\underline{X}$ como da população ou grupo π_2 quando na verdade ele é do grupo π_1, dadas a regiões representativas das populações R_2 e R_1 e com $R_1 \cup R_2 = \Omega$ é:

$$P(2|1) = P(\underline{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\underline{x})d\underline{x} \tag{58}$$

e da mesma forma:

$$P(1|2) = P(\underline{X} \in R_1|\pi_2) = \int_{R_1} f_2(\underline{x})d\underline{x} \tag{59}$$

P(2|1) representa o volume formado pela f.d.p. $f_1(\underline{x})$ na região R_2 e com p_1 sendo a probabilidade a *priori* de π_1 e p_2 a probabilidade a *priori* de π_2, onde $p_1 + p_2 = 1$, as probabilidades de reconhecer corretamente ou incorretamente são dadas por:

$$\begin{aligned}
&P(\text{reconhecimento correto como } \pi_1) = P(\underline{x} \in \Pi_1 \text{ e é rec. corr. como } \pi_1) \\
&= P(\underline{x} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \\
&P(\text{reconhecimento incorreto como } \pi_1) = P(\underline{x} \in \Pi\pi_2 \text{ e é rec. incorr. como } \pi_1) \\
&= P(\underline{x} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \\
&P(\text{reconhecimento correto como } \pi_2) = P(\underline{x} \in \Pi_2 \text{ e é rec. corr. como } \pi_2) \\
&= P(\underline{x} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \\
&P(\text{reconhecimento incorreto como } \pi_2) = P(\underline{x} \in \pi_1 \text{ e é rec. incorr. como } \pi_2) \\
&= P(\underline{x} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1
\end{aligned} \tag{60}$$

Regras de reconhecimento de padrões e classificação são comumente avaliadas considerando as probabilidades de erros de classificação e os custos associados a essas decisões, denotados como $\underline{c}_{(j|i)}$. O custo médio ou esperado de uma decisão incorreta pode ser expresso como a soma dos produtos entre os custos dessas decisões e as respectivas probabilidades de ocorrência.

$$ECM = c(2|1).p(2|1)p_1 + c(1|2).p(1|2)p_2 \tag{61}$$

Uma boa regra de reconhecimento deve ter ECM muito baixo, tanto quanto possível é definida pelas desigualdades.

$$R_1: \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p_1}\right] \text{ que é } \left[\begin{matrix}\text{Razão das}\\ \text{densidades}\end{matrix}\right] \geq \left[\begin{matrix}\text{Razão dos}\\ \text{custos}\end{matrix}\right] . \left[\begin{matrix}\text{Razão das}\\ \text{probabilidades}\\ \text{à priori}\end{matrix}\right] \tag{62}$$

$$R_2: \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \left[\frac{c(1|2)}{c(2|1)}\right] \cdot \left[\frac{p_2}{p_1}\right] \text{ que é } \left[\begin{matrix}\text{Razão das}\\ \text{densidades}\end{matrix}\right] < \left[\begin{matrix}\text{Razão dos}\\ \text{custos}\end{matrix}\right] . \left[\begin{matrix}\text{Razão das}\\ \text{probabilidades}\\ \text{à priori}\end{matrix}\right] \tag{63}$$

Outro critério, além do ECM, pode ser usado para construir procedimentos ótimos. Trata-se daquele que escolhe as regiões R_1 e R_2 que minimizam a Probabilidade Total de Erro de Classificação (TPM).

$$\begin{aligned} \text{TPM} = {} & P(\underline{x} \in \pi_1 \text{ e é classificada errada}) + P(\underline{x} \in \pi_2 \text{ e é classificada errada}) \\ \text{TPM} = {} & p_1 \int_{R_2} f_1(\underline{x}) d\underline{x} + p_2 \int_{R_1} f_2(\underline{x}) d\underline{x} \end{aligned} \tag{64}$$

Assim, é equivalente a minimizar ECM quando os custos de classificação errada são iguais. Portanto, aloca-se uma nova observação $\underline{X}_0$ para a população com a maior probabilidade *posteriori* $P(\pi_i|\underline{x}_0)$, onde:

$$P(\pi_1|\underline{x}_0) = \frac{P(\pi_1 \text{ ocorre e se observa } \underline{x}_0)}{P(\text{se observa } \underline{x}_0)} = \frac{P(\text{se observa } \underline{x}_0|\pi_1)p(\pi_1)}{P(\text{se observa } \underline{x}_0|\pi_1)p(\pi_1) + P(\text{se observa } \underline{x}_0|\pi_2)p(\pi_2)} \tag{65}$$

$$P(\pi_1|\underline{x}_0) = \frac{p_1 f_1(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)} \text{ e } P(\pi_2|\underline{x}_0) = 1 - P(\pi_1|\underline{x}_0) = \frac{p_2 f_2(\underline{x}_0)}{p_1 f_1(\underline{x}_0) + p_2 f_2(\underline{x}_0)} \tag{66}$$

E, classifica-se $\underline{x}_0$ em π_1 quando:

$$P(\pi_1|\underline{x}_0) > P(\pi_2|\underline{x}_0) \tag{67}$$

Uma forma prática de visualizar a eficácia de um modelo de classificação, como a função discriminante linear de Fisher, é através da matriz de confusão. Essa matriz organiza os resultados da classificação em uma tabela, onde as frequências de alocações corretas aparecem

na diagonal principal, enquanto as alocações incorretas ficam nas demais posições. Cada célula reflete a relação entre as classes verdadeiras e as classes preditas, permitindo uma avaliação direta do desempenho do modelo ao separar grupos distintos.

3.5.9.5 Regressão Linear Múltipla

Hair Jr *et al.* (2009) definem a Regressão Linear Múltipla como um modelo que envolve duas ou mais variáveis independentes (X) para analisar a relação com uma única variável dependente (Y). Desta forma, busca-se usar as variáveis independentes, cujos valores são conhecidos, para prever os valores da variável dependente selecionada pelo pesquisador. Denota-se o modelo de Regressão Múltipla por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{p-1i} + \varepsilon_i \tag{68}$$

onde $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{p-1} X_{p-1i}$ é a parte sistemática do modelo, com β_0 sendo o intercepto, e ε_i $(i = 1,2,\cdots,n)$ o erro aleatório para a observação i.

Na forma matricial o mesmo modelo é descrito por:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \tag{69}$$

onde:

$\underline{Y}$: é o vetor resposta de dimensão n;

X: é a matriz do modelo de ordem $n \times p$;

$\underline{\beta}$: é o vetor de parâmetros de dimensão p;

$\underline{\varepsilon}$: é o vetor de erros de dimensão n.

A solução da equação (65) é a combinação linear $X\underline{\beta}$ que é altamente correlacionada com $\underline{Y}$ (STEFFEN, 2021).

3.5.9.5.1 Métricas de Erros em Regressão Múltipla Comparando com Dados Reais

Após a conclusão do modelo de regressão linear múltipla, é possível avaliar a precisão das previsões para verificar a adequação do modelo aos dados observados, utilizando os valores reais como parâmetro de análise. As métricas de erro mais utilizadas incluem o Erro Quadrático

Médio da Raiz (RMSE), o Erro Médio Absoluto (MAE) e o Erro Percentual Absoluto Médio (MAPE). Assim, seja $\underline{y_i}$ o valor real observado para a i-ésima amostra e $\underline{\hat{y}_i}$ o valor previsto pelo modelo, então essas métricas são definidas da seguinte forma:

a) Erro Quadrático Médio da Raiz (RMSE – *Root Mean Squared Error*)

O MSE mede a média dos quadrados das diferenças entre os valores previstos e observados. Um MSE baixo indica que as previsões estão próximas dos valores observados. Sua fórmula é dada por:

$$MSR = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\underline{y_i} - \underline{\hat{y}_i}\right)^2} \tag{70}$$

b) **Erro Médio Absoluto (MAE - *Mean Absolute Error*)**

O MAE calcula a média das diferenças absolutas entre os valores previstos e observados, expressando o erro médio nas mesmas unidades da variável dependente. Sua fórmula é:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\underline{y_i} - \underline{\hat{y}_i}\right| \tag{71}$$

c) **Erro Percentual Absoluto Médio (MAPE - *Mean Absolute Percentage Error*)**

O MAPE mede o erro percentual médio das previsões, expressando o erro como uma porcentagem dos valores observados. Sua fórmula é:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\underline{y_i} - \underline{\hat{y}_i}}{\underline{y_i}}\right| \times 100 \tag{72}$$

3.6 SÉRIES TEMPORAIS - METODOLOGIA BOX & JENKINS

Segundo Morettin e Toloi (2004), uma série temporal é definida como um conjunto de observações ordenadas no tempo. Essa característica a distingue das amostras aleatórias, em que a ordem das observações não é considerada relevante. Krajewski, Ritzman e Malhotra

(2009) acrescentam que as séries temporais constituem uma abordagem estatística fundamentada em dados históricos, com o objetivo de identificar padrões ao longo do tempo.

Pires (2001) complementa que, na Análise de Séries Temporais, o instante em que cada observação é registrada é fundamental, pois influencia diretamente a modelagem dos dados. As séries temporais ocorrem em diversas áreas do conhecimento, como economia, medicina, meteorologia e ciências sociais, e podem ser classificadas segundo a natureza do tempo (discreto ou contínuo) e dos valores observados (discretos ou contínuos) (PIRES, 2001). Assim, uma série temporal pode ser formalmente descrita por:

$$\{Z_1, Z_2, Z_3, \cdots, Z_{n-1}, Z_n\} \text{ ou } \{Z_t, t = 1,2,3, \cdots, n\} \tag{73}$$

Ainda de acordo com Pires (2001), os principais objetivos da Análise de Série Temporais são descrever o comportamento dos dados ao longo do tempo, ajustar modelos que representem essa dinâmica, realizar previsões futuras e, quando necessário, aplicar mecanismos de controle sobre o processo observado. Diante disso, alguns conceitos são necessários e serão abordados a seguir.

3.6.1 Modelos Estocásticos Estacionários e Não-Estacionários

Anselmo Chaves Neto (2025) define o processo estocástico como uma família de trajetórias para cada evento ω fixado. Em outras palavras, trata-se de uma coleção de variáveis aleatórias definidas sobre um mesmo espaço de probabilidade.

Morettin e Tolói (1981) destacam que uma das suposições mais comuns em séries temporais é a estacionariedade. Essa condição assegura que o processo estocástico parte de um estado de equilíbrio, no qual as propriedades estatísticas, como a média e a variância, permanecem constantes ao longo do tempo. Quando essa condição não é satisfeita, ou seja, quando a série apresenta tendência ou variância não constante, um dos procedimentos mais utilizados para torná-la estacionária é a aplicação de diferenças sucessivas da série original (MORETTIN e TOLÓI, 1981).

Assim, a primeira e segunda diferenciação é dada, respectivamente, por:

$$\Delta Z_t = Z_t - Z_{t-1} \tag{74}$$

$$\Delta^2 Z_t = \Delta[\,\Delta Z_t] = \Delta[Z_t - Z_{t-1}] \tag{75}$$

De modo geral, a n-ésima diferença de Z_t é:

$$\Delta^n Z_t = \Delta[\Delta^{n-1} Z_t] \tag{76}$$

Morettin e Tolói (2004) afirmam ainda que, para a maioria dos casos, será suficiente tomar uma ou duas diferenças para que a série se torne estacionária.

Adicionalmente, um dos métodos mais utilizados para avaliar a estacionariedade de séries temporais é o Teste de Dickey-Fuller Aumentado (ADF). Conforme Gujarati (2011, p. 751), o teste Dickey-Fuller tradicional pressupõe que os resíduos do modelo não apresentem autocorrelação. A autocorrelação ocorre quando os valores de uma série temporal estão correlacionados com seus próprios valores históricos, ou seja, quando há dependência sistemática entre uma observação e seus períodos anteriores. Quando a autocorrelação não é tratada, o modelo pode produzir estimativas enviesadas ou inválidas.

Para situações em que os resíduos apresentam autocorrelação, Dickey e Fuller propuseram a extensão do teste, conhecida como Dickey-Fuller Aumentado (ADF), que incorpora termos históricos da variável dependente para corrigir este problema.

A formulação geral do teste Dickey-Fuller Aumentado é representada pela seguinte equação:

$$\Delta Z_t = \beta_1 + \beta_2 t + \delta Z_{t-1} + \sum_{i=1}^{m} \theta_i \Delta Z_{t-i} + a_t \tag{77}$$

onde a_t é um termo de ruído branco, isto é, uma sequência de erros aleatórios com média zero, variância constante e ausência de autocorrelação e ΔZ_t a primeira diferenciação da série (MORETTIN; TOLOI, 1981; GUJARATTI, 2011).

A hipótese nula do teste Dickey-Fuller Aumentado é de que a série possui uma raiz unitária, ou seja, não é estacionária ($H_0: \delta = 0$). A hipótese alternativa indica que a série é estacionária ($H_1: \delta < 0$). Assim, de acordo com Morettin e Toloi (2004), quando a hipótese nula não é rejeitada, a série é considerada não estacionária, sendo necessário aplicar diferenciação sucessiva até alcançar estacionariedade.

3.6.2 Modelos de Média Móveis

Nos modelos denominados de Média Móvel (MA), a variável analisada é expressa em função da média dos termos de erro aleatório ocorridos nos períodos mais recentes. De acordo

com Gujarati (2011), esse tipo de modelo considera que o valor da série em determinado período t resulta da soma de uma constante e de uma média móvel dos erros aleatórios (ruído branco), atuais e anteriores. Morettin e Toloi (2004) complementam que, ao subtrair a média μ da observação Z_t, obtém-se uma nova variável $\tilde{Z}_t$, a qual depende linearmente de um número finito q de termos de erro passados. Sendo θ o coeficiente de média móvel, a formulação geral do modelo de média móvel de ordem q, denotado por MA(q), é apresentada por:

$$\tilde{Z}_t = a_t - \theta_1 a_1 - \theta_2 a_{t-2} - \theta_q a_{t-q} \tag{78}$$

Assim, o modelo MA(q) caracteriza-se como um processo estocástico em que o valor presente da série depende exclusivamente da influência dos q últimos choques aleatórios.

3.6.3 Modelo Auto Regressivo (AR)

De acordo com Gujarati (2011), os modelos Auto Regressivos (AR) e Média Móvel (MA) partem da suposição de que a série temporal é gerada por um sistema linear, o qual apresenta um termo de erro aleatório não correlacionado, com média zero e variância constante, caracterizando-se, como um ruído branco.

Conforme explicado por Morettin e Toloi (2004), em um modelo Auto Regressivo (AR), a série temporal Z_t é representada em função de seus próprios valores prévios, $Z_{t-1}, Z_{t-2}, \cdots, Z_{t-p}$, além do termo de erro aleatório a_t (ruído branco). A estrutura do modelo AR é expressa da seguinte forma:

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t \tag{79}$$

onde p é a ordem do modelo, caracterizando um processo auto-regressivo de ordem p, denotado por AR (p). De maneira semelhante, Gujarati (2011) destaca que, nesse tipo de modelo, o valor atual da série depende diretamente de seus p valores passados.

3.8.4 Metodologia Box-Jenkins e ARIMA (p, d, q)

Conforme apontado por Musial (2016), os modelos criados por Box e Jenkins partem do princípio de que os valores de uma série temporal apresentam forte dependência entre si, ou seja, cada observação pode ser explicada por dados anteriores da mesma série. Esses modelos

são conhecidos como Modelos Autoregressivos Integrados de Média Móvel, ou simplesmente ARIMA (*Autoregressive Integrated Moving Average*).

O modelo ARIMA surge da necessidade de lidar com séries temporárias que apresentarem características não estacionárias. Desta forma, o modelo é representado por três parâmetros: p, d e q. O parâmetro p corresponde ao número de termos autoregressivos (AR), d indica o número de diferenciações necessárias para tornar a série estacionária, e q representa o número de termos de média móvel (MA). Esses elementos formam a estrutura do modelo ARIMA(p, d, q), cuja escolha adequada depende do comportamento da série observada. Assim, seja $w_t = \Delta^d Z_t$, o modelo ARIMA(p, d, q), é dado pela equação:

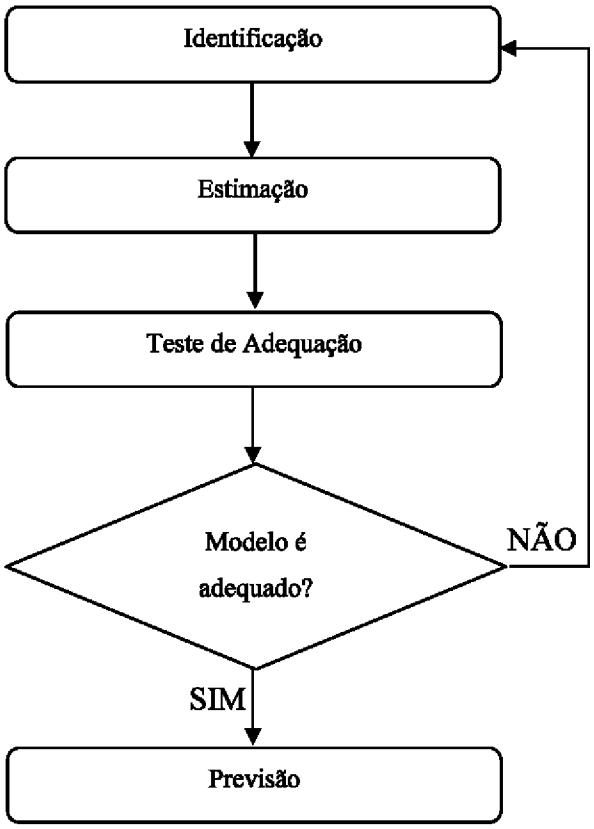
$$w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \cdots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \qquad (80)$$

Importa destacar que antes da aplicação do ARIMA são necessários alguns processos, sendo a Metodologia Box-Jenkins a mais utilizada. Anselmo Chaves Neto (2025) destaca que essa metodologia é composta por três fases:

- Identificação do Modelo;
- Estimação do Modelo;
- Verificação da adequação do Modelo ou testes.

De modo geral, o fluxograma (FIGURA 7) apresenta os vários estágios dessa Metodologia.

FIGURA 7 - ESTÁGIOS DA METODOLOGIA BOX E JENKINS



Fonte: Marchezan (2007).

A primeira etapa consiste na análise exploratória da série para identificar a estrutura ARIMA mais apropriada. Tradicionalmente, esta identificação é realizada por meio do exame das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), que fornecem indícios sobre os componentes autorregressivos (AR) e de médias móveis (MA) (BOX; JENKINS; REINSEL, 2008; MORETTIN; TOLOI, 2004). Além da análise gráfica, existem critérios quantitativos que auxiliam na escolha do modelo mais robusto. Deste modo, Matos (2018), destaca o Critério de Informação de Akaike (AIC), o qual considera o balanço entre ajuste e complexidade do modelo. O AIC é definido por:

$$AIC(p) = -2logL_p + 2(k+1) \tag{81}$$

em que L_p representa o valor máximo da função verossimilhança do modelo ajustado k corresponde ao número de parâmetros estimados.

Segundo Morettin e Toloi (2004), a máxima verossimilhança é um método que determina os valores dos parâmetros que maximizam a probabilidade de ocorrência dos dados observados, considerando a distribuição assumida pelo modelo.

Seja $L(\vartheta|Z)$ é a função de verossimilhança condicional à amostra Z, em função do vetor de parâmetros ϑ, o estimador por máxima verossimilhança é expresso por:

$$\hat{\vartheta} = \text{argmax}_{\vartheta}\, L(\vartheta|Z) \tag{82}$$

O processo de estimação, que também utiliza a função de verossimilhança, corresponde ao cálculo dos parâmetros que definem a estrutura do modelo, usando os termos autorregressivos, de médias móveis e, quando aplicável, a variância dos resíduos. Já a etapa verificação busca avaliar a adequação do modelo ajustado, por meio da análise dos resíduos, com o intuito de confirmar se o modelo consegue capturar adequadamente as características da série temporal, sem deixar padrões auto correlacionados não explicados. Por fim, na etapa final, são feitas projeções futuras com base no modelo previamente validado. Essa fase só é executada quando os testes de diagnóstico confirmam a adequação do modelo estimado.

Caso a verificação aponte inconsistências ou inadequações, o procedimento iterativo é retomado, revisando as etapas anteriores até que se obtenha um modelo satisfatório, capaz de representar o comportamento da série analisada (MORETTIN; TOLOI, 2004).

4 PROCEDIMENTOS METODOLÓGICOS

Este capítulo tem por finalidade descrever os métodos e procedimentos aplicados neste trabalho, especialmente sobre o banco de dados utilizados e a aplicação das Técnicas Multivariadas apresentadas nas Seções 3.5 e 3.6.

4.1 BANCO DE DADOS

Os dados usados na presente pesquisa foram obtidos por meio de *sites* como o *Yahoo Finance* e *Fred,* e também fornecidos pelo Empresa A. Para a formação do banco de dados, foram consideradas variáveis previamente discutidas em estudos de autores apresentados na Seção 3.4, com base em sua relevância teórica e potencial influência sobre os preços dos Óleos Básicos.

A Empresa A também forneceu os dados da série histórica de preços de 31 Óleos Básicos, sendo alguns pertencentes ao Grupo 1, outros ao Grupo 2 e os demais ao Grupo 3. Entre esses produtos, um Óleo Básico específico apresenta a série histórica mais longa, com registros mensais de janeiro de 2010 à fevereiro de 2024. Os demais 30 Óleos Básicos contam com séries históricas mensais que abrangem o período de janeiro de 2015 à junho de 2024. No entanto, para esta pesquisa, a análise com dados reais de todas as do Óleos Básicos foi restrita aos dados reais até julho de 2023, para que se alinhasse com a disponibilidade das variáveis independentes.

Assim, foi construído um banco de dados com 227 variáveis que vão de janeiro de 2010 (jan/2010) a julho de 2023(jul/2023), contendo informações sobre:

- oferta, demanda, estoques de petróleo e derivados, e capacidade de refino em diferentes continentes;
- commodities minerais e de energia derivados do petróleo,
- índices Dow Jones setoriais;
- ações de empresas mundiais de energia;
- paridade entre moedas mundiais e o dólar;
- índices de bolsa globais e;
- quantidade de moeda em circulação como indicador de liquidez mundial.

Para o Óleo Básico com a série mais longa (iniciando em jan/2010), foram utilizadas as variáveis independentes desde esse mesmo período. Para os outros 30 Óleos Básicos, as análises com variáveis independentes iniciaram em janeiro de 2015.

As projeções geradas para períodos com dados reais disponíveis foram comparadas aos valores observados, e os erros foram calculados por meio da métrica RMSE. Como o modelo apresentou bom desempenho nas projeções testadas, para a esse Óleo de maior série, optou-se por realizar previsões para períodos posteriores a julho de 2023, mesmo com a ausência de dados reais para validação imediata. Isso porque, as variáveis dependentes têm dados até fevereiro de 2024, no caso da série mais longa, e até junho de 2024 para as outras, o que permite calcular a precisão dos modelos utilizados e descritos nas seções seguir.

4.2 ESCOLHA DOS MÉTODOS DE ANÁLISE MULTIVARIADA

O Banco de Dados neste estudo apresenta alta multidimensionalidade, que pode atrapalhar a análise direta do impacto das variáveis. Assim, como discutido em Seções anteriores e como afirmado por Guerreiro (2021), o uso de técnicas de redução de dimensionalidade torna-se uma alternativa viável para minimizar essa variabilidade.

Nesse contexto, a escolha dos Métodos Multivariados justifica-se pela necessidade de reduzir a complexidade do banco e, ao mesmo tempo, preservar a variância explicada pelas variáveis originais. Técnicas como a Análise das Componentes Principais (PCA) e a Análise Fatorial (AF) permitem atingir esse objetivo (SIQUEIRA *et al.*, 2013).

Adicionalmente, foi empregada a Análise de Agrupamento (*Cluster*) com o intuito de verificar se, a partir das 227 variáveis iniciais, seria possível identificar subconjuntos com características semelhantes. O objetivo foi construir um novo banco de dados reduzido, composto pelas variáveis pertencentes ao grupo com maior número de observações, e posteriormente aplicar a PCA nesse novo conjunto para avaliar se teriam bons resultados preditivos.

Para verificar a validade das variáveis obtidas por *cluster*, foi utilizada a Análise Discriminante, para avaliar se os grupos formados foram corretamente classificados. Caso o resultado mostrasse um baixo índice de acurácia na alocação das variáveis usando as Componentes, seria necessário refazer a Análise de Agrupamento alterando o número de grupos.

4.3 PROCEDIMENTOS E APLICAÇÃO DOS MÉTODOS ESTATÍSTICO

4.3.1 O uso de Séries Temporais

Inicialmente, a base de dados original com série mais longa (jan/2010 a fev/2024) foi dividida em diferentes janelas temporais, como jan/2010-dez/2022, jan/2010-jan/2023, e sucessivamente até se chegar no período de jan/2010-jul/2023, totalizando oito janelas. Para cada um destes subconjuntos temporais, todos com as mesmas variáveis, foi aplicado o modelo ARIMA a fim de projetar valores futuros para horizontes de 3 e 6 meses. Para o horizonte de 12 meses, a divisão das janelas foi de jan/2010-jul/2022 até fev/2023. Os períodos projetados foram então adicionados aos respectivos conjuntos de dados reais, originando bases ampliadas.

Todas as análises foram feitas no *Python* e o procedimento metodológico consistiu em três etapas principais: identificação da estacionariedade, seleção do modelo ARIMA ótimo e realização das projeções futuras.

A identificação da estacionariedade foi realizada por meio do Teste de Dickey-Fuller Aumentado (ADF), disponível na função *adfuller()* da biblioteca *statsmodels.tsa.stattools*. Para cada série temporal pertencente às janelas temporais definidas, foi verificada a presença de raiz unitária, determinando o número mínimo de diferenciações (d) necessário para estacionarizar a série.

A seleção do modelo ARIMA foi feita através da varredura automatizada de combinações dos parâmetros (p, d, q), com p e q variando de 0 a 3, utilizando a função ARIMA() da biblioteca *statsmodels.tsa.arima.model*. Para cada combinação possível, o critério de avaliação adotado foi o Critério de Informação de Akaike (AIC), buscando o modelo com menor AIC.

A estimação dos parâmetros do modelo foi realizada via Método da Máxima Verossimilhança, implementado automaticamente pela função *fit()* do *statsmodels*. A validação do modelo foi feita pela análise dos resíduos, assegurando a inexistência de autocorrelação, o que caracteriza um bom ajuste do modelo aos dados históricos.

Por fim, a previsão dos valores futuros foi realizada através da função *forecast()*, também da biblioteca *statsmodels*, gerando as projeções para horizontes de 3, 6 e 12 meses, dependendo da janela temporal analisada.

4.3.1.1 Análise Diagnóstica e Validação dos Modelos ARIMA

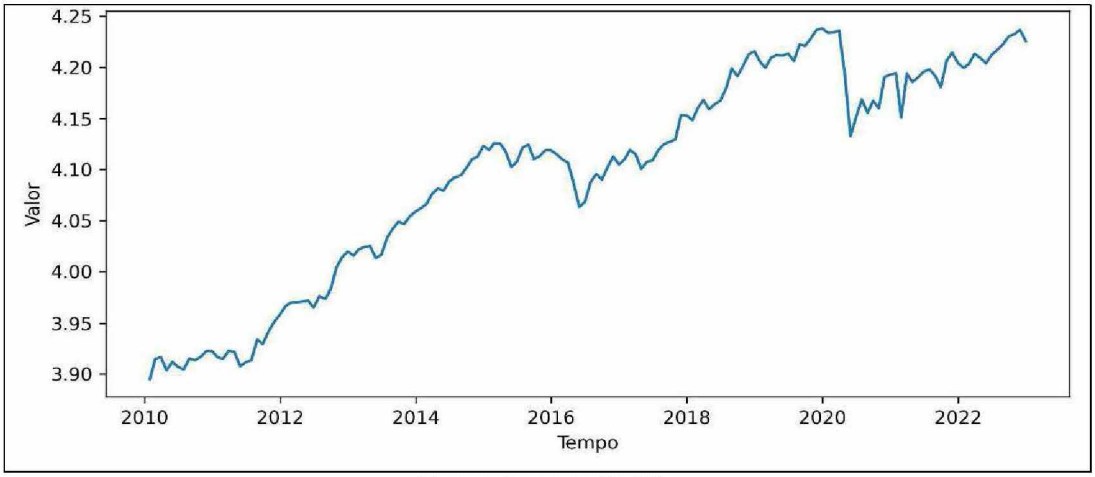
Após a estimação dos modelos ARIMA selecionados com base no critério de informação de Akaike (AIC), foi realizada uma análise diagnóstica com o objetivo de avaliar a adequação dos modelos ajustados às séries temporais analisadas. Nesta etapa foi analisado o gráfico da série temporal original, a análise das funções de autocorrelação (FAC) e autocorrelação parcial (FACP), utilizadas tanto no apoio à identificação do modelo quanto na avaliação do comportamento dos resíduos.

Adicionalmente, foi realizada a análise do periodograma acumulado dos resíduos, com o objetivo de avaliar o comportamento espectral dos erros do modelo. Observou-se que o periodograma acumulado apresentou trajetória próxima à linha teórica esperada para um processo de ruído branco, indicando distribuição aproximadamente uniforme da potência espectral ao longo das frequências. Esse resultado indica uma boa adequação dos modelos ARIMA ajustados.

De forma complementar, a qualidade preditiva dos modelos foi avaliada por meio do Erro Quadrático Médio da Raiz (RMSE) e Erro Médio Absoluto (MAE), calculado a partir de conjuntos de validação correspondentes aos últimos períodos observados em cada janela temporal. Os valores obtidos indicaram desempenho satisfatório dos modelos para a maioria das séries analisadas.

Abaixo tem-se o gráfico do comportamento da série correspondente a uma variável específica do banco de dados com horizonte de seis meses. Com isso, foi possível observar um comportamento caracterizado por tendência crescente ao longo do período de 2010 a 2023, evidenciando a não estacionariedade da série. Além disso, houve flutuações de curto prazo em torno da tendência, indicando a presença de dependência temporal entre as observações. Destaca-se ainda a ocorrência de um choque mais pronunciado por volta de 2020, seguido de recuperação gradual, comportamento típico de séries econômicas sujeitas a eventos exógenos, como o caso da COVID-19. Essas características justificaram a adoção de modelos ARIMA, com aplicação de diferenciação para estabilização da média e inclusão de termos autorregressivos e de médias móveis para capturar a dinâmica temporal da série.

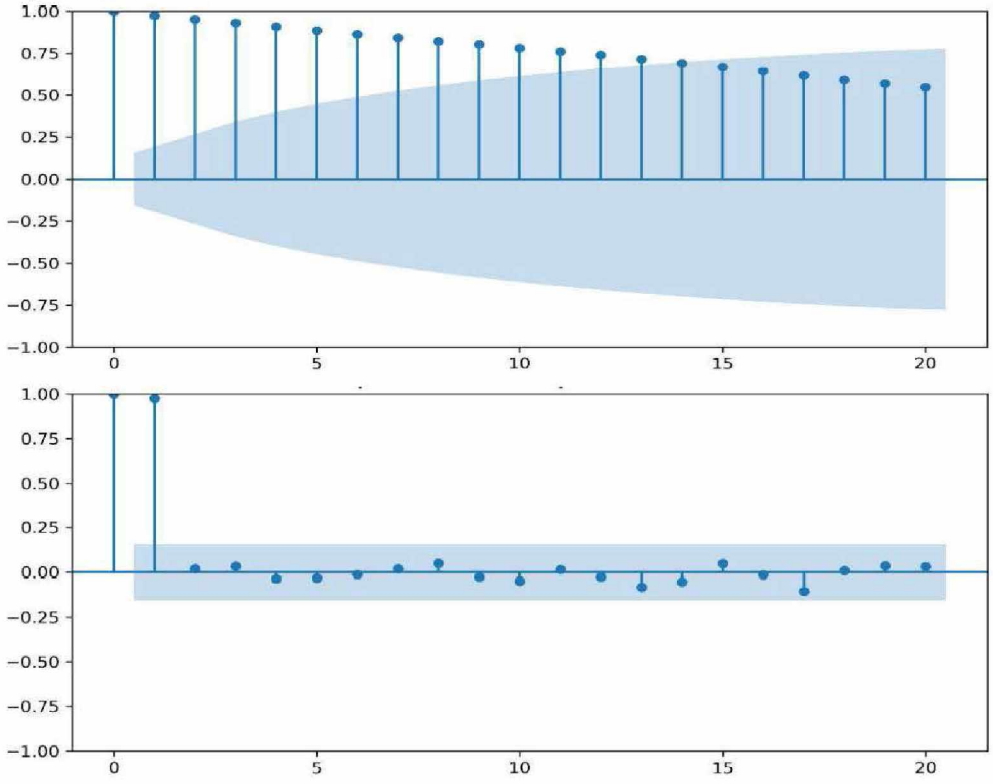
FIGURA 8 - SÉRIE TEMPORAL - VARIÁVEL REPRESENTATIVA



Fonte: A autora (2025).

Além disso, a análise conjunta da FAC e da FACP permitiu identificar a necessidade de diferenciação da série, bem como a presença de um número reduzido de termos autorregressivos, orientando a escolha de modelos ARIMA, posteriormente validados por critérios de informação e desempenho preditivo. A FIGURA 9 ilustra os resultados para a variável analisada.

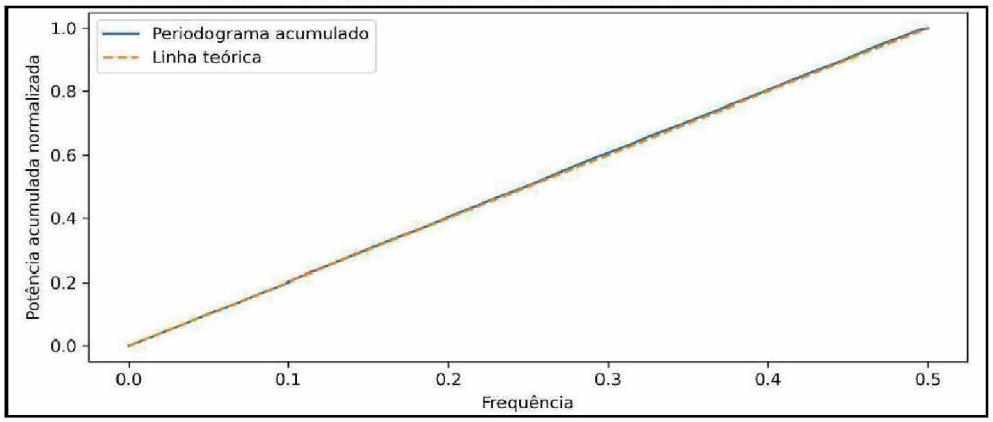
FIGURA 9 - FUNÇÃO DE AUCORRELAÇÃO E AUTOCORRELAÇÃO PARCIAL (FAC E FACP)



Fonte: A autora (2025).

Após o ajuste da série por meio do modelo ARIMA, o periodograma acumulado foi analisado. Observa-se que a curva empírica apresenta comportamento próximo à linha teórica, indicando uma distribuição aproximadamente uniforme da potência ao longo das frequências (FIGURA 10). Esse resultado sugere que não há componentes periódicos relevantes remanescentes, o que reforça que a principal estrutura temporal da série foi adequadamente capturada pelo modelo.

FIGURA 10 - PERIODOGRAMA ACUMULADO DOS RESÍDUOS



Fonte: A autora (2025).

Por fim, para o horizonte considerado, o erro médio absoluto (MAE) foi de 0,0961 e a raiz do erro quadrático médio (RMSE) foi de 0,1084, indicando desvios relativamente baixos entre os valores observados e os estimados pelo modelo. Esses resultados reforçaram a capacidade do modelo em representar a dinâmica temporal da série e em fornecer previsões consistentes.

4.3.2 O uso de Técnicas Multivariada

4.3.2.1 Análise de Componentes Principais (PCA) e Análise Fatorial (AF)

Como o banco de dados reúne informações de diferentes setores e diversos países, foram aplicados alguns tratamentos tais como a padronização dos valores monetários para o dólar americano e aplicação do logaritmo na base 10. Esses procedimentos visaram reduzir a amplitude de variação entre as escalas das variáveis, permitindo rodar Análise das Componentes Principais e Fatorial de forma que variáveis com maiores valores não distorcessem sua importância na formação das Componentes e dos Fatores.

Para a aplicação da Análise Fatorial, que exige correlação entre as variáveis originais, foram selecionadas, dentre as 227 disponíveis, 135 variáveis relacionadas ao mercado

financeiro, tais como índices de bolsa, paridade de moedas, preço de ações de empresas de energia e preços dos diferentes tipos de petróleo. A seleção desse subconjunto se deu pelo fato de que todas essas variáveis são negociadas em bolsas de valores globais e, por isso, pode haver uma alta colinearidade entre estas variáveis.

Nas bases com dados reais, sem a adição das projeções, foram aplicadas a PCA ao banco de 227 variáveis e a AF com as 135 variáveis, de modo a extrair, respectivamente, as Componentes Principais e os Fatores. A partir dessas extrações, foram obtidas equações de Regressão Múltipla para cada janela de tempo, sendo as variáveis explicativas formadas (a) pelas Componentes Principais e (b) pelos Fatores.

As equações de regressão foram ajustadas utilizando apenas os períodos correspondentes aos dados reais. A acurácia dos modelos foi avaliada por meio do MAPE aplicado aos períodos projetados, tomando como referência os valores reais observados.

Para cada uma das técnicas foram usadas bibliotecas específicas do Python. Na Análise das Componentes Principais, foram utilizados os pacotes *sklearn.decomposition* e *sklearn.preprocessing*, com o método *StandardScaler*, com o objetivo de centralizar as variáveis em média zero e desvio padrão igual a um. Já na Análise Fatorial, foram utilizados os pacotes *factor_analyzer* e *sklearn.decomposition* para extrair os fatores. Para verificar suposições específicas desse método, recorreu-se as funções *calculate_bartlett_sphericity* e *calculate_kmo*.

A PCA segue um modelo formativo, no qual as variáveis observadas compõem as Componentes Principais, enquanto a AF adota um modelo reflexivo, em que as variáveis são tratadas como indicadores causados por Variáveis Latentes. Por essa razão, no caso da AF, procedimentos adicionais como a rotação ortogonal do tipo VARIMAX foram realizados para facilitar a interpretação dos Fatores.

A aplicação dessas técnicas seguiu uma sequência padrão: carregamento dos dados, pré-processamento, cálculo do vetor médio e centralização, obtenção da matriz de covariância ou correlação, cálculo de autovalores e autovetores e, por fim, extração das componentes (PCA) ou dos fatores (AF), ordenados conforme a variância explicada.

4.3.2.2 Análise de Agrupamento (*Cluster Analysis*)

Importa destacar que os procedimentos descritos anteriormente, foram executados exclusivamente para o Óleo Básico com a série temporal mais longa (jan/2010 a fev/2024),

pertencente ao Grupo 1. Isso se deve ao fato de que, para os demais 30 Óleos Básicos, cujas séries começam em jan/2015, os testes com o banco completo de 227 variáveis e com o de 135 variáveis resultaram em erros mais elevados, mesmo após a aplicação da PCA e AF.

Diante disso, com o banco de 227 variáveis com as projeções de três meses nas 8 janelas de tempo, foi aplicado a técnica de Análise de Agrupamento (*Cluster*) para agrupar variáveis com base nos padrões de associação presentes nas Componentes Principais extraídas pela PCA. Através do pacote *sklearn.cluster*, o algoritmo *K-means* foi aplicado aos *loadings*, com número de clusters variando entre 2 e 14. Para definição do número ideal de agrupamentos, foram utilizados dois métodos: Curva do Cotovelo (*Elbow Curve*) e o Coeficiente de Cilhueta (*Silhueta Score*), chegando a um número ideal de sete Cluster. O primeiro foi implementado com o auxílio do pacote *seaborn* para visualização e *sklearn.cluster* para cálculo da inércia e, o segundo, implementado via *yellowbrick.cluster.SilhouetteVisualizer*.

Considerando que esse processo de agrupamento foi realizado para cada uma das janelas temporais com o horizonte de 3 meses, e a partir da análise das variáveis presentes em cada grupo resultante em cada janela, foram adotados os seguintes critérios para a definição e escolha do conjunto final de variáveis:

- Priorizaram-se as variáveis que apareceram com maior frequência entre os grupos identificados nos diferentes períodos analisados.
- Optou-se por selecionar o grupo (*cluster*) com o maior número de variáveis recorrentes na maioria das janelas de tempo analisada, que neste caso foi o Grupo 1, com 71 variáveis.

A segunda e a quinta janelas de tempo apresentaram uma classificação distinta das demais para o *Cluster* 1 (Tabela 4). Por isso, a seleção final foi feita com base na intersecção das variáveis presentes nas outras seis janelas.

TABELA 4 - VARIÁVEIS POR CLUSTER

Cluster	Período								
	Jan – Mar/2223	Fev – Abr/2023	Mar – Mai/2023	Abr – Jun/2023	Mai – Jul/2023	Jun – Ago/2023	Jul – Set/2023	Ago – Out/2023	Número de variáveis recorrentes
1	75	21	74	75	12	74	76	80	71
2	41	36	36	35	43	39	41	38	30
3	9	30	36	26	29	19	17	8	6
4	35	28	28	28	33	31	32	31	0
5	30	28	30	31	29	28	28	17	0
6	20	16	16	16	14	13	10	25	0
7	17	68	17	16	67	23	23	28	0
Total	227	227	227	227	227	227	227	227	

Fonte: A autora (2025).

4.3.2.3 Análise Discriminante

Para verificar a qualidade dos agrupamentos das variáveis, foi utilizada a Análise Discriminante Linear (LDA). Os *loadings* serviram como variáveis explicativas e os rótulos de cada *cluster* como variável resposta. A LDA foi implementada com o pacote *sklearn.discriminant_analysis* e teve como objetivo avaliar a separação entre os grupos formados. A classificação dos dados foi comparada às originais por meio da matriz de confusão (pacote *sklearn.metrics.confusion_matrix*) e da acurácia do modelo (pacote *accuracy_score*). Os resultados confirmaram a adequação da classificação para as variáveis do Grupo.

Desse modo, esse novo conjunto de 71 variáveis mantém a diversidade temática do banco original, contendo informações sobre dinâmica de oferta e demanda de petróleo e derivados, preços de commodities energéticas, indicadores financeiros e cambiais, ações de empresas do setor energético, entre outros fatores relevantes. Trata-se, portanto, de um subconjunto representativo das 227 variáveis iniciais, com a vantagem de menor dimensionalidade e menor risco de multicolinearidade.

4.3.2.4 Aplicação das Técnicas ao Subconjunto de 71 Variáveis e aos 30 Óleos Básicos

Com esse novo subconjunto, o mesmo procedimento foi repetido para os 30 Óleos Básicos. Inicialmente, foram feitas projeções para 3, 6 e 12 meses sem técnicas de redução de dimensionalidade, e os erros ficaram, em sua maioria, acima de 30%. Já com a aplicação das

Componentes Principais, os erros foram consideravelmente reduzidos. Desta forma, optou-se por utilizar esse subconjunto com PCA para a modelagem desses Óleos.

Cabe destacar que para o banco de dados, seja com 227, 135 ou 71 variáveis, foi adotado o mesmo procedimento metodológico: a aplicação de Regressão Múltipla a partir das variáveis selecionadas.

A partir da combinação das 71 variáveis com a PCA, foi possível identificar quais variáveis apresentaram impacto positivo (aumento de preço) ou negativo (queda de preço) sobre cada Óleo Básico analisado. Para isso, os *loadings* obtidos na PCA foram combinados com os coeficientes das equações de Regressão Múltipla, permitindo avaliar a direção e intensidade do impacto de cada variável sobre os preços.

A Tabela 5 mostra as divisões de período de análise da aplicação de cada Técnica:

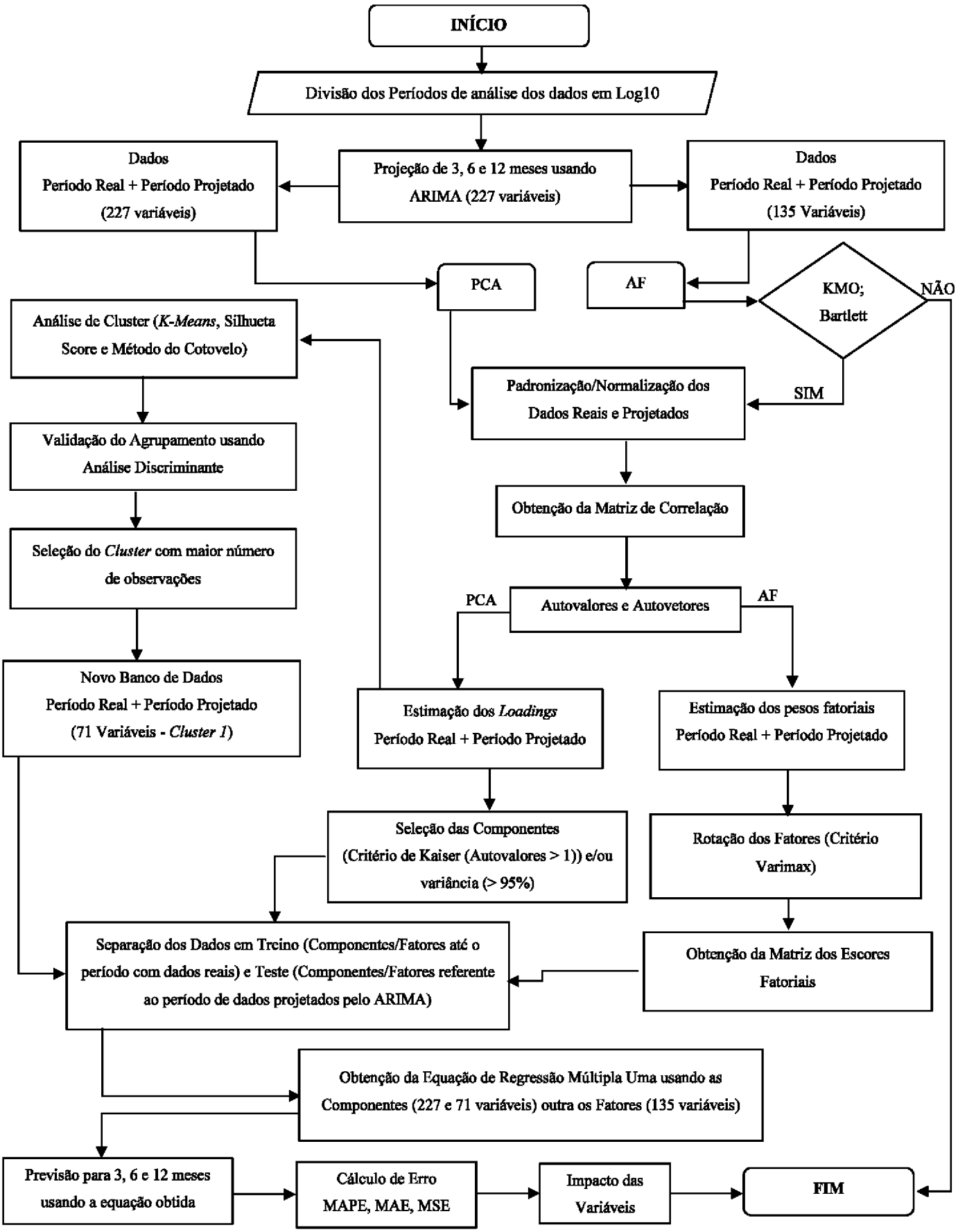
TABELA 5 - PERÍODO E APLICAÇÃO DAS TÉCNICAS MULTIVARIADAS

Janela de Tempo	**Conjunto de Variáveis**	**Óleos Básicos Analisado(s)**	**Métodos Aplicados**	**Projeções/Avaliações**
Jan/2010 – Dez/2022 até Jan/2010 – Jul/2023)	227 variáveis (PCA) 135 variáveis (AF)	Um Óleo Básico específico (Grupo 1 - Empresa A)	ARIMA, PCA, AF, Análise de Agrupamento, Análise Discriminante, Regressão Múltipla	Projeções: 3 e 6 meses/ LDA, Matriz de Confusão, Acurácia. Avaliação via MAPE, MAE, MSE
Jan/2010 – Jul/2022 até Jan/2010 – Fev/2023)	227 variáveis (PCA) 135 variáveis (AF)	Um Óleo Básico específico (Grupo 1 - Empresa A)	ARIMA, PCA, AF, Regressão Múltipla	Projeções: 12 meses Avaliação via MAPE, MAE, MSE
Jan/2015 – Dez/2022 até Jan/2015 – Jul/2023)	71 variáveis (selecionadas via cluster)	Os outros 30 Óleos Básicos	ARIMA PCA (para *loadings*) Regressão Múltipla	Projeções: 3 e 6 meses Avaliação via MAPE, MAE, MSE. Análise do Impacto das Variáveis
Jan/2015 – Jul/2022 até Jan/2015 – Fev/2023)	71 variáveis (selecionadas via cluster)	Os outros 30 Óleos Básicos	ARIMA PCA (para *loadings*) Regressão Múltipla	Projeções: 12 meses Avaliação via MAPE, MAE, MSE

Fonte: A Autora (2025).

De modo geral, o fluxograma a seguir, FIGURA 11, mostra todos os procedimentos metodológicos adotados neste estudo.

FIGURA 11 - FLUXOGRAMA DOS PROCEDIMENTOS UTILIZADOS



Fonte: A autora (2025).

5 TESTES COMPUTACIONAIS, RESULTADOS E DISCUSSÕES

Os testes computacionais foram implementados em *Python* com auxílio do *Microsoft* Excel para análise de planilhas, em um computador com sistema operacional Microsoft *Windows 11 Home Single Language* (64 bits), equipado com processador Intel® Core™ i5-11300H de 11ª geração (3.10 GHz), 8 GB de memória RAM e placa de vídeo dedicada com 4 GB de memória.

Todos os testes foram realizados com base nas três versões de banco de dados: (a) com 227 variáveis, (b) com 135 variáveis e (c) com 71 variáveis, sendo os dois últimos subconjuntos do banco original (a). Assim, neste capítulo, são apresentados e discutidos os resultados obtidos por meio das técnicas de Análise Multivariada aplicadas a cada base de dados. São abordados os resultados das etapas de Redução de Dimensionalidade (PCA e AF), Reconhecimento de Padrões e Classificação (Análise de *Cluster* e Discriminante), assim como a identificação das variáveis de maior impacto sobre o preço de cada Óleo Básico analisado.

5.1 RESULTADOS DA ANÁLISE DAS COMPONENTES PRINCIPAIS COM O ÓLEO BÁSICO DE MAIOR SÉRIE TEMPORAL

A Tabela 6 apresenta os valores do erro percentual absoluto obtidos nas projeções com horizonte de três meses, utilizando Análise das Componentes Principais (PCA) com o banco de 227 variáveis. O número de Componentes foi selecionado a partir do critério de Kaiser. Logo, em todas as janelas de tempo se manteve o número de Componentes Principais em dezesseis, variando-se apenas o período utilizado para a geração das Componentes e os meses projetados.

TABELA 6 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO PCA - 3 MESES

Período de Geração das Componentes	**Período Projetado**	**Meses Projetados (Erro Percentual Absoluto %)**
Jan/2010 – Dez/2022	Jan/2023 – Mar/2023	Jan: 4,05% / Fev: 0,03% / Mar: 0,32%
Jan/2010 – Jan/2023	Fev/2023 – Abr/2023	Fev: 3,68% / Mar: 2,91% / Abr: 0,34%
Jan/2010 – Fev/2023	Mar/2023 – Mai/2023	Mar: 6,69% / Abr: 4,08% / Mai: 1,05%
Jan/2010 – Mar/2023	Abr/2023 – Jun/2023	Abr: 4,80% / Mai: 7,72% / Jun: 10,30%
Jan/2010 – Abr/2023	Mai/2023 – Jul/2023	Mai: 9,25% / Jun: 11,56% / Jul: 12,95%
Jan/2010 – Mai/2023	Jun/2023 – Ago/2023	Jun: 7,28% / Jul: 8,09% / Ago: 1,85%
Jan/2010 – Jun/2023	Jul/2023 – Set/2023	Jul: 13,31% / Ago: 6,63% / Set: 4,05%

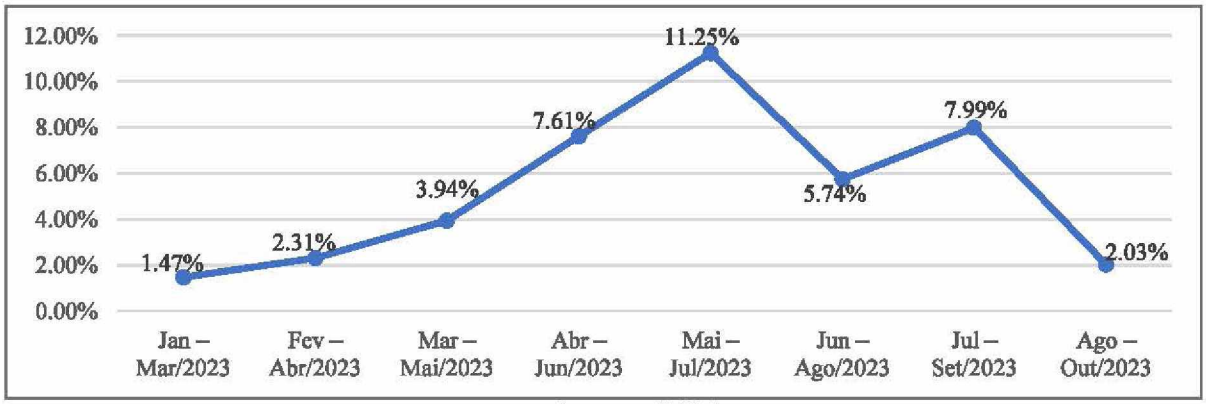
Jan/2010 – Jul/2023	Ago/2023 – Out/2023	Ago: 3,39% / Set: 1,51% / Out: 1,17%

Fonte: A autora (2025).

É possível observar no Gráfico 1 que os resultados do modelo apresentaram, em sua maioria, MAPE inferiores a 5%, indicando que o modelo apresentou um bom desempenho. Além disso, os menores erros ocorreram nas primeiras janelas de tempo, especialmente no modelo treinado com dados até dez/2022, cuja média de erro foi de 1,47%.

Em algumas janelas de tempo os erros forma maiores, como jan/2010-abr/2023, jan/2010-mai/2023 e jan/2010-jul/2023. Essas variações podem estar associadas a mudanças na dinâmica das variáveis ou à presença de ruídos que não foram completamente capturados pelas Componentes Principais.

GRÁFICO 1 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 3 MESES - PCA



Fonte: A autora (2025).

A Tabela 7 apresenta o Erro Médio Percentual de 30, 60 e 90 dias. Observa-se que a média de erros de 30 dias foi 6,56%, para 60 dias foi de 5,32% e para 90 dias, 4%. Para este caso, o aumento do horizonte de projeção não implicou em um aumento do erro médio de projeção.

TABELA 7 - ERRO MÉDIO PARA PERÍODO DE 30, 60 E 90 DIAS USANDO PCA

Período	Mês 1	Mês 2	Mês 3	Mês 4	Mês 5	Mês 6	Mês 7	Mês 8	MAPE
30 Dias	4,05%	3,68%	6,69%	4,80%	9,25%	7,28%	13,31%	3,39%	**6,56%**
60 Dias	0,03%	2,91%	4,08%	7,72%	11,56%	8,09%	6,63%	1,51%	**5,32%**
90 Dias	0,32%	0,34%	1,05%	10,30%	12,95%	1,85%	4,05%	1,17%	**4,00%**

Fonte: A autora (2025).

Mantendo-se os mesmos períodos para geração das Componentes, os resultados das projeções com horizonte de seis meses, apresentados na Tabela 8, demonstram um bom desempenho que, na maioria dos casos, é superior ao observado nas projeções para três meses.

TABELA 8 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO PCA – 6 MESES

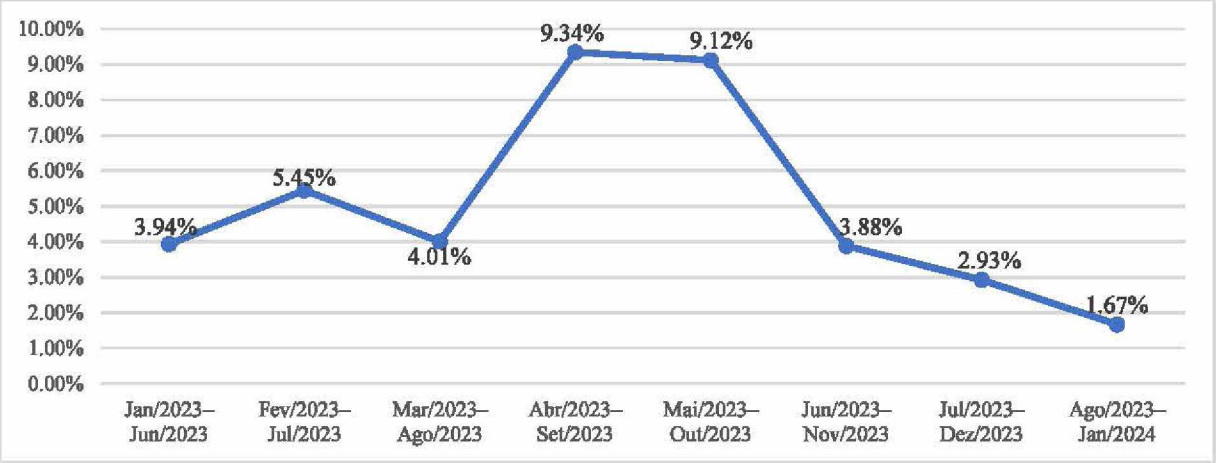
Período Projetado	Meses Projetados (Erro Percentual Absoluto %)
Jan/2023– Jun/2023	Jan: 3,55% / Fev: 0,78% / Mar: 0,16% / Abr: 3,19% / Mai: 6,23% / Jun: 9,72%
Fev/2023– Jul/2023	Fev: 2,76% / Mar: 2,13% / Abr: 1,96% / Mai: 5,54% / Jun: 9,19% / Jul: 11,14%
Mar/2023– Ago/2023	Mar: 3,98% / Abr: 0,89% / Mai: 2,32% / Jun: 5,99% / Jul: 8,16% / Ago: 2,74%
Abr/2023– Set/2023	Abr: 6,10% / Mai: 9,17% / Jun: 12,49% / Jul: 14,10% / Ago: 7,92% / Set: 6,26%
Mai/2023– Out/2023	Mai: 9,53% / Jun: 12,75% / Jul: 14,62% / Ago: 8,28% / Set: 6,09% / Out: 3,47%
Jun/2023– Nov/2023	Jun: 7,39% / Jul: 8,66% / Ago: 2,40% / Set: 0,51% / Out: 1,98% / Nov: 2,36%
Jul/2023 – Dez/2023	Jul: 9,77% / Ago: 3,08% / Set: 0,85% / Out: 1,76% / Nov: 2,11% / Dez: 0,03%
Ago/2023 – Jan/2024	Ago: 3,15% / Set: 1,27% / Out: 1,43% / Nov: 1,64% / Dez: 0,84% / Jan: 1,68%

Fonte: A autora (2025).

No gráfico 2 é possível notar que cinco dos oito cenários avaliados apresentaram MAPE inferior a 5%, com o melhor desempenho registrado no modelo treinado até julho de 2023, cuja média de erro foi de apenas 1,67%, mesmo projetando até janeiro de 2024.

Embora dois períodos tenham concentrado os maiores erros médios, jan/2010-mar/2023 e jan/2010-abr/2023, ambos com valores superiores a 9%, essas exceções não comprometem a tendência geral de melhora. Adicionalmente, a Tabela 9 demonstra que o MAPE para cada uma das janelas de previsão analisadas (30, 60, 120, 150 e 180 dias) permaneceu abaixo de 6%.

GRÁFICO 2 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 6 MESES - PCA



Fonte: A autora (2025).

TABELA 9 - ERRO MÉDIO PARA PERÍODO DE 30, 60, 90, 150 E 180 DIAS USANDO PCA

Período	Mês 1	Mês 2	Mês 3	Mês 4	Mês 5	Mês 6	Mês 7	Mês 8	MAPE
30 Dias	3,55%	2,76%	3,98%	6,10%	9,53%	7,39%	9,77%	3,15%	**5,78%**
60 Dias	0,78%	2,13%	0,89%	9,17%	12,75%	8,66%	3,08%	1,27%	**4,84%**
90 Dias	0,16%	1,96%	2,32%	12,49%	14,62%	2,40%	0,85%	1,43%	**4,53%**
120 Dias	3,19%	5,54%	5,99%	14,10%	8,28%	0,51%	1,76%	1,64%	**5,13%**
150 Dias	6,23%	9,19%	8,16%	7,92%	6,09%	1,98%	2,11%	0,84%	**5,32%**
180 Dias	9,72%	11,14%	2,74%	6,26%	3,47%	2,36%	0,03%	1,68%	**4,67%**

Fonte: A autora (2025).

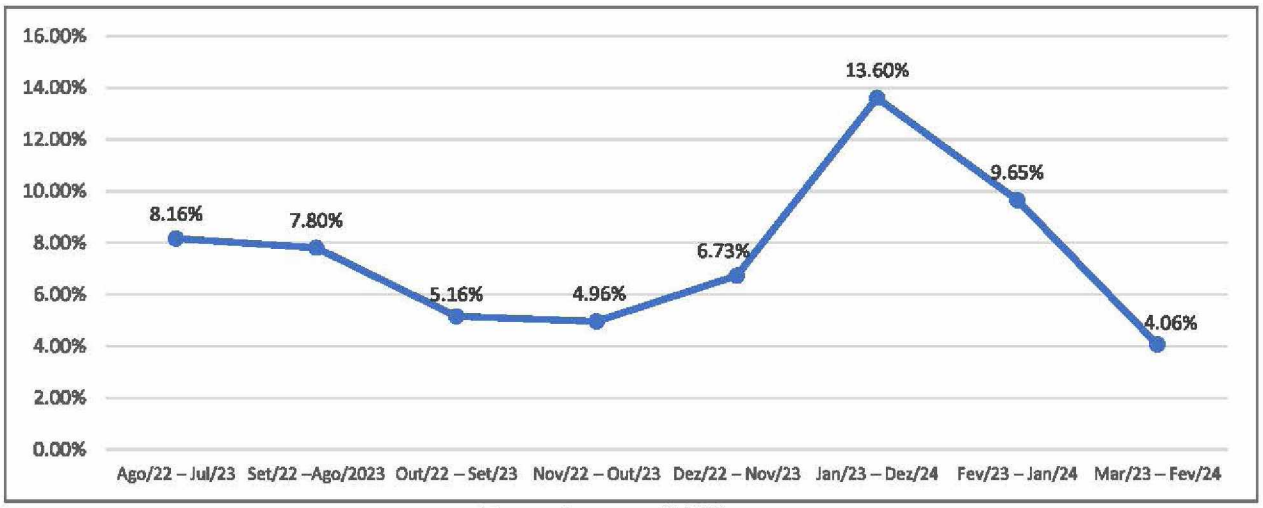
Os resultados do modelo para o horizonte de 12 meses, com diferentes períodos de projeção, estão apresentados na Tabela 10. O MAPE dos erros de cada período projetado, comparado aos valores reais, é exibido no Gráfico 3. Assim como nas projeções de três e seis meses, os erros variaram ao longo das janelas analisadas. As janelas iniciadas em 2022 apresentaram maiores erros, principalmente entre o 6° e o 10° mês projetado. A partir de nov/2022, os valores de erro foram reduzidos. Por exemplo, no período de mar/2023 a fev/2024, todos os meses projetados registraram erro absoluto inferior a 10%. Considerando os valores de MAPE ao longo dos períodos, a maioria dos erros também permaneceu abaixo de 10%, com o maior valor registrado sendo 13,60%.

TABELA 10 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO PCA – 12 MESES

Período Projetado	Meses Projetados (Erro Percentual Absoluto %)
Ago/2022–Jul/2023	Ago: 15,92% / Set: 14,62% / Out: 12,09% / Nov: 12,40% / Dez: 12,54% / Jan: 8,76% / Fev: 4,68% / Mar: 4,70%, Abr: 2,16%, Mai: 0,83%, Jun: 3,65%, Jul: 5,57%
Set/2022–Ago/2023	Set: 4,79% / Out: 2,38% / Nov: 3,10% / Dez: 3,35% / Jan: 0,94% / Fev: 5,84% / Mar: 5,67% / Abr: 8,87% / Mai: 12,49% / Jun: 15,86% / Jul: 17,82% / Ago: 12,52%
Out/2022–Set/2023	Out: 7,35% / Nov: 8,40% / Dez: 8,77% / Jan: 5,02% / Fev: 0,87% / Mar: 1,09% / Abr: 1,57% / Mai: 4,56% / Jun: 7,52% / Jul: 9,67% / Ago: 4,45% / Set: 2,62%
Nov/2022–Out/2023	Nov: 5,46% / Dez: 6,83% / Jan: 4,51% / Fev: 0,19% / Mar: 0,21% / Abr: 2,39% / Mai: 6,06% / Jun: 9,25% / Jul: 11,22% / Ago: 6,32% / Set: 4,87% / Out: 2,19%
Dez/2022–Nov/2023	Dez: 5,42% / Jan: 2,26% / Fev: 2,74% / Mar: 2,88% / Abr: 5,38% / Mai: 8,86% / Jun: 12,25% / Jul: 14,17% / Ago: 8,98% / Set: 7,54% / Out: 5,00% / Nov: 5,29%
Jan/2023–Dez/2024	Jan: 5,28% / Fev: 9,92% / Mar: 9,32% / Abr: 11,91% / Mai: 15,57% / Jun: 19,18% / Jul: 21,33% / Ago: 15,87% / Set: 14,65% / Out: 11,97% / Nov: 12,26% / Dez: 15,89%
Fev/2023–Jan/2024	Fev: 5,73% / Mar: 5,34% / Abr: 8,32% / Mai: 11,85% / Jun: 14,99% / Jul: 16,98% / Ago: 11,28% / Set: 9,27% / Out: 6,29% / Nov: 6,28% / Dez: 9,30% / Jan: 10,17%
Mar/2023–Fev/2024	Mar: 0,34% / Abr: 2,89% / Mai: 5,77% / Jun: 8,61% / Jul: 10,27% / Ago: 5,11% / Set: 3,23% / Out: 0,48% / Nov: 0,49% / Dez: 3,66% / Jan: 4,35% / Fev: 3,45%

Fonte: A autora (2025).

GRÁFICO 3 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 12 MESES - PCA



Fonte: A autora (2025).

Os erros de projeção (MAPE) mês a mês para horizontes de 30 até 350 dias são apresentados na Tabela 11. Diferente das projeções de 3 e 6 meses, observa-se que os menores erros médios de projeção ocorrem nos primeiros meses, especialmente até o terceiro ou quarto mês. A partir de horizontes mais longos, como acima de 180 dias, os erros tendem a aumentar, indicando redução da precisão do modelo com o avanço do tempo de projeção.

TABELA 11 - ERRO MÉDIO PARA PERÍODO DE 30, 60 ATÉ 350 DIAS USANDO PCA

Período	Janela 1	Janela 2	Janela 3	Janela 4	Janela 5	Janela 6	Janela 7	Janela 8	MAPE
30 Dias	15,92%	4,79%	7,35%	5,46%	5,42%	5,28%	5,73%	0,34%	**6,29%**
60 Dias	14,63%	2,38%	8,40%	6,83%	2,26%	9,92%	5,34%	2,89%	**6,58%**
90 Dias	12,09%	3,10%	8,77%	4,51%	2,74%	9,32%	8,32%	5,77%	**6,83%**
120 Dias	12,40%	3,35%	5,02%	0,19%	2,88%	11,91%	11,85%	8,61%	**7,03%**
150 Dias	12,54%	0,94%	0,87%	0,21%	5,38%	15,57%	14,99%	10,27%	**7,60%**
180 Dias	8,77%	5,84%	1,09%	2,39%	8,86%	19,18%	16,98%	5,11%	**8,53%**
210 Dias	4,68%	5,67%	1,57%	6,06%	12,25%	21,33%	11,28%	3,23%	**8,26%**
230 Dias	4,70%	8,87%	4,56%	9,25%	14,17%	15,87%	9,27%	0,48%	**8,40%**
260 Dias	2,16%	12,49%	7,52%	11,22%	8,98%	14,65%	6,29%	0,49%	**7,98%**
290 Dias	0,83%	15,86%	9,67%	6,32%	7,54%	11,97%	6,28%	3,66%	**7,77%**
320 Dias	3,65%	17,82%	4,45%	4,87%	5,00%	12,26%	9,30%	4,35%	**7,71%**
350 Dias	5,57%	12,52%	2,62%	2,19%	5,29%	15,89%	10,17%	3,45%	**7,21%**

Fonte: A autora (2025).

5.2 RESULTADOS DA ANÁLISE FATORIAL COM O ÓLEO BÁSICO DE MAIOR SÉRIE TEMPORAL

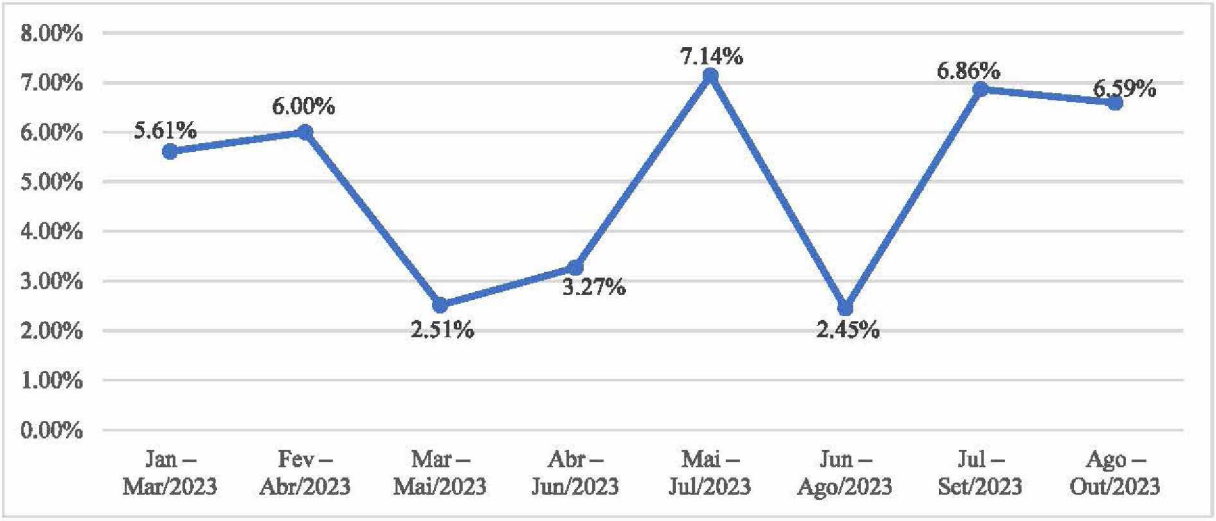
Empregando a mesma metodologia também para a Análise Fatorial, obteve-se os resultados apresentados na Tabela 12 e os MAPE das projeções com horizonte de três meses (Gráfico 4). Cabe relembrar que a Análise Fatorial foi realizada com o banco de dados com 135 variáveis, extraído do banco de dados original de 227 variáveis. Além disso, o teste KMO apresentou valor de 0,87, e o teste de esfericidade de Bartlett resultou em p-valor igual a 0, confirmando a adequação dos dados para a aplicação da Análise Fatorial. Utilizando o critério de Kaiser, o número ideal de fatores permaneceu em nove para todas as janelas de tempo.

TABELA 12 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO AF – 3 MESES

Período de Geração das Componentes	Período Projetado	Meses Projetados (Erro Percentual Absoluto %)
Jan/2010 – Dez/2022	Jan/2023 – Mar/2023	Jan: 2,18% / Fev: 7,29% / Mar: 7,35%
Jan/2010 – Jan/2023	Fev/2023 – Abr/2023	Fev: 5,01% / Mar: 4,95% / Abr: 8,05%
Jan/2010 – Fev/2023	Mar/2023 – Mai/2023	Mar: 1,60% / Abr: 1,43% / Mai: 4,51%
Jan/2010 – Mar/2023	Abr/2023 – Jun/2023	Abr: 0,30% / Mai: 3,20% / Jun: 6,32%
Jan/2010 – Abr/2023	Mai/2023 – Jul/2023	Mai: 4,18% / Jun: 7,48% / Jul: 9,75%
Jan/2010 – Mai/2023	Jun/2023 – Ago/2023	Jun: 2,49% / Jul: 4,72% / Ago: 0,13%
Jan/2010 – Jun/2023	Jul/2023 – Set/2023	Jul: 10,77% / Ago: 5,58% / Set: 4,22%
Jan/2010 – Jul/2023	Ago/2023 – Out/2023	Ago: 7,88% / Set: 6,85% / Out: 5,05%

Fonte: A autora (2025)

GRÁFICO 4 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 3 MESES - AF



Fonte: A autora (2025).

Na Tabela 13 estão evidenciados os erros para 30, 60 e 90 dias para cada janela de tempo utilizada.

TABELA 13 - ERRO MÉDIO PARA PERÍODO DE 30, 60 E 90 DIAS USANDO AF

Período	30 Dias	60 Dias	90 Dias
MAPE	4,30%	5,19%	5,67%

Fonte: A autora (2025).

Assim como na Análise das Componentes Principais (PCA), os resultados da Análise Fatorial demonstram boa acurácia, o que reforça a eficácia das técnicas multivariadas para fins de previsão (Tabela 14).

TABELA 14 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO AF – 6 MESES

Período Projetado	**Meses Projetados (Erro Percentual Absoluto %)**
Jan/2023– Jun/2023	Jan: 2,43% / Fev: 7,24% / Mar: 6,52% / Abr: 9,23% / Mai: 12,22% / Jun: 15,49%
Fev/2023– Jul/2023	Fev: 4,95% / Mar: 4,69% / Abr: 7,53% / Mai: 10,55% / Jun: 13,46% / Jul: 15,68%
Mar/2023– Ago/2023	Mar: 1,73% / Abr: 0,62% / Mai: 3,00% / Jun: 5,76% / Jul: 7,87% / Ago: 2,77%
Abr/2023– Set/2023	Abr: 0,10% / Mai: 2,52% / Jun: 5,62% / Jul: 7,53% / Ago: 2,29% / Set: 1,16%
Mai/2023– Out/2023	Mai: 4,26% / Jun: 7,37% / Jul: 9,63% / Ago: 4,29% / Set: 2,92% / Out: 0,51%
Jun/2023– Nov/2023	Jun: 3,05% / Jul: 5,27% / Ago: 0,08% / Set: 1,08% / Out: 3,27% / Nov: 2,93%
Jul/2023 – Dez/2023	Jul: 7,01% / Ago: 1,31% / Set: 0,14% / Out: 2,46% / Nov: 2,52% / Dez: 0,17%
Ago/2023 – Jan/2024	Ago: 7,10% / Set: 5,97% / Out: 3,89% / Nov: 4,15% / Dez: 7,26% / Jan: 8,29%

Fonte: A autora (2025).

O Gráfico 5 ilustra a média dos erros apresentados na Tabela 12. Nota-se que, em cinco das oito janelas de tempo, os resultados foram considerados bons, com valores de MAPE inferiores a 5%, especialmente nos modelos mais recentes, como os treinados até junho e julho de 2023, que apresentaram médias de erro de 2,61% e 2,27%, respectivamente. A média de erros para os horizontes de 30 e 60 dias foi de 3,83% e 4,37%, respectivamente, mantendo-se abaixo de 5% na maior parte das janelas. Para horizontes mais longos, como 180 dias, o erro médio foi ligeiramente superior (5,88%).

GRÁFICO 5 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 6 MESES - AF

Fonte: A autora (2025).

Seguindo a mesma lógica dos casos anteriores, a Tabela 15 mostra os erros para os períodos de 30 à 180 dias.

TABELA 15 - ERRO MÉDIO PARA PERÍODO DE 30, 60, 90, 150 E 180 DIAS USANDO AF

Período	30 Dias	60 Dias	90 Dias	120 Dias	150 Dias	180 Dias
MAPE	3,83%	4,37%	4,55%	5,63%	6,48%	5,88%

Fonte: A autora (2025).

Na Tabela 16 estão os erros de projeção para o período de 12 meses, usando Análise Fatorial.

TABELA 16 - CÁLCULO DO ERRO DE PROJEÇÃO USANDO PCA – 12 MESES

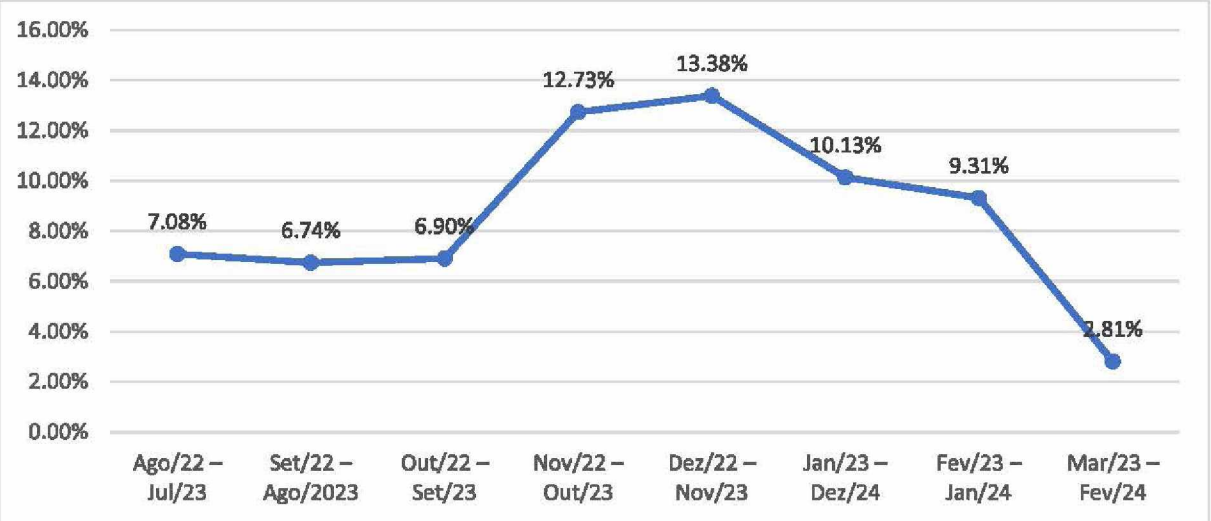
Período Projetado	**Meses Projetados (Erro Percentual Absoluto %)**
Ago/2022–Jul/2023	Ago: 14,19% / Set: 12,63% / Out: 10,02% / Nov: 10,50% / Dez: 10,65% / Jan: 6,74% / Fev: 2,88% / Mar: 3,43% / Abr: 0,71% / Mai: 1,92% / Jun: 4,67% / Jul: 6,59%
Set/2022–Ago/2023	Set: 6,49% / Out: 3,86% / Nov: 4,02% / Dez: 4,13% / Jan: 0,07% / Fev: 4,34% / Mar: 4,09% / Abr: 6,66% / Mai: 9,74% / Jun: 13,17% / Jul: 15,18% / Ago: 9,19%
Out/2022–Set/2023	Out: 5,26% / Nov: 4,84% / Dez: 5,14% / Jan: 1,08% / Fev: 3,67% / Mar: 3,90% / Abr: 6,17% / Mai: 9,15% / Jun: 12,64% / Jul: 14,59% / Ago: 8,91% / Set: 7,49%

Nov/2022–Out/2023	Nov: 1,38% / Dez: 1,33% / Jan: 6,37% / Fev: 11,34% / Mar: 11,01% / Abr: 13,83% / Mai: 17,39% / Jun: 21,13% / Jul: 23,01% / Ago: 17,09% / Set: 15,70% / Out: 13,24%
Dez/2022–Nov/2023	Dez: 1,27% / Jan: 6,02% / Fev: 11,26% / Mar: 10,85% / Abr: 13,76% / Mai: 16,91% / Jun: 20,37% / Jul: 22,28% / Ago: 16,42% / Set: 15,10% / Out: 13,06% / Nov: 13,31%
Jan/2023–Dez/2024	Jan: 2,76% / Fev: 7,55% / Mar: 6,84% / Abr: 9,56% / Mai: 12,55% / Jun: 15,81% / Jul: 17,61% / Ago: 11,97% / Set: 10,40% / Out: 7,90% / Nov: 7,76% / Dez: 10,91%
Fev/2023–Jan/2024	Fev: 5,23% / Mar: 4,97% / Abr: 7,81% / Mai: 10,89% / Jun: 13,80% / Jul: 16,00% / Ago: 10,61% / Set: 9,14% / Out: 6,44% / Nov: 6,52% / Dez: 9,74% / Jan: 10,54%
Mar/2023–Fev/2024	Mar: 1,46% / Abr: 0,86% / Mai: 3,26% / Jun: 6,03% / Jul: 8,13% / Ago: 3,02% / Set: 1,79% / Out: 0,21% / Nov: 0,03% / Dez: 2,84% / Jan: 3,64% / Fev: 2,44%

Fonte: A autora (2025).

A média dos erros de projeções estão elencados no Gráfico 6 e os erros para o período de 30 à 350 dias na Tabela 17.

GRÁFICO 6 - MÉDIA DOS ERROS DE PROJEÇÕES PARA 12 MESES - AF



Fonte: A autora (2025).

TABELA 17 - ERRO MÉDIO PARA PERÍODO DE 30, 60 ATÉ 350 DIAS USANDO AF

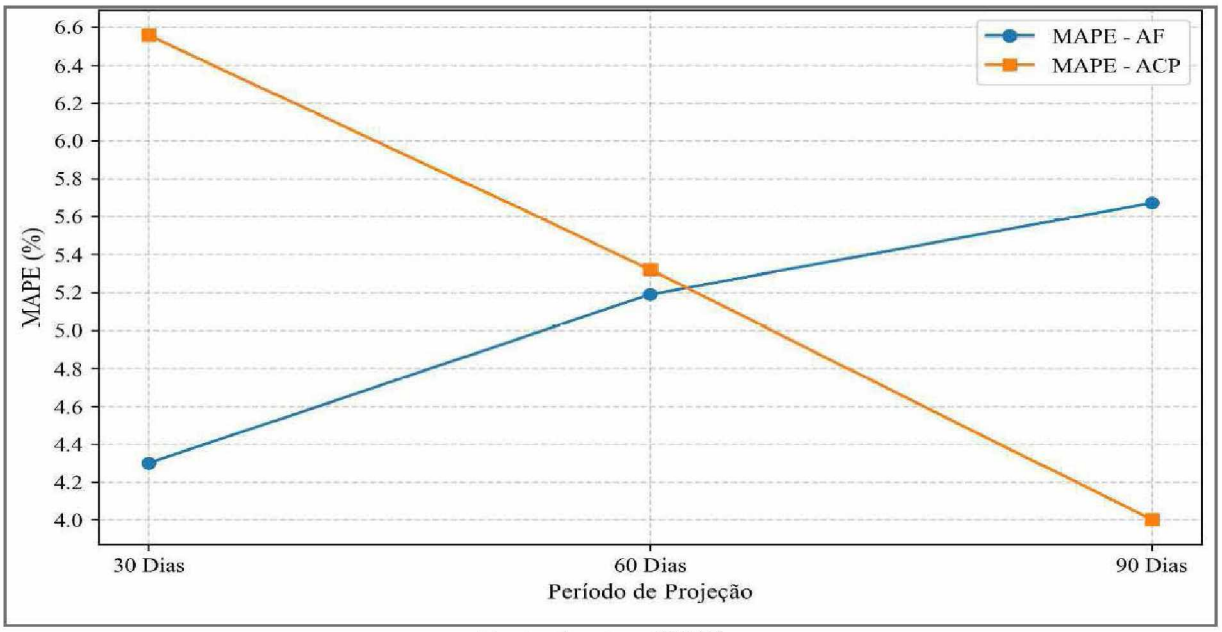
Período (Dias)	**30**	**60**	**90**	**120**	**150**	**180**	**210**	**230**	**260**	**290**	**320**	**350**
MAPE (%)	4,75	5,26	6,84	8,05	9,20	10,07	10,11	10,50	9,92	9,89	9,83	9,21

Fonte: A autora (2025).

Através dos testes realizados para o Óleo Básico de maior sério temporal, pode-se verificar, a partir, dos valores de MAPE por período projetado, que a Análise Fatorial (AF) apresentou melhor acurácia para projeções de curto prazo. Por exemplo, para projeções de três meses, no horizonte de 30 dias, a AF apresentou menor erro (4,30%) em relação à PCA (6,56%). No período de 60 dias, os resultados foram semelhantes, com leve vantagem para a AF (5,19%) em relação à PCA (5,32%). Já em 90 dias, a PCA obteve menor erro (4,00%) em comparação à AF (5,67%). Esses resultados indicam que a AF apresenta melhor desempenho nas primeiras projeções, enquanto a PCA é mais adequada para horizontes mais longos.

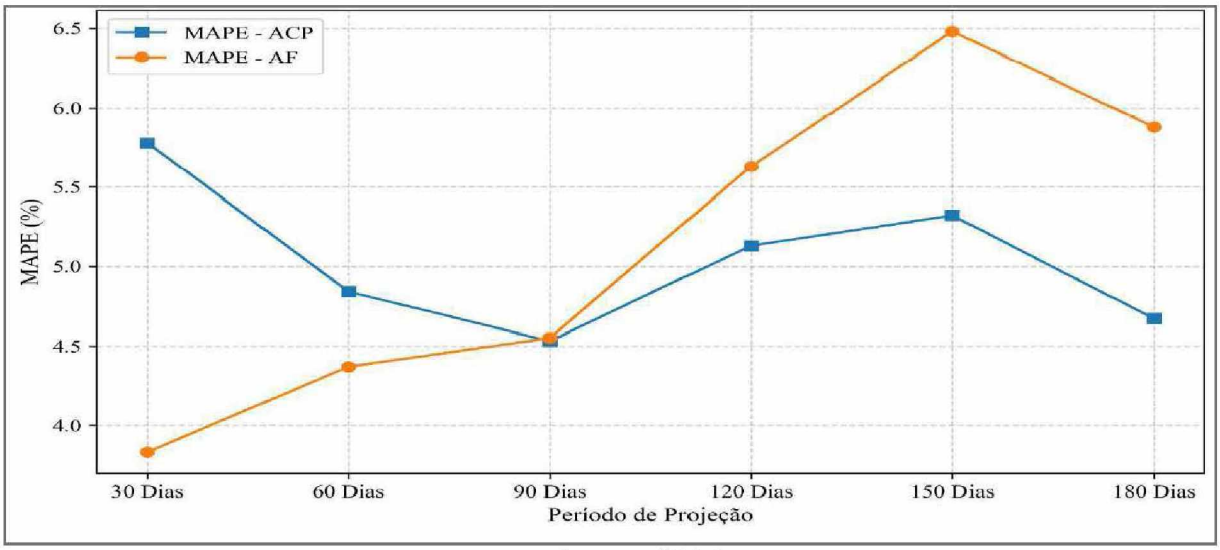
Adicionalmente, o mesmo ocorre com as projeções de seis e doze meses, sendo o modelo baseado em Análise das Componentes Principais (PCA) o que apresenta maior estabilidade e menor crescimento do erro ao longo do tempo, tornando-se mais adequado para projeções de médio e longo prazo. Nota-se que neste caso há maior estabilidade e menor crescimento dos erros de projeção ao longo do tempo. Os gráficos 7, 8 e 9 ilustram o comportamento desses erros, para os horizontes analisados.

GRÁFICO 7 - COMPARAÇÃO DOS ERROS MÉDIOS DE 30 A 90 DIAS OBTIDOS PELA PCA E AF (HORIZONTE DE 3 MESES)



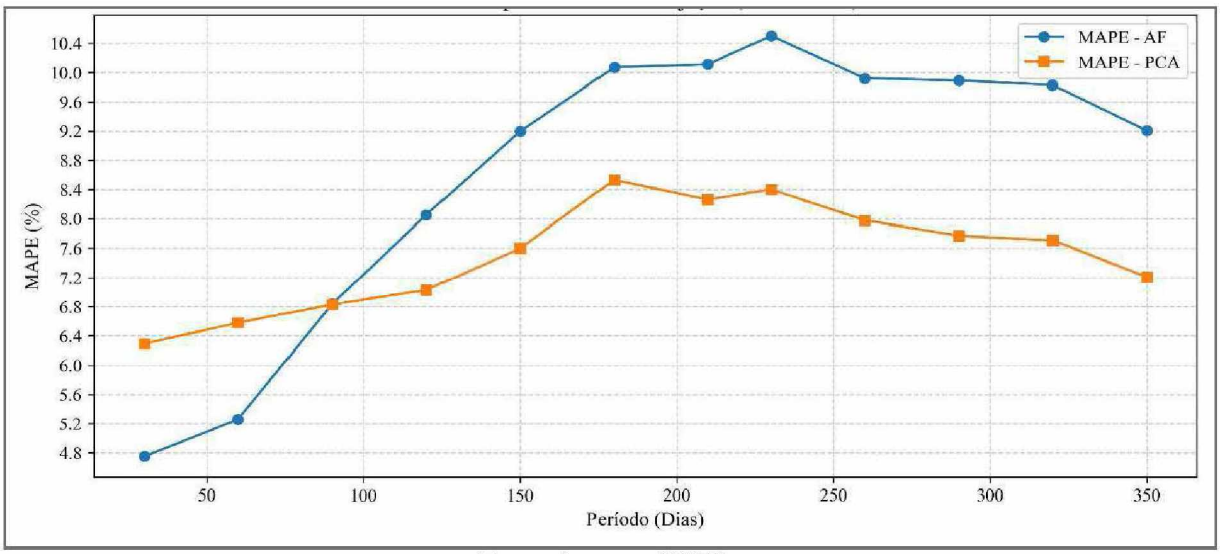
Fonte: A autora (2025).

GRÁFICO 8 - COMPARAÇÃO ERROS MÉDIOS DE 30 À 180 DIAS OBTIDOS PELA PCA E AF (HORIZONTE DE 6 MESES)



Fonte: A autora (2025).

GRÁFICO 9 - COMPARAÇÃO ERROS MÉDIOS DE 30 À 180 DIAS OBTIDOS PELA PCA E AF (HORIZONTE DE 12 MESES)



Fonte: A autora (2025).

É importante destacar que a AF foi aplicada sobre um subconjunto de 135 variáveis focadas no mercado financeiro, como índices de bolsa, paridade de moedas, ações de empresas de energia e preços de diferentes tipos de petróleo. Enquanto isso, a PCA considerou um conjunto mais amplo, com 227 variáveis que incluíram, além das anteriores, informações relacionadas à oferta, demanda e estoques de petróleo e derivados, capacidade de refino por continente, commodities minerais e energéticas, índices setoriais Dow Jones, paridade cambial global, índices de bolsas internacionais e indicadores de liquidez mundial. A diferença do banco

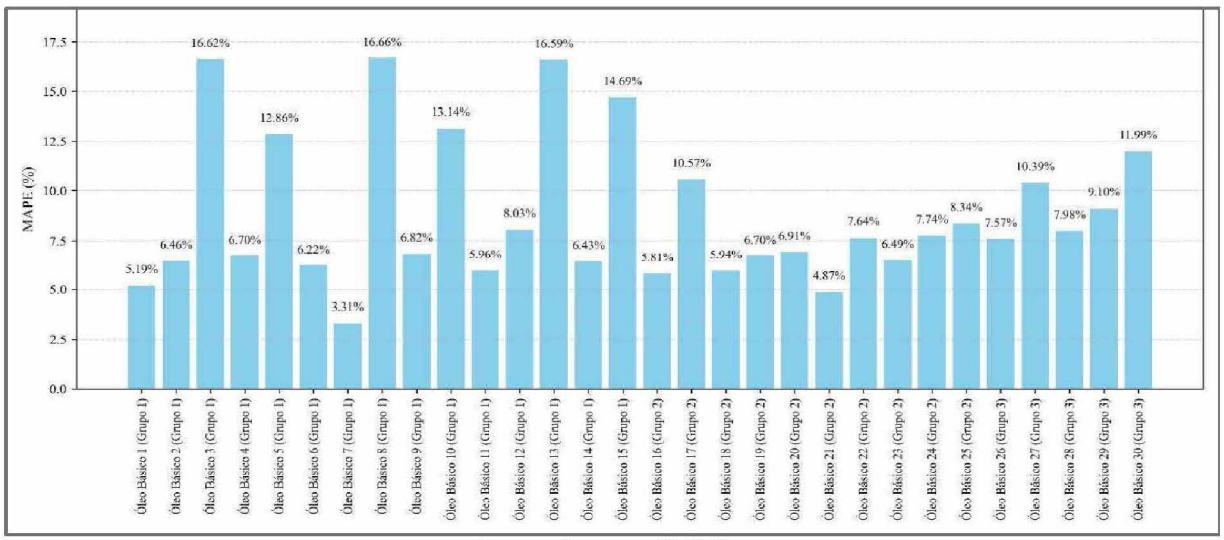
de dados pode ter influenciado diretamente o comportamento dos modelos: enquanto a AF mostrou-se eficiente na previsão de curto prazo do mercado financeiro, a PCA, ao adicionar informações estruturais e setoriais a mais, apresentou maior acurácia para horizontes temporais mais longos.

5.3 RESULTADOS OBTIDOS PARA OS 30 ÓLEOS BÁSICOS PARA HORIZONTES DE 3, 6 E 12 MESES

Como dito no capítulo anterior, os demais Óleos Básicos, com série mensal menor, não apresentaram resultados interessantes com o banco de 227 variáveis, seja com ou sem aplicação da PCA, assim como também com o banco de 135 variáveis com a utilização da Análise Fatorial. Em contrapartida, com o banco de 71 variáveis com o uso da Análise das Componentes Principais, os resultados, para a maioria dos Óleos Básicos foram satisfatórios.

As projeções e as janelas de tempo utilizadas foram as mesmas do Óleo de maior série temporal, a diferença está no início do período inicial dos dados, sendo janeiro de 2015. Assim, a partir das médias de MAPE obtidas para cada um dos 30 Óleos Básicos, observou-se que para o horizonte de três meses, 22 apresentaram MAPE abaixo de 10%, indicando bom desempenho do modelo baseado em PCA. No Grupo 1, embora a maioria tenha registrado erros baixos, quatro óleos tiveram MAPE acima de 14%, possivelmente devido à maior volatilidade dos preços ou à menor adequação das variáveis explicativas. No Grupo 2, um óleo ultrapassou 10%, enquanto os demais apresentaram erros mais controlados. Já os óleos do Grupo 3 tiveram desempenho intermediário, com MAPE entre 7% e 12% (Gráfico 10).

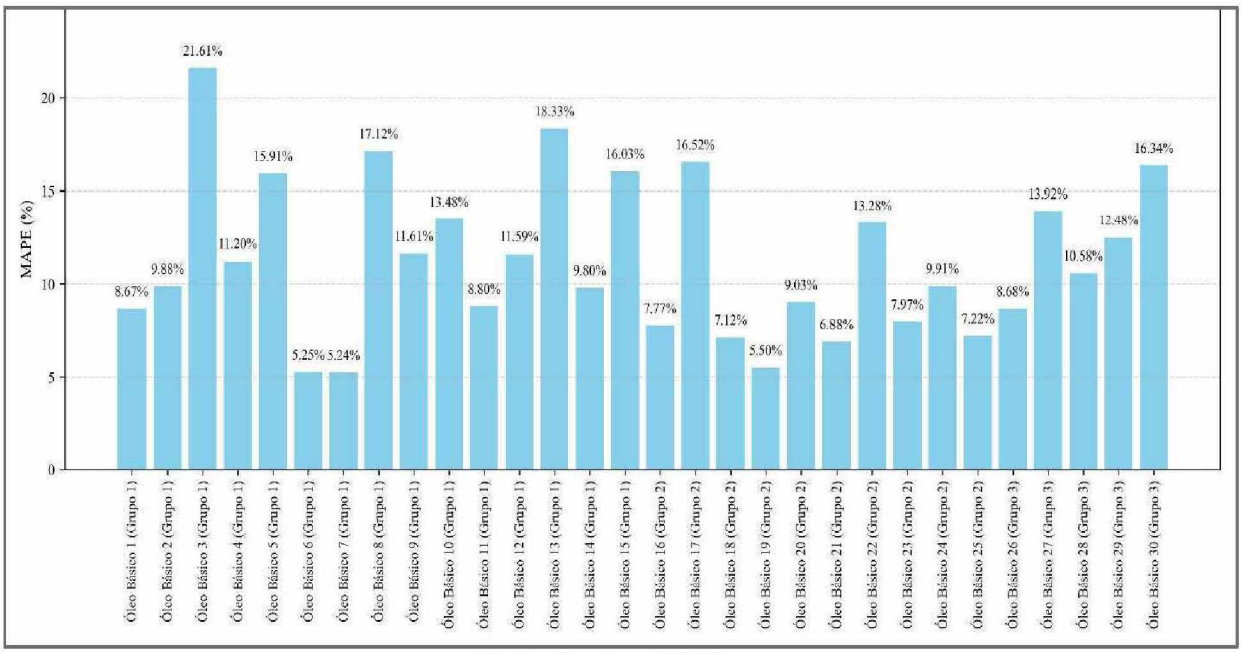
GRÁFICO 10 - MAPE POR ÓLEO BÁSICO – PREVISÃO INIDIVIDUAL (HORIZONTE DE 3 MESES)



Fonte: A autora (2025).

Os resultados obtidos para as projeções de seis meses estão elencados no Gráfico 11. Observa-se que, ao aumentar o período de projeção, o MAPE para todos os 30 Óleos Básicos também aumentou quando comparado com o horizonte de três meses. Para este caso, 15 óleos apresentaram erros abaixo de 10%. Além disso, em grande parte dos casos, a tendência das séries se manteve semelhante à observada nas projeções de três meses.

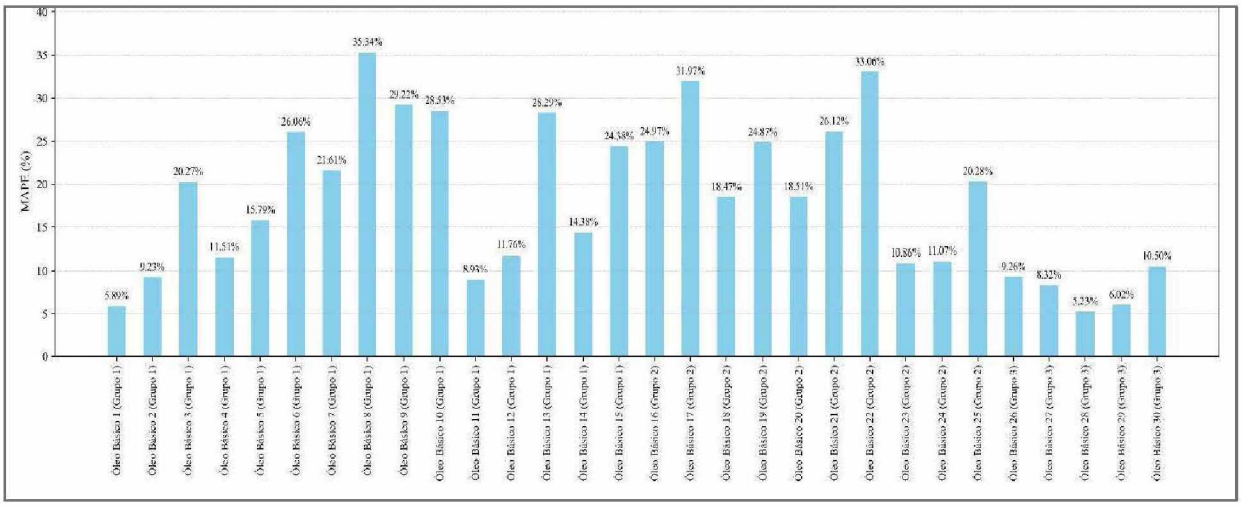
GRÁFICO 11 - MAPE POR ÓLEO BÁSICO – PREVISÃO INIDIVIDUAL (HORIZONTE DE 6 MESES)



Fonte: A autora (2025).

Os resultados para o horizonte de 12 meses, apresentados no Gráfico 12, revelam um aumento expressivo do MAPE em relação às previsões de curto e médio prazo. Apenas 7 Óleos Básicos apresentaram erros abaixo de 10%, enquanto cerca de 30% dos casos registraram MAPE acima de 20%. Esse comportamento reforça o efeito do horizonte sobre a acurácia do modelo, indicando que previsões mais longas tendem a ser menos confiáveis. Ainda assim, uma parcela dos óleos manteve bom desempenho, sugerindo que, embora o modelo tenha limitações, ele pode ser útil para projeções de longo prazo em casos específicos.

GRÁFICO 12 - MAPE POR ÓLEO BÁSICO – PREVISÃO INIDIVIDUAL (HORIZONTE DE 12 MESES)



Fonte: A autora (2025).

5.4 VARIÁVEIS DE IMPACTO POSITIVO E NEGATIVO PARA CADA ÓLEO BÁSICO

A verificação do impacto das variáveis, seja positivo (indicando aumento no preço) ou negativo (indicando queda), foi realizada por meio da aplicação da Análise das Componentes Principais (PCA) sobre o banco de dados com 71 variáveis, tanto para o óleo com maior série histórica quanto para os demais.

Como discutido no capítulo anterior, os *loadings* das Componentes Principais e os coeficientes das equações de regressão serviram como base para esta análise. Abaixo, apresenta-se uma das equações de regressão obtidas (em base logarítmica) para um Óleo Básico do Grupo 1, com projeção de três meses:

Óleo Básico Grupo 1 = 3.0550 + (0.1034)*PC1 + (0.1784)*PC2 + (-0.1115)*PC3 + (-0.0590)*PC4 + (0.2855)*PC5 + (0.0141)*PC6 + (-0.1198)*PC7 + (0.0945)*PC8 + (0.0632)*PC9 + (0.0113)*PC10 + (-0.3390)*PC11 + (-0.0945)*PC12 + (0.0604)*PC13 + (-0.0005)*PC14

Para calcular o impacto individual de cada variável original sobre o preço do óleo, foi adotado o seguinte procedimento:

- Inicialmente, foram considerados apenas aqueles superiores a 0,09 ou inferiores a −0,09, de forma a garantir significância na associação entre a variável e a Componente Principal.
- Verificou-se se a mesma variável mantinha a relação com a respectiva Componente ao longo de todas as janelas de tempo analisadas.

- Como para cada janela de tempo foi estimada uma equação de regressão distinta para cada óleo básico, avaliou-se a combinação entre o sinal do loading e o sinal do coeficiente da Componente na equação. Assim, quando ambos os sinais eram positivos ou negativos, o impacto da variável sobre o preço foi classificado como positivo; quando apresentavam sinais opostos, o impacto foi considerado negativo.

Esse processo foi aplicado a todas as variáveis do banco de dados. Ainda com relação ao Óleo da equação anterior, a Tabela 14 a seguir ilustra as primeiras linhas da matriz de *loadings* obtida para uma janela de projeção:

GRÁFICO 13 - *LOADINGS* DE UM ÓLEO BÁSICO DO GRUPO 1

Variável Original	PC1	PC2	PC3	⋯	PC14
x_1	0,084875	-0,02388	0,023427	⋯	0,117643
x_2	0,066976	-0,00443	-0,06082	⋯	0,000242
x_3	0,134313	-0,00908	0,07837	⋯	0,011343
⋮	⋮	⋮	⋮	⋯	⋯
x_{71}	0,137745	-0,06997	0,107407	⋯	-0,08612

Fonte: A autora (2025).

Nesta tabela, cada linha representa uma variável original (com os nomes omitidos por confidencialidade), e cada célula indica seu *loading* em relação à respectiva Componentes Principal. Para este caso, a variável x_1 tem maior relação com a componente 14. Ao avaliar o impacto dessa variável na equação de regressão apresentada anteriormente, o efeito seria negativo, porque o *loading* é positivo (0,1176) e o coeficiente negativo (−0,0005).

Esse mesmo raciocínio foi aplicado a todas as variáveis e em todas as janelas de tempo. Como, neste caso, o efeito se manteve consistente ao longo dos testes, o impacto final foi consolidado como negativo.

A análise foi realizada em todas janelas com horizontes de 3 e 6 e 12 meses. Observou-se que algumas variáveis mantiveram um impacto consistente (positivo ou negativo) ao longo dos diferentes períodos, enquanto outras apresentaram mudanças. Desta forma, foram selecionadas as variáveis que apresentaram impacto dominante mais frequente ao longo das janelas temporais consideradas.

A análise dos setores mais frequentes identificou padrões comuns nos fatores que impactam os preços dos Óleos Básicos, com variações apenas na frequência e intensidade entre os grupos analisados.

No Grupo 1, as variáveis mais recorrentes estão associadas aos setores de combustíveis e refino, minerais e metais estratégicos, mercados asiáticos e energia renovável. Além das variáveis ligadas ao mercado de combustíveis líquidos, como o ÁsiaPack Fuel Oil, foram identificadas influências de minerais como lítio e zinco, bem como de indicadores econômicos de países emergentes, representados, por exemplo, pelo índice NIFYT 50 (Índia). Por outro lado, também se verificou a presença de variáveis com impacto negativo, associadas principalmente à atividade de refino na região Ásia-Pacífico, como a produção de gasolina, GLP e a taxa de utilização das refinarias, além da produção de GLP na América do Norte, do índice Dow Jones de empresas de semicondutores e do desempenho das ações da *Reliance Industries Limited* (Índia).

No Grupo 2, os setores mais frequentes mantêm perfil semelhante ao Grupo 1, com destaque para combustíveis e derivados e minerais/metais. Contudo, para associações positivas, observou-se outras variáveis como as ações da Neste Oyj (Finlândia) e da Vestas Wind Systems (Dinamarca), ligadas à energia renovável, bem como os índices acionários Tadawul All Share (Arábia Saudita) e VN 30 (Vietnã). Adicionalmente, as variáveis de impacto negativo, além das apresentadas referente ao grupo 1, estavam associadas a indicadores econômicos e financeiros, como o índice da Bolsa de Budapeste (Hungria), o Shanghai Shenzhen CSI 300 (China) e as ações Innovation Co. (Coreia do Sul).

No Grupo 3, a diversidade setorial foi menor. A influência de fatores como o lítio, o zinco e os índices acionários asiáticos mantiveram-se presente. Contudo, surgiram novas associações positivas ligadas ao setor energético europeu e à indústria de bens de consumo duráveis, como as ações da Acea Energia e da Hera S.p.A. (Itália), ambas do segmento de energia e serviços ambientais, além dos índices *Dow Jones Automobiles e Dow Jones Footwear*. As variáveis com impacto negativo permaneceram concentradas em indicadores de refino, oferta de derivados e em mercados financeiros emergentes, representados pelos índices BIST 100 (Turquia), Budapest SE (Hungria) e Shanghai Shenzhen CSI 300 (China).

A repetição dos setores com maior impacto entre os três grupos pode ser atribuída à composição do banco de dados utilizado. Com 71 variáveis, o conjunto de dados é composto majoritariamente por informações sobre energia, commodities e indicadores macroeconômicos, o que contribui para uma concentração nos mesmos fatores de influência. Nesse contexto, as

diferenças observadas entre os grupos referem-se principalmente à frequência e ao peso relativo de cada setor.

6 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

O presente trabalho teve como objetivo desenvolver um modelo de Regressão Múltipla baseado em técnicas de Análise Estatística Multivariada, aplicadas à previsão de preços de Óleos Lubrificantes Básicos e à identificação das variáveis com maior influência sobre essas variações. Para isto, utilizou-se um banco de dados composto por 227 variáveis de diferentes setores.

As análises realizadas demonstraram que o uso combinado de métodos de redução de dimensionalidade, como a Análise das Componentes Principais (PCA) e a Análise Fatorial (AF), aliado às projeções em diferentes janelas de tempo usando ARIMA, contribuiu para o tratamento de bases de dados complexas, com grande número de variáveis inter-relacionadas, permitindo a obtenção de previsões e interpretações quantitativas das relações entre os fatores explicativos e as variáveis dependentes.

A integração de Técnicas Multivariadas com Séries Temporais permitiu formar novos bancos de dados, o que proporcionou melhores resultados. Com a verificação das variáveis que atendem aos critérios para aplicação da Análise Fatorial, foi possível compor um banco com 135 variáveis. Já com o uso das Análises de Cluster e Discriminante, obteve-se um grupo com 71 variáveis. Ambos os bancos mostraram-se válidos para os diferentes tipos de verificações.

Os resultados obtidos com o óleo básico de maior série temporal mostraram que tanto a PCA quanto a AF alcançaram baixos erros médios percentuais absolutos (MAPE), em sua maioria inferiores a 10%, o que indica boa capacidade de ajuste e previsão. Observou-se que a AF apresentou melhor desempenho para horizontes de curto prazo, especialmente até três meses, com menores valores médios de erro em comparação à PCA. Por outro lado, a PCA demonstrou maior estabilidade e menor crescimento dos erros ao longo do tempo, sendo mais adequada para horizontes de médio e longo prazo. Essa diferença de comportamento reflete, em parte, a natureza distinta das bases de dados utilizadas: a AF foi aplicada a um subconjunto de 135 variáveis predominantemente financeiras, enquanto a PCA considerou um conjunto mais abrangente de 227 variáveis, incluindo indicadores de oferta e demanda de petróleo, capacidade de refino, commodities energéticas e minerais, índices de bolsas internacionais e paridades cambiais. Essa maior diversidade de informações permitiu à PCA capturar de forma mais robusta as variações estruturais do mercado.

Já com o banco de 71 variáveis com a aplicação da PCA, foi possível obter melhores resultados para os outros 30 óleos básicos. Os resultados evidenciaram que, para o horizonte de três meses, a maioria apresentou MAPE inferior a 10%, demonstrando desempenho satisfatório

do modelo baseado em PCA. Com o aumento do horizonte de previsão para seis e doze meses, observou-se crescimento gradual do erro, o que era esperado em função da maior incerteza associada a projeções temporais mais longas. Ainda assim, uma parcela dos produtos manteve erros abaixo de 10%, indicando que o modelo conserva utilidade preditiva mesmo em prazos estendidos.

A análise das variáveis com impacto positivo e negativo permitiu identificar padrões setoriais comuns entre os grupos de óleos básicos. As variáveis associadas aos setores de combustíveis e refino, minerais estratégicos e índices financeiros asiáticos foram as mais recorrentes. Destacaram-se, entre os impactos positivos, os índices e ações ligados aos mercados asiáticos e à energia renovável, como Acea Energia, Neste Oyj, Vestas Wind Systems, Tadawul All Share (Arábia Saudita) e VN 30 (Vietnã). Entre os impactos negativos, prevaleceram variáveis relacionadas à atividade de refino, à produção de derivados e a índices de bolsas emergentes, como o BIST 100 (Turquia), Budapest SE (Hungria) e Shanghai Shenzhen CSI 300 (China). Essa recorrência entre os grupos indica que os preços dos óleos básicos são influenciados por fatores macroeconômicos e setoriais interdependentes, refletindo a integração entre os mercados de energia, metais e capitais.

Os resultados obtidos confirmam que o modelo desenvolvido representa de forma consistente a relação entre variáveis financeiras, energéticas e industriais e os preços dos óleos lubrificantes básicos. A metodologia proposta, baseada na combinação entre modelagem de Séries Temporais e Análise Multivariada, mostrou-se adequada e aplicável a diferentes conjuntos de dados. Sua implementação em linguagem Python e integração com bases estruturadas permitem a automação dos testes e a replicação dos resultados, favorecendo o uso do modelo em processos de apoio à decisão. Além disso, a combinação entre essas técnicas constitui uma contribuição original, pois não foram encontrados estudos semelhantes na literatura consultada.

Em síntese, o estudo mostrou que o uso de Técnicas Multivariadas permite reduzir a dimensionalidade de bases de dados e identificar variáveis que influenciam os preços de mercado. Conclui-se que a abordagem adotada contribui para o entendimento das interações entre fatores globais que afetam o setor de Óleos Básicos, auxiliando na análise e previsão de preços.

Sugere-se, para trabalhos futuros, a integração da Análise Multivariada com métodos híbridos, como redes neurais e decomposição por wavelets, com o objetivo de aprimorar o desempenho e a capacidade de modelagem de relações não lineares. Também se propõe a ampliação do banco de dados, priorizando variáveis de acesso público e com séries temporais

mais recentes, além da exclusão daquelas provenientes de fontes privadas. Essa abordagem permitirá maior transparência e possibilitará a replicação dos resultados por diferentes setores e pesquisadores.

REFERÊNCIAS

ALMEIDA, R. A. **Estudo da recuperação de óleos lubrificantes minerais usados utilizando solventes polares**. 2011. Dissertação (Mestrado em Engenharia Química) – Universidade Federal de Campina Grande, Centro de Ciências e Tecnologia, Unidade Acadêmica de Engenharia Química, Campina Grande, 2011.

AL-ZAHRANI, S. M.; PUTRA, M. D. Used lubricating oil regeneration by various solvent extraction techniques. **Journal of Industrial and Engineering Chemistry**, v. 19, n. 2, p. 536–539, 2013.

ALALI, Z. H.; HORNE, R. N. A comparative study of deep learning models and traditional methods in forecasting oil production in the Volve Field. In: SPE Annual Technical Conference and Exhibition, 2023, San Antonio. Texas: Society of Petroleum Engineers, 2023. Disponível em: https://doi.org/10.2118/214881-MS. Acesso em: 26 fev. 2025.

ALVES, V. **Avaliação de imóveis urbanos baseada em métodos estatísticos multivariados**. 2005. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal do Paraná, Campo Mourão, 2005.

ALQUIST, R.; KILIAN, L. What do we learn from the price of crude oil futures? **Journal of Applied Econometrics**, v. 25, p. 539–573, 2010.

AUSTRALIAN COMPETITION TRIBUNAL. Re: AGL Cooper Basin Natural Gas Supply Arrangements. ATPR 41-593, 1997.

AYUB, D. **Análise preditiva da eficiência global do equipamento: uma abordagem multivariada**. Curitiba: Universidade Federal do Paraná, 2019. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal do Paraná, Curitiba, 2019.

BAJOTTO, A. C. **Validação da tipologia geoquímica das rochas vulcânicas do Grupo Serra Geral no estado do Paraná, por meio de técnicas de estatística multivariada**. 2025. Tese (Doutorado) – Universidade Federal do Paraná, Curitiba, 2025.

BRAULIO, S. N. **Proposta de uma metodologia para a avaliação de imóveis urbanos baseada em métodos estatísticos multivariados**. 2005. Dissertação (Mestrado em Ciências) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2005.

BOX, G. E. P.; JENKINS, G. M. **Time series analysis: forecasting and control**. San Francisco: Holden-Day, 1970.

BRASIL. Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP). **Resolução ANP nº 911, de 2022**. Estabelece as especificações e os requisitos aplicáveis aos óleos básicos no território nacional. Diário Oficial da União, Brasília, 2022.

CARRETEIRO, R. P.; BELMIRO, P. N. A. **Lubrificantes e Lubrificação Industrial**. Editora Interciência; Rio de Janeiro, 2008.

CAVALCANTI, S. L. L. **Caracterização do óleo de carnaúba para uso como biolubrificante**. 2014. Dissertação (Mestrado em Engenharia Química) – Universidade Federal do Rio Grande do Norte, Natal, 2014.

CHAVES NETO, A. **Análise multivariada aplicada à pesquisa**. Notas de aula. Curitiba: Universidade Federal do Paraná, 2025.

CHIAM, L.; AHUJA, V. Long-term supply contracts: time for review. Australian Competition Tribunal. Caso: **Re AGL Cooper Basin Natural Gas Supply Arrangements**. ATPR 41-593, 1997.

CRIVISQUI, M. **Analyse factorielle et classification**. Paris: Technip, 1993.

DAVE, Kushal; LAWRENCE, Steve; PENNOCK, David M. **Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews**. *In:* Proceedings of the 12th International Conference on World Wide Web (WWW '03), 2003. DOI: 10.1145/775152.775226.

DING, Z., *et al.* A State-of-the-Art Survey on Reconfigurable Intelligent Surface-Assisted Non-Orthogonal Multiple Access Networks. **Proceedings of the IEEE**, 110, 1358-1379.

DINNICK, P. Re: AGL Cooper Basin natural gas supply arrangements: application for a review of a determination of the Australian Competition and Consumer Commission. **AMPU Case Notes**, v. 16, p. 258, 1997.

DU, X; TANG, Z; CHEN, K. A novel crude oil futures trading strategy based on time-frequency decomposition with ensemble deep reinforcement learning. **Energy**, v. 285, 2023. Disponível em: https://doi.org/10.1016/j.energy.2023.129394.

DYM RESOURCES. **Base Oil Market Trends and Challenges**. Disponível em: https://dymresources.com/news/base-oil_and_lubes/base-oil-market-in-2023-challenges-trends-key-market-drivers/. Acesso em: 05 fev. 2025.

EISENBEIS, R. Pitfalls in the application of discriminant analysis in business, finance and economics. **The Journal of Finance**, v. XXXII, n. 3, p. 875–900, 1997.

STACHOWIAK, G. W.; BATCHELOR, A. W. **Engineering Tribology**. 3. ed. Oxford: Butterworth-Heinemann, 2005.

FACHEL, J. M. G. **Análise Fatorial**. 1976. Dissertação (Mestrado em Estatística) – Universidade de São Paulo, São Paulo, 1976.

FIGUEIREDO, E. *et al.* Swarm intelligence for clustering—A systematic review with new perspectives on data mining. **Engineering Applications of Artificial Intelligence**, v. 82, p. 313–329, 2019.

FREITAS, R. V. *et al.* Remoção de metais em óleos lubrificantes usados utilizando argila ativada. In: **Anais do 2° Congresso Brasileiro de P&D em Petróleo e Gás**. Rio de Janeiro, 2003.

GARTNER. 2014. **BI: Analytics Moves To The Core.** Digital Business and Business Analytics – Timo Elliott's Blog. < https://timoelliott.com/blog/2013/02/gartnerbi-emea-2013-part-1-analytics-moves-to-the-core.html>. Acesso em 10 mar 2018.

GONTIJO, C.; AGUIRRE, A. Elementos para uma tipologia do uso do solo agrícola no Brasil. **Revista Brasileira de Economia**, v. 42, n. 1, p. 13–49, 1988.

GOUVÊA, M. A.; PREARO, L. C.; ROMEIRO, M. C. Avaliação da adequação de aplicação de técnicas multivariadas em estudos do comportamento do consumidor em teses e dissertações de duas instituições de ensino superior. **R.Adm.**, São Paulo, v. 47, n. 2, p. 338–355, abr./maio/jun. 2012. DOI: 10.5700/rausp1043.

GUERREIRO, M. T. **Análise de métodos de agrupamento de dados para detecção de anomalias na precificação e categorização de peças da indústria automotiva**. 2021. Dissertação (Mestrado em Ciência da Computação) – UTFPR, Ponta Grossa, 2021.

GUJARATI, D. N. **Econometria básica**. São Paulo: Makron Books, 2000.

GUNARTO, T. *et al.* Accurate Estimated Model of Volatility Crude Oil Price. **International Journal of Energy Economics and Policy**. 10, 5 (Aug. 2020), 228–233.

HAESBAERT, F. **Testes de multicolinearidade em variáveis morfológicas e produtivas de tomateiro**. 2016. Dissertação (Mestrado em Agronomia) – UFSM, 2016.

HAIR JR., J. F.; WILLIAM, B.; BABIN, B.; ANDERSON, R. E. **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman, 2009.

HASHEM, I. A. T. *et al.* The rise of Big Data on cloud computing. **Information Systems**, v. 47, p. 98–115, 2015.

HUANG, B., SUN, Y., & WANG, S. A new two-stage approach with boosting and model averaging for interval-valued crude oil prices forecasting in uncertainty environments. *Frontiers in Energy Research, 9*, Article 707937, 2021.HUBERTY, C. J. **Applied Discriminant Analysis**. New York: John Wiley, 1994.

HUBERTY, C. J. **Applied discriminant analysis**. New York: Wiley, 1994.

JHA, N. *et al.* Multivariate analysis and forecasting of the crude oil prices: Part I – Classical machine learning approaches. **Energy**, v. 296, p. 131185, 2024.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. 4. ed. Englewood Cliffs: Prentice Hall, 1998.

JOBBERSWORLD. **Price adjustments: lubricants, base oils, additives, and others**. Year-in-Review 2024. 30 dez. 2024. Disponível em: https://jobbersworld.com/lubricant-price-adjustments/.

JONES, C. M.; KAUL, G. Oil and the stock markets. **Journal of Finance**, v. 51, n. 2, p. 463–491, 1996.

JOSÉ-GARCÍA, A.; GÓMEZ-FLORES, W. Automatic clustering using nature-inspired metaheuristics. **Applied Soft Computing**, v. 41, p. 192–213, 2016.

KAYALAR, D. E.; KÜÇÜKÖZMEN, C. C.; SELCUK-KESTEL, A. S. The impact of crude oil prices on financial market indicators. **Energy Economics**, v. 61, p. 162–173, 2017.

KENDALL, M. G.; BUCKLAND, W. R. **A Dictionary of Statistical Terms**. New York: Hafner, 1971.

KLINE & COMPANY. **Global lubricant basestocks: market analysis and opportunities**. Relatório de mercado. [S.l.], [s.d.].

KUBRUSLY, L. S. **O Modelo de Análise Fatorial**. Dissertação (Mestrado em Ciências) – UFRJ, 1981.

KRAJEWSKI, L. J.; RITZMAN, L. P.; MALHOTRA, M. K. **Operations management: processes and supply chains**. 9th ed. Upper Saddle River: Pearson, 2009.

LATTIN, J.; CARROL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. São Paulo: Cengage Learning, 2011.

LIBES, D.; SHIN, S.; WOO, J. Considerations and recommendations for data availability for data analytics for manufacturing. In: **IEEE Big Data 2015**, p. 68–75, 2015.

LIMA, A. E. A. **Avaliação e otimização do processo de recuperação de óleos lubrificantes automotivos usados**. 2016. Tese (Doutorado) – UFPB, João Pessoa, 2016.

REVISTA LUBES EM FOCO. Padronização dos óleos básicos e suas aplicações industriais e automotivas. **Revista Lubes em Foco**, São Paulo, 2010.

MAROCO, J. **Análise Estatística: Com a Utilização do SPSS**. Lisboa: Sílabo, 2003.

MARCHEZAN, A.; SOUZA, A. M. Previsão do preço dos principais grãos produzidos no Rio Grande do Sul. **Ciência Rural**, v. 40, n. 11, p. 2368–2374, 2010.

MARQUES, J. M. **Apostila de Análise Multivariada Aplicada à Pesquisa**. UFPR, 2017.

MACGREGOR, J. F.; KOURTI, T. Statistical process control of multivariate processes. **Control Engineering Practice**, v. 3, n. 3, p. 403–414, 1995.

MAXIMIZE MARKET RESEARCH. **Base Oil Market – Global Industry Analysis and Forecast**. Pune, Índia, out. 2024. Relatório de mercado. Disponível em: https://www.maximizemarketresearch.com/market-report/base-oil-market/105579/. Acesso em: jan. 2025.

MINGOTI, S. A. **Análise de Dados Através de Métodos de Estatística Multivariada**. Belo Horizonte: UFMG, 2005.

MORETTIN, P. A.; TOLOI, C. M. C. **Modelos para previsão de séries temporais**. Rio de Janeiro: IMPA, 1981.

MORRISON, D. F. **Multivariate statistical methods**. 2. ed. New York: McGraw-Hill, 1976.

MUSIAL, N. T. K. **Metodologia Box & Jenkins, modelos ARCH-GARCH, redes neurais de camada recorrente e análise de dados em painel na previsão de séries financeiras**. 2016. Dissertação (Mestrado) – UFPR, Curitiba, 2016.

NASCIMENTO, C. A. O.; GUARDANI, R. **Análise Estatística Multivariada aplicada a processos químicos**. Notas de aula. Escola Politécnica da USP, 2007.

OLIVA FILHO, S. Visão da Petrobras sobre o mercado nacional de óleos lubrificantes básicos. In: **Congresso Simepetro**, 3., 2010. Anais [...]. Disponível em: http://www.simepetro.com.br/wp-content/uploads/Mercado-oleos-basicos_Congresso-Simepetro.pdf.

PEBAY, P.; THOMPSON, D.; BENNETT, J.; MASCARENHAS, A. Design and performance of a scalable, parallel statistics toolkit. In: **IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum**, p. 1475–1484, 2011.

PIRES, A. P. Séries temporais. **Notas de aula do LMAC**. Lisboa, 2001.

PLA, L. E. **Analysis multivariado: Método de componentes principales.** Washington: Secretaria General de la Organización de Los Estados Americanos, D. C. 1986.

REGAZZI, A. J. **Análise Multivariada**. Notas de aula. Universidade Federal de Viçosa, 2000.

SADORSKY, P. Oil price shocks and stock market activity. **Energy Economics**, v. 21, n. 5, p. 449–469, 1999.

SAFARI, A.; DAVALLOU, M. Oil Price Forecasting Using a Hybrid Model. **Energy**, *148*, 49-58. https://doi.org/10.1016/j.energy.2018.01.007, 2018.

SAINI, V.; BIJWE, J.; SETH, S.; RAMAKUMAR, S. S. V. Role of base oils in developing extreme pressure lubricants. **Tribology International**, v. 143, 2020.

SAZZAD, Md Rahatul Islam *et al.* Advancing sustainable lubricating oil management. **Heliyon**, v. 10, n. 10, 2024.

SHAO, G.; SHIN, S. J.; JAIN, S. Data analytics using simulation for smart manufacturing. Proceedings - Winter Simulation Conference, v. 2015, n. December, p. 2192–2203, 2015.SHAN, Jessin P. A.; KIRUTHIGA, G. Crude oil price forecasting using ARIMA model. **IRJET – International Research Journal of Engineering and Technology**, v. 7, 2020.

SHARMA, S. **Applied Multivariate Techniques**. John Wiley & Sons, 1996.

SIQUEIRA, A. L.; TIBÚRCIO, J. D.; ALMEIDA, R. M. **Métodos estatísticos multivariados aplicados**. São Paulo: Blucher, 2013.

STEFFEN, D. **Redes neurais e análise multivariada: estudo e aplicação em classificação**. 2021. Tese (Doutorado em Métodos Numéricos em Engenharia) – Programa de Pós-Graduação em Métodos Numéricos em Engenharia, Universidade Federal do Paraná, Curitiba, 2021.

TOON, V. *et al.* Economic exposure to oil price shocks and the fragility of oil-exporting countries. **Energies**, 2018.

VASCONCELOS, S. **Análise das Componentes Principais (PCA)**. Notas de Aula. UFF, 2025.

VICINI, L. **Análise multivariada: da teoria à prática**. 2005. Monografia (Graduação em Estatística) – Universidade Federal de Santa Maria, Santa Maria, 2005.

VO, D. H. *et al.* Modeling the relationship between crude oil and agricultural commodity prices. **Energies**, 12(7), 1344, 2019.

WANG, Y; WANG, C. Oil price shocks and stock market activities: Evidence from oil-importing and oil-exporting countries. **Journal of Comparative Economics**, v. 44, n. 2, p. 420–437, 2016.

YANG, C. *et al.* Characterization and differentiation of chemical fingerprints of virgin and used lubricating oils for identification of contamination or adulteration sources. **Fuel**, v. 163, p. 271-281, out. 2016. DOI: 10.1016/j.fuel.2015.09.070. Disponível em: ResearchGate. Acesso em: 03 mar 2025.

YANG, Y. T. *et al.* Exploring the non-linearity of West Texas Intermediate crude oil price from exchange rate of US dollar and West Texas Intermediate crude oil production. **Energy Strategy Reviews**, v. 41, p. 100854, 2022.

ZHAO, S.; CHANCELLOR, W.; JACKSON, T; BOULT, C. Productivity as a measure of performance: ABARES perspective. [S.l.], set. 2021. Disponível em: https://www.researchgate.net/publication/354802124. Acesso em: 15 de mar. 2025.

ZHU, P. *et al.* Multidimensional risk spillovers among crude oil, the US and Chinese stock markets: Evidence during the COVID-19 epidemic. **Energy**, v. 237, 120949, dez. 2021. DOI: 10.1016/j.energy.2021.120949. Disponível em: https://www.sciencedirect.com/science/article/pii/S036054422101197X. Acesso em: 25 jan. 2025.

ANEXO A - CÓDIGO DE PROJEÇÃO USANDO ARIMA

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from factor_analyzer import FactorAnalyzer, calculate_kmo, calculate_bartlett_sphericity
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_squared_error, mean_absolute_error
from statsmodels.stats.stattools import durbin_watson
from statsmodels.tsa.stattools import acf
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.api import ExponentialSmoothing
import math
from statsmodels.tsa.stattools import adfuller
import warnings

warnings.filterwarnings("ignore")

# Função para encontrar o valor de 'd' (diferenciação) para tornar a série estacionária
def find_d(timeseries):
    d = 0
    while adfuller(timeseries)[1] > 0.05:
        d += 1
        timeseries = np.diff(timeseries, n=1)
        if d > 5:
            break
    return d

# Função para encontrar o melhor modelo ARIMA com base no AIC
def find_best_arima(timeseries):
    best_aic = np.inf
    best_model = None
    best_params = None

    d = find_d(timeseries)

    for p in range(4):
        for q in range(4):
            try:
                model = ARIMA(timeseries, order=(p, d, q))
                model_fit = model.fit()
                if model_fit.aic < best_aic:
                    best_model_name = f'ARIMA({p},{d},{q})'
                    best_aic = model_fit.aic
                    best_model = model_fit
```

```
                best_params = (p, d, q)
            except:
                continue

    return best_model_name, best_model, best_params

# Carregar os dados
dados = pd.read_excel(arquivo com os dados em log10 para realizar as projeções)

# Verificar se a coluna 'DATA' está no formato datetime
dados['DATA'] = pd.to_datetime(dados['DATA'])

# Definir a coluna 'DATA' como índice
dados.set_index('DATA', inplace=True)

# Criar dicionários para armazenar os resultados
resultados_modelos = {}
previsoes_df = pd.DataFrame()
metricas_erro = {}

# Nome da variável que queremos ignorar
coluna_ignorada = "Base oil Group 1"

# Aplicar a modelagem para cada série temporal no DataFrame
for coluna in dados.columns:
    if coluna == coluna_ignorada:
        print(f"Ignorando a coluna: {coluna}")
        continue  # Pula essa variável e segue para a próxima

    timeseries = dados[coluna].dropna()

    if len(timeseries) > 10:
        try:
            # Separar treino e teste (últimos x meses são teste)
            treino = timeseries.iloc[:-x]
            teste = timeseries.iloc[-x:]

            # Encontrar o melhor modelo ARIMA
            best_model_name, best_model, best_params = find_best_arima(treino)
            resultados_modelos[coluna] = {
                'Melhor modelo': best_model_name,
                'Parâmetros': best_params
            }

            # Fazer a previsão para os últimos x meses
            previsao = best_model.forecast(steps=x)
            previsoes_df[coluna] = previsao.values

            # Calcular métricas de erro (MAE, RMSE)
            mae = mean_absolute_error(teste, previsao)
```

```
            rmse = mean_squared_error(teste, previsao, squared=False)

            metricas_erro[coluna] = {'MAE': mae, 'RMSE': rmse}

            print(f"\nColuna: {coluna}")
            print(f"Melhor ARIMA: {best_model_name} | Parâmetros: {best_params}")
            print(f"MAE: {mae:.4f} | RMSE: {rmse:.4f}")

        except Exception as e:
            print(f"Erro ao ajustar ARIMA para '{coluna}': {e}")

# Definir o índice do DataFrame como as datas reais do período de teste
previsoes_df.index = dados.index[-x:]

# Exibir as previsões
print("\nPrevisões para os últimos x meses:")
print(previsoes_df)

# Salvar as previsões e as métricas no Excel
caminho_previsoes = caminho do arquivo
previsoes_df.to_excel(caminho_previsoes)

# Criar DataFrame para salvar as métricas de erro
df_metricas = pd.DataFrame.from_dict(metricas_erro, orient='index')
caminho_metricas = caminho com métricas
df_metricas.to_excel(caminho_metricas)

print(f"\nPrevisões salvas em: {caminho_previsoes}")
print(f"Métricas de erro salvas em: {caminho_metricas}")
```

ANEXO B – CÓDIGO DA ANÁLISE DE COMPONENTES PRINCIPAIS – ÓLEO BÁSICO DE MAIOR PROJEÇÃO

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Ler o conjunto de dados
dados = pd.read_excel(r"arquivo com período real e com o período de projeção")

# Verificar se a coluna 'Data' está no formato datetime e definir como índice
dados['DATA'] = pd.to_datetime(dados['DATA'])
dados.set_index('DATA', inplace=True)

# Separar a variável dependente
Y_log = dados['Base oil Group 1']

# Separar as variáveis independentes e escaloná-las
X = dados.drop(columns=['Base oil Group 1'])
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Aplicação do PCA para manter todos os componentes
pca = PCA(n_components=None)
X_pca = pca.fit_transform(X_scaled)

# Obter os autovalores (variância explicada por componente)
autovalores = pca.explained_variance_

# Selecionar os componentes com autovalores > 1 (critério de Kaiser)
num_componentes_kaiser = sum(autovalores > 1)

# Reduzir o PCA para o número de componentes que atendem ao critério de Kaiser
pca_reduzido = PCA(n_components=num_componentes_kaiser)
X_pca_reduzido = pca_reduzido.fit_transform(X_scaled)

# Dividir o conjunto de dados em treino e teste
X_train = X_pca_reduzido[:-x, :]  # Excluindo os últimos x meses para treino
Y_train = Y_log[:-x]              # Excluindo os últimos x meses

# Ajustar o modelo de regressão linear múltipla
model = LinearRegression()
model.fit(X_train, Y_train)

# Previsões para os últimos x meses usando o modelo ajustado
X_test = X_pca_reduzido[-x:, :]  # Últimos x meses
```

```
Y_pred_test_log = model.predict(X_test)

# Reverter a transformação logarítmica para obter os valores na escala original
Y_pred_test = np.power(10, Y_pred_test_log)
Y_real_test_log = Y_log.iloc[-x:].values.flatten()  # Os últimos x valores reais
Y_real_test = np.power(10, Y_real_test_log)

# Valores reais para os últimos x meses (na escala original)
Y_real_test = Y_log.iloc[-x:].values.flatten()

# Reverter os valores reais de volta para a escala original (caso estejam no log)
Y_real_test = 10 ** Y_real_test
# Calcular o erro percentual absoluto para cada um dos últimos x meses
erro_percentual_absoluto = np.abs((Y_real_test - Y_pred_test) / Y_real_test) * 100

# Exibir os erros absolutos
meses = [f'Mês {i+1}' for i in range(x)]
for i, erro in enumerate(erro_percentual_absoluto):
    print(f'Erro percentual absoluto para {meses[i]}: {erro:.2f}%')

# Calcular as métricas (MSE, MAE, MAPE) na escala original
mse = mean_squared_error(Y_real_test, Y_pred_test)
mae = mean_absolute_error(Y_real_test, Y_pred_test)
mape = np.mean(np.abs((Y_real_test - Y_pred_test) / Y_real_test)) * 100

# Exibir as métricas
print(f'MSE: {mse}')
print(f'MAE: {mae}')
print(f'MAPE: {mape:.2f}%')

# Comparação entre valores reais e previstos na escala original
result_df = pd.DataFrame({
    'Real': Y_real_test,
    'Previsto': Y_pred_test.flatten()
})
print(result_df)

# Criar um DataFrame para armazenar os erros absolutos e os valores reais e previstos
erro_percentual_df = pd.DataFrame({
    'Mês': meses,
    'Erro Percentual Absoluto (%)': erro_percentual_absoluto,
    'Real': Y_real_test,
    'Previsto': Y_pred_test.flatten()
})

# Criar um DataFrame para armazenar o MSE, MAE e MAPE
metricas_df = pd.DataFrame({
    'Métrica': ['MSE', 'MAE', 'MAPE (%)'],
    'Valor': [mse, mae, mape]
})
```

```
# Definir o caminho do arquivo Excel para salvar os dados
caminho_arquivo = r"arquivo salvo.xlsx"

# Criar um objeto ExcelWriter para salvar os dados em diferentes planilhas
with pd.ExcelWriter(caminho_arquivo, engine='xlsxwriter') as writer:
    # Salvar os erros percentuais absolutos e os valores reais e previstos
    erro_percentual_df.to_excel(writer, sheet_name='Erros e Previstos', index=False)

    # Salvar as métricas de MSE, MAE e MAPE
    metricas_df.to_excel(writer, sheet_name='Métricas', index=False)

print(f"Os dados foram salvos com sucesso em: {caminho_arquivo}")
print(f"O número de componentes é: {num_componentes_kaiser}")

# Obter a equação da regressão com os componentes principais selecionados
intercepto = model.intercept_
coeficientes = model.coef_

# Reverter o intercepto para a escala original
intercepto_original = intercepto

# Construir a equação na escala original
equacao = f"y = {intercepto_original:.4f}"
for i, coef in enumerate(coeficientes):
    # Reverter os coeficientes usando a transformação exponencial para escala original
    coef_original = coef
    equacao += f" + ({coef_original:.4f}) * PC{i+1}"

print("Equação da regressão com os componentes principais selecionados:")
print(equacao)

print(f"Os dados foram salvos com sucesso em: {caminho_arquivo}")
print(f"O número de componentes é: {num_componentes_kaiser}")
```

ANEXO C – CÓDIGO DA ANÁLISE FATORIAL – ÓLEO BÁSICO DE MAIOR PROJEÇÃO

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from factor_analyzer import FactorAnalyzer, calculate_kmo, calculate_bartlett_sphericity
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_squared_error, mean_absolute_error
from statsmodels.stats.stattools import durbin_watson
from statsmodels.tsa.stattools import acf


# Ler o conjunto de dados
dados = pd.read_excel(r"arquivo com período real e com período de projeção.xlsx")

# Verificar se a coluna 'Data' está no formato datetime
dados['DATA'] = pd.to_datetime(dados['DATA'])

# Definir a coluna 'Data' como índice
dados.set_index('DATA', inplace=True)

# Separar a variável dependente (Base oil Group 1)
Y = dados['Base oil Group 1 100']

# Separar as variáveis independentes (todas as outras colunas)
X = dados.drop(columns=['Base oil Group 1'])

# Padronizar as variáveis independentes
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Realizar a análise fatorial
fa = FactorAnalyzer(n_factors=X.shape[1], rotation=None)
fa.fit(X_scaled)

# Obter os autovalores para determinar o número de fatores
eigenvalues, _ = fa.get_eigenvalues()
n_factors = sum(eigenvalues > 1)  # Critério de Kaiser

# Ajustar o modelo de análise fatorial com rotação varimax
fa = FactorAnalyzer(n_factors=n_factors, rotation='varimax')
fa.fit(X_scaled)
```

```
# Obter os loadings e os scores fatoriais
loadings = fa.loadings_
scores = np.dot(X_scaled, loadings)

# Transformar os scores fatoriais em um DataFrame
scores_df = pd.DataFrame(scores, columns=[f'Fator{i+1}' for i in range(n_factors)])

# Remover as últimas x linhas de scores_df e Y para o ajuste do modelo de regressão
scores_df_train = scores_df.iloc[:-x, :]  # Todos, exceto os últimos x meses
Y_train = Y.iloc[:-x]          # Todos, exceto os últimos x meses (com log10)

# Ajustar o modelo de regressão linear múltipla
model = LinearRegression()
model.fit(scores_df_train, Y_train)

# Previsões para todos os dados de treino (na escala logarítmica)
Y_pred_train_log = model.predict(scores_df_train)

# Previsões para os últimos x meses (usando as últimas x linhas de scores_df)
Y_pred_test_log = model.predict(scores_df.iloc[-x:, :])

# Reverter as previsões para a escala original (exponencial base 10)
Y_pred_test = 10 ** Y_pred_test_log
Y_pred_train = 10 ** Y_pred_train_log

# Valores reais para os últimos x meses (na escala original)
Y_real_test = Y.iloc[-x:].values.flatten()

# Reverter os valores reais de volta para a escala original (caso estejam no log)
Y_real_test = 10 ** Y_real_test

# Calcular o erro absoluto percentual para cada um dos últimos x meses
erro_percentual_absoluto = np.abs((Y_real_test - Y_pred_test) / Y_real_test) * 100

# Exibir o erro absoluto percentual para cada um dos últimos x meses
meses = [lista os meses de análise]
for i, erro in enumerate(erro_percentual_absoluto):
   print(f'Erro percentual absoluto para {meses[i]}: {erro:.2f}%')

# Calcular os erros (MSE, MAE, MAPE) na escala original
mse = mean_squared_error(Y_real_test, Y_pred_test)
mae = mean_absolute_error(Y_real_test, Y_pred_test)
mape = np.mean(np.abs((Y_real_test - Y_pred_test) / Y_real_test)) * 100

# Exibir os erros
print(f'MSE: {mse}')
print(f'MAE: {mae}')
print(f'MAPE: {mape:.2f}%')
```

```
# Comparação entre valores reais e previstos
result_df = pd.DataFrame({
    'Real': Y_real_test,
    'Previsto': Y_pred_test.flatten()
})
print(result_df)

# Criar um DataFrame para armazenar os erros absolutos e os valores reais e previstos
erro_percentual_df = pd.DataFrame({
    'Mês': meses,
    'Erro Percentual Absoluto (%)': erro_percentual_absoluto,
    'Real': Y_real_test,
    'Previsto': Y_pred_test.flatten()
})

# Criar um DataFrame para armazenar o MSE, MAE e MAPE
metricas_df = pd.DataFrame({
    'Métrica': ['MSE', 'MAE', 'MAPE (%)'],
    'Valor': [mse, mae, mape]
})

# Definir o caminho do arquivo Excel para salvar os dados
caminho_arquivo = r"caminho do arquivo para salvar os dados"

# Criar um objeto ExcelWriter para salvar os dados em diferentes planilhas
with pd.ExcelWriter(caminho_arquivo, engine='xlsxwriter') as writer:
    # Salvar os erros absolutos e os valores reais e previstos
    erro_percentual_df.to_excel(writer, sheet_name='Erros e Previstos', index=False)

    # Salvar as métricas de MSE, MAE e MAPE
    metricas_df.to_excel(writer, sheet_name='Métricas', index=False)

print(f"Os dados foram salvos com sucesso em: {caminho_arquivo}")
```

ANEXO D – CÓDIGO DA ANÁLISE DE AGRUPAMENTO E DISCRIMINANTE

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from factor_analyzer import FactorAnalyzer, calculate_kmo, calculate_bartlett_sphericity
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import mean_squared_error, mean_absolute_error
from statsmodels.stats.stattools import durbin_watson
from statsmodels.tsa.stattools import acf
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.api import ExponentialSmoothing
import math
from statsmodels.tsa.stattools import adfuller
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from scipy.spatial.distance import cdist, pdist
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from yellowbrick.cluster import SilhouetteVisualizer

def silhouettePlot(range_, data):
    '''
    Mostra os gráficos do coeficiente de silhueta para avaliar coesão intra-agrupamento.
    '''
    n_plots = len(range_)
    n_rows = math.ceil(n_plots / 2)  # Número de linhas para os subplots (2 colunas por linha)
    fig, ax = plt.subplots(n_rows, 2, figsize=(15, 15))
    ax = ax.flatten()  # Garantir que o array de eixos seja unidimensional

    for idx, n_clusters in enumerate(range_):
        kmeans = KMeans(n_clusters=n_clusters, random_state=42)
        sv = SilhouetteVisualizer(kmeans, colors="yellowbrick", ax=ax[idx])
        sv.fit(data)
        ax[idx].set_title(f"Silhouette Plot with n={n_clusters} Clusters")

    # Ajustar layout e exibir
    fig.tight_layout()
    plt.show()
    fig.savefig("silhouette_plot.png")
def elbowPlot(range_, data, figsize=(3.5,3.5)):
    '''
    a função para produzir o gráfico do método do cotovelo vai ajudar-nos a determinar o número
adequado de agrupamentos para o nosso conjunto de dados
    '''
    inertia_list = []
    for n in range_:
```

```
        kmeans = KMeans(n_clusters=n, random_state=42)
        kmeans.fit(data)
        inertia_list.append(kmeans.inertia_)

    # plotting
    fig = plt.figure(figsize=figsize)
    ax = fig.add_subplot(111)
    sns.lineplot(y=inertia_list, x=range_, ax=ax)
    ax.set_xlabel("Cluster")
    ax.set_ylabel("Inertia")
    ax.set_xticks(list(range_))
    fig.show()
    fig.savefig("elbow_plot.png")

# Carregar os dados e ignorar a primeira coluna (posição 0)
dados = pd.read_excel(
    arquivo com os dados reais e o projetado"
)
dados = dados.iloc[:, 1:]  # Ignorar a primeira coluna com base na posição

print(dados)

# Normalizar os dados
scaler = StandardScaler()
dados_norm = scaler.fit_transform(dados)

# Aplicar PCA para redução de dimensionalidade
pca = PCA(n_components=x)  # Escolha o número de componentes principais desejado
dados_pca = pca.fit_transform(dados_norm)

# Exibir a variância explicada pelas componentes principais
explained_variance_ratio = pca.explained_variance_ratio_
print(f"Variância explicada por cada componente principal: {explained_variance_ratio}")
print(f"Variância total explicada: {np.sum(explained_variance_ratio)}")

# Calcular os loadings (cargas fatoriais)
# Loadings = Componentes principais * raiz(variância explicada)
loadings = pca.components_.T * np.sqrt(pca.explained_variance_)

# Criar um DataFrame para exibir os loadings
df_loadings = pd.DataFrame(
    loadings,
    columns=[f'PC{i+1}' for i in range(loadings.shape[1])],
    index=dados.columns  # Usar os nomes das variáveis originais como índice
)
print("Loadings (cargas fatoriais):")
print(df_loadings)

# Salvar os loadings em um arquivo Excel
df_loadings.to_excel("Salvar arquivo com loading")
```

```
elbowPlot(range(2, 15), df_loadings)
silhouettePlot(range(2, 15), df_loadings)

# Aplicar o algoritmo K-Means
n_clusters_otimo = 7  # Número ótimo de clusters
kmeans = KMeans(n_clusters=n_clusters_otimo, random_state=42)
kmeans.fit(df_loadings)

# Adicionar os rótulos de cluster ao DataFrame de loadings
df_loadings['Cluster'] = kmeans.labels_

# Exibir a contagem de observações por cluster
print(df_loadings['Cluster'].value_counts())

# Visualização dos clusters
sns.pairplot(df_loadings, hue='Cluster', palette='viridis')
plt.show()

# Criar um dicionário para armazenar as variáveis de cada cluster
clusters_dict = {}

for i in range(n_clusters_otimo):
    cluster_vars = df_loadings[df_loadings['Cluster'] == i].index.tolist()  # Obter as variáveis do
cluster
    clusters_dict[f"Cluster {i}"] = cluster_vars

# Converter o dicionário em um DataFrame
# Preenchendo com NaN para garantir que todas as colunas tenham o mesmo tamanho
df_clusters = pd.DataFrame(dict([(k, pd.Series(v)) for k, v in clusters_dict.items()]))

# Salvar o DataFrame reorganizado em um arquivo Excel
df_clusters.to_excel("Salvar arquivo com os dados obtidos", index=False)

print("Arquivo Excel com as variáveis agrupadas por cluster salvo com sucesso!")

# Adicionar os rótulos de cluster ao DataFrame de loadings
df_loadings['Cluster'] = kmeans.labels_

# Criar um DataFrame consolidado com as variáveis, pesos e número do cluster
# Resetar o índice para trazer as variáveis como uma coluna
final_df = df_loadings.reset_index()

# Renomear a coluna do índice para "Variável"
final_df.rename(columns={'index': 'Variável'}, inplace=True)

# Salvar o DataFrame consolidado em um arquivo Excel
output_path = "Salvar arquivo com dados.xlsx"
final_df.to_excel(output_path, index=False)
```

```
print(f"Arquivo salvo com sucesso em: {output_path}")


# Carregar o arquivo organizado
input_path= "Salvar arquivo com os dados.xlsx"
data = pd.read_excel(input_path)

# Separar as variáveis preditoras (pesos dos componentes principais) e a variável alvo (Cluster)
X = data.iloc[:, 1:-1]  # Colunas com os pesos (PC1 a PCX)
y = data['Cluster']     # Coluna com os clusters originais

# Treinar o modelo LDA
lda = LinearDiscriminantAnalysis()
lda.fit(X, y)

# Prever os clusters com base no modelo LDA
y_pred = lda.predict(X)

# Avaliar a acurácia da classificação atual
accuracy = accuracy_score(y, y_pred)
conf_matrix = confusion_matrix(y, y_pred)
print(f"Acurácia do modelo: {accuracy:.2f}")
print("Matriz de confusão:")
print(conf_matrix)

# Adicionar os clusters ajustados ao DataFrame
data['Cluster Ajustado'] = y_pred

# Salvar o arquivo ajustado
output_path = "arquivo salvo com as informações"
data.to_excel(output_path, index=False)

print(f"Arquivo ajustado salvo com sucesso em: {output_path}")
```

ANEXO E – CÓDIGO DA ANÁLISE E VERIFICAÇÃO DAS VARIÁVEIS DE IMPACTO – 30 ÓLEOS BÁSICOS

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, mean_absolute_percentage_error
import matplotlib.pyplot as plt

# ==========================
# 1. Ler e preparar os dados
# ==========================
caminho_dados = r"arquivo com dados reais"
dados = pd.read_excel(caminho_dados)

# Substituir NaN pela média de cada coluna
dados = dados.fillna(dados.mean())

# Converter coluna DATA para datetime e definir como índice
dados['DATA'] = pd.to_datetime(dados['DATA'])
dados.set_index('DATA', inplace=True)

# Lista dos 30 óleos básicos (dependentes)
oleos_dependentes = [
  Lista dos Óleos 30 Óleos Básicos
]

# ==========================
# 2. Variáveis independentes (X)
# ==========================
X = dados.drop(columns=oleos_dependentes)  # remove os óleos básicos

# Escalonar
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# ==========================
# 3. PCA e seleção de componentes
# ==========================
pca = PCA(n_components=None)
X_pca = pca.fit_transform(X_scaled)
autovalores = pca.explained_variance_

# Critério de Kaiser
```

```
num_componentes_kaiser = sum(autovalores > 1)

# PCA reduzido
pca_reduzido = PCA(n_components=num_componentes_kaiser)
X_pca_reduzido = pca_reduzido.fit_transform(X_scaled)

# Loadings completos
loadings = pca_reduzido.components_.T * np.sqrt(pca_reduzido.explained_variance_)
df_loadings = pd.DataFrame(
    loadings,
    index=X.columns,
    columns=[f'PC{i+1}' for i in range(num_componentes_kaiser)]
)

# Variância explicada e acumulada
var_exp = pca_reduzido.explained_variance_ratio_ * 100
var_exp_acum = np.cumsum(var_exp)
df_var_exp = pd.DataFrame({
    'Componente': [f'PC{i+1}' for i in range(num_componentes_kaiser)],
    'Variância Explicada (%)': var_exp,
    'Variância Acumulada (%)': var_exp_acum
})

# ============================
# 4. Regressão para cada óleo básico
# ============================
todos_erros = []
todas_metricas = []
todas_equacoes = []
todos_previstos = []

for oleo in oleos_dependentes:
    # Variável dependente (em log10) - substituir NaN pela média
    y_log = dados[oleo].fillna(dados[oleo].mean())

    # Divisão treino/teste (últimos x meses como teste)
    X_train = X_pca_reduzido[:-x, :]
    Y_train_log = y_log[:-x]
    X_test = X_pca_reduzido[-x:, :]
    Y_real_test_log = y_log[-x:].values.flatten()

    # Ajustar regressão linear
    modelo = LinearRegression()
    modelo.fit(X_train, Y_train_log)
    Y_pred_test_log = modelo.predict(X_test)

    # Reverter para escala real
    Y_pred_test = 10 ** Y_pred_test_log
    Y_real_test = 10 ** Y_real_test_log
```

```
    # Métricas na escala real
    mse = mean_squared_error(Y_real_test, Y_pred_test)
    mae = mean_absolute_error(Y_real_test, Y_pred_test)
    mape = mean_absolute_percentage_error(Y_real_test, Y_pred_test)

    todas_metricas.append({
        "Oleo": oleo, "MSE": mse, "MAE": mae, "MAPE": mape
    })

    # Previstos e reais
    df_prev = pd.DataFrame({
        "Data": dados.index[-x:],
        "Oleo": oleo,
        "Real": Y_real_test,
        "Previsto": Y_pred_test
    })
    todos_previstos.append(df_prev)

    # Equação da regressão (em log10)
    eq = f"{oleo}: log10(y) = {modelo.intercept_:.4f}"
    for i, coef in enumerate(modelo.coef_):
        eq += f" + ({coef:.4f})*PC{i+1}"
    todas_equacoes.append(eq)

    # Exibir erros no console
    erro_percentual = np.abs((Y_real_test - Y_pred_test) / Y_real_test) * 100
    for i, e in enumerate(erro_percentual):
        print(f"{oleo} - Mês {i+1}: Erro percentual = {e:.2f}%")

# Consolidar resultados
previstos_df = pd.concat(todos_previstos, ignore_index=True)
# Consolidar métricas para todos os óleos
metricas_df = pd.DataFrame(todas_metricas)

# Converter MAPE para porcentagem
metricas_df['MAPE (%)'] = metricas_df['MAPE'] * 100
metricas_df.drop(columns=['MAPE'], inplace=True)
equacoes_df = pd.DataFrame({"Oleo": oleos_dependentes, "Equacao": todas_equacoes})

# ============================
# 5. Filtrar Loadings relevantes
# ============================
limite = 0.08
df_loadings_filtrados = df_loadings.copy()
df_loadings_filtrados = df_loadings_filtrados[(df_loadings_filtrados.abs() >= limite)]
df_loadings_filtrados = df_loadings_filtrados.dropna(how='all')

# ============================
# 6. Exportar para Excel
# ============================
```

```
caminho_saida = r"arquivo com as informações"
with pd.ExcelWriter(caminho_saida, engine="xlsxwriter") as writer:
    df_loadings.to_excel(writer, sheet_name="Loadings")
    df_loadings_filtrados.to_excel(writer, sheet_name="Loadings Filtrados")
    df_var_exp.to_excel(writer, sheet_name="Variancia Explicada", index=False)
    metricas_df.to_excel(writer, sheet_name="Metricas", index=False)
    previstos_df.to_excel(writer, sheet_name="Previstos", index=False)
    equacoes_df.to_excel(writer, sheet_name="Equacoes", index=False)

print(f" Resultados salvos em: {caminho_saida}")
print(f"Número de componentes selecionados (Kaiser): {num_componentes_kaiser}")

# ==========================
# 7. Gráficos
# ==========================
# Scree Plot
plt.figure(figsize=(10, 6))
plt.plot(np.arange(1, len(autovalores)+1), autovalores, marker='o')
plt.axhline(y=1, color='r', linestyle='--', label='Autovalor = 1')
plt.title('Curva de Declive (Scree Plot)')
plt.xlabel('Número do Componente Principal')
plt.ylabel('Autovalor (Variância Explicada)')
plt.legend()
plt.grid(True)
plt.show()

# Predito vs Observado (exemplo com o primeiro óleo)
plt.figure(figsize=(7, 5))
plt.scatter(previstos_df[previstos_df['Oleo'] == oleos_dependentes[0]]['Previsto'],
        previstos_df[previstos_df['Oleo'] == oleos_dependentes[0]]['Real'],
        color='blue', edgecolor='k')
plt.plot([previstos_df['Previsto'].min(), previstos_df['Previsto'].max()],
       [previstos_df['Previsto'].min(), previstos_df['Previsto'].max()],
       color='red', linestyle='--')
plt.title(f'Predito vs Observado ({oleos_dependentes[0]})')
plt.xlabel('Previsto')
plt.ylabel('Real')
plt.grid(True, linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```