UNIVERSIDADE ESTADUAL DO PIAUÍ
CENTRO DE TECNOLOGIA E URBANISMO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Gabriel Santana da Silva

A Influência do Investimento Estrangeiro no Mercado Acionário Brasileiro: Uma Análise com Técnicas Estatísticas e de Machine Learning

TERESINA
2025

Gabriel Santana da Silva

A Influência do Investimento Estrangeiro no Mercado Acionário Brasileiro: Uma Análise com Técnicas Estatísticas e de Machine Learning

Monografia de Trabalho de Conclusão de Curso apresentado na Universidade Estadual do Piauí – UESPI como parte dos requisitos para conclusão do Curso de Bacharelado em Ciência da Computação.

Orientador: Prof. Dr. Carlos Giovanni Nunes de Carvalho

TERESINA

2025

S586i Silva, Gabriel Santana da.
A influência do investimento estrangeiro no mercado acionário brasileiro: uma análise com técnicas estatísticas e de machine learning / Gabriel Santana da Silva. - 2025.
56f.: il.

Monografia (Graduação) - Universidade Estadual do Piauí - UESPI, Centro de Tecnologia e Urbanismo (CTU), Bacharelado em Ciência da Computação, Teresina, 2025.
"Orientador: Prof. Dr. Carlos Giovanni Nunes de Carvalho".

1. Correlação. 2. Ibovespa. 3. Investimento Estrangeiro. 4. K-means. 5. Causalidade de Granger. I. Carvalho, Carlos Giovanni Nunes de . II. Título.

CDD 006.31

Ficha elaborada pelo Serviço de Catalogação da Biblioteca da UESPI
JOSÉ EDIMAR LOPES DE SOUSA JÚNIOR (Bibliotecário) CRB-3ª/1512

Gabriel Santana da Silva

A Influência do Investimento Estrangeiro no Mercado Acionário Brasileiro: Uma Análise com Técnicas Estatísticas e de Machine Learning

Monografia de Trabalho de Conclusão de Curso apresentado na Universidade Estadual do Piauí – UESPI como parte dos requisitos para conclusão do Curso de Bacharelado em Ciência da Computação.

Aprovada em 03 de Julho de 2025.

BANCA EXAMINADORA:

gov.br Documento assinado digitalmente
CARLOS GIOVANNI NUNES DE CARVALHO
Data: 09/07/2025 20:24:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Carlos Giovanni Nunes de Carvalho- Orientador.
Universidade Estadual do Piauí - UESPI

gov.br Documento assinado digitalmente
ALDIR SILVA SOUSA
Data: 14/07/2025 11:54:40-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Aldir Silva Sousa - Membro
Universidade Estadual do Piauí - UESPI

gov.br Documento assinado digitalmente
SERGIO BARROS DE SOUSA
Data: 10/07/2025 11:36:11-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Sérgio Barros de Sousa - Membro
Universidade Estadual do Piauí - UESPI

Dedico este trabalho à minha família

AGRADECIMENTOS

A minha mãe e meu pai por me incentivarem a cada dia a alcançar meus sonhos, paciência, amor e suporte na minha árdua caminhada até aqui.

Aos professores do curso de Ciência da Computação, em especial ao professor Carlos Giovanni Nunes de Carvalho, por toda a orientação e apoio fundamentais nas etapas finais desta jornada.

Aos colegas de curso que compartilharam conhecimento e bons momentos durante minha jornada na UESPI.

RESUMO

Este trabalho investiga a relação entre o fluxo de investimento estrangeiro na B3, o índice Ibovespa (IBOV) e as ações que o compõem, buscando identificar se os fluxos estrangeiros podem servir como indicadores antecipados de movimentos no mercado acionário brasileiro. Para isso, são aplicadas técnicas estatísticas como correlação de Pearson, regressão linear e causalidade de Granger, além de métodos de aprendizado de máquina, como o algoritmo K-Means e a análise de componentes principais (PCA), sobre séries temporais normalizadas.Os dados analisados abrangem o período de 19 de agosto de 2024 a 14 de junho de 2025, incluindo valores diários do fluxo estrangeiro fornecidos pela B3, e os preços de fechamento do IBOV e de ações selecionadas obtidos no Yahoo Finance. Foram consideradas diferentes janelas temporais para verificar a robustez das relações e a presença de padrões dinâmicos. Os resultados mostram que, quanto maior o volume de dados, menor é a correlação de Pearson entre o fluxo estrangeiro e o IBOV, mas maior é a evidência de causalidade de Granger. A análise com PCA e K-Means revelou agrupamentos com alta participação estrangeira associada a altos níveis do IBOV, sugerindo uma relação positiva entre essas variáveis. No caso das ações que compõem o IBOV, observou-se independência em janelas curtas, mas uma correlação mais forte no médio e longo prazo, indicando coesão estrutural do índice. Já a correlação entre o fluxo estrangeiro e ações individuais foi baixa, sugerindo que o impacto direto dos investidores estrangeiros sobre papéis específicos é limitado.

Palavras-chaves: ações; ibovespa, investimento estrangeiro, correlação, causalidade de Granger, k-means, séries temporais.

ABSTRACT

This study investigates the relationship between foreign investment flow on B3, the Ibovespa index (IBOV), and its constituent stocks, aiming to determine whether foreign flows can serve as leading indicators of movements in the Brazilian stock market. To this end, statistical techniques such as Pearson correlation, linear regression, and Granger causality are applied, along with machine learning methods including the K-Means algorithm and principal component analysis (PCA), using normalized time series. The data analyzed covers the period from August 19, 2024, to June 14, 2025, including daily foreign investment flow values provided by B3 and the closing prices of the IBOV and selected stocks retrieved from Yahoo Finance. Different time windows are considered to assess the robustness of the relationships and the presence of dynamic patterns. The results show that the larger the dataset, the lower the Pearson correlation between foreign flow and the IBOV, but the greater the evidence of Granger causality. The PCA and K-Means analysis revealed clusters with high foreign participation associated with high IBOV levels, suggesting a positive relationship between these variables. Regarding the IBOV constituent stocks, independence was observed in short-term windows, but stronger correlation emerged in the medium and long term, indicating structural cohesion of the index. However, the correlation between foreign flow and individual stocks was low, suggesting that the direct impact of foreign investors on specific stocks is limited.

Keywords: stocks, ibovespa, foreign investment, correlation, Granger causality, k-means, time series

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
B3	Bolsa de Valores do Brasil
IBOV	Ibovespa
PCA	*Principal Component Analysis*
SEE	*Sum of Squared Errors*

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Em mercados emergentes como o Brasil, os investidores estrangeiros exercem influência significativa sobre o comportamento do mercado financeiro, especialmente nas oscilações dos preços dos ativos negociados na Bolsa de Valores do Brasil (B3). Esses investidores respondem por uma parcela significativa do volume financeiro movimentado diariamente e impactam diretamente os movimentos do principal índice acionário do país, o Ibovespa (IBOV). Para ilustrar, no início de 2025, o fluxo de investidores estrangeiros na B3 apresentou um saldo positivo de R$ 6,82 bilhões e, no mesmo período, a Ibovespa acumulou uma alta de 4,86%[1]. Em geral, um aumento no fluxo estrangeiro tende a impulsionar ações com maior liquidez e capitalização; o contrário também tende a ocorrer em momentos de retirada de capital. Um alto grau de correlação cruzada entre a simultaneidade de um conjunto de retornos de ações é um fato empírico bem conhecido (Imeni; Bao; Nozick, 2024). Logo, a alta participação de investidores estrangeiros nas negociações da B3 torna seus fluxos de capital um possível indicador antecedente de movimentos do mercado.

Compreender a relação entre o desempenho do Ibovespa e os fatores que influenciam seu comportamento é de grande importância para investidores e analistas, uma vez que permite interpretar com mais precisão os movimentos do mercado e perceber possíveis impactos Castro (2022, p. 15). O Ibovespa é o principal indicador de desempenho das ações das maiores empresas do país, que reflete o desempenho médio das ações mais negociadas na B3, serve como referência para o mercado de ações brasileiro, sendo composto por empresas com alto volume de negociação e grande representatividade. Assim, compreender a correlação entre o fluxo de capital estrangeiro, o IBOV e as ações que o compõem é fundamental para interpretar os movimentos do mercado, sendo relevante para a construção de estratégias de investimento, escolha de ativos para a diversificação de portfólios e mitigação de riscos.

De acordo com Meng et al. (2024, p. 1058), os estudos atuais sobre correlações concentram-se principalmente em aspectos como competição e colaboração dentro da indústria esses estudos abordam, em sua maioria, temas como indústria, cadeia de suprimentos, investimentos, competição e colaboração.

A correlação é construída ao longo do tempo. Segundo Marti et al. (2017, p. 15), existem duas dinâmicas distintas para a formação dessas correlações: uma lenta e outra rápida. Na dinâmica lenta, os relacionamentos se desenvolvem gradativamente.

[1] Estrangeiros aportam R$ 6,82 bi em janeiro na B3, maior valor desde agosto de 2024 (2025), disponível em <https://www.infomoney.com.br/mercados/estrangeiros-aportam-r-682-bi-em-janeiro-na-b3-maior-valor-desde-agosto-de-2024/>

Já na dinâmica rápida, as correlações são formadas de maneira abrupta, geralmente impulsionadas por eventos de grande impacto, como crises financeiras, conflitos geopolíticos, aumento das tarifas entre países, desastres naturais ou pandemias.

No entanto, ainda não está claro a existência de uma relação direta e estável entre os fluxos estrangeiros e a variação do Ibovespa ou das ações brasileiras que a compõem, bem como se essa possível influência é forte o suficiente para sustentar modelos preditivos ou estratégias de investimento.

Diante desse desafio, a utilização de metodologias tradicionais, como a correlação de Pearson e causalidade de Granger, aliadas a técnicas de *machine learning* (aprendizado de máquina), podem esclarecer essas dúvidas. Segundo Avelar et al. (2022 apud Leal, 2024, p. 11),"Nos últimos dez anos, houve um aumento significativo no uso de algoritmos de inteligência artificial para análise de padrões".

A correlação de Pearson é uma medida estatística que avalia a força e a direção da relação linear entre duas variáveis e é a estatística de correlação mais amplamente utilizada para isso (Imeni; Bao; Nozick, 2024). Já o teste de Granger avalia se uma série temporal possui valor preditivo em relação a outra.

Já as técnicas de *clustering* permitem ajustes dinâmicos e em tempo real às mudanças do mercado, além de possibilitarem a descoberta de novos padrões e a construção de modelos preditivos mais robustos. Segundo (Vanhala; Järvi; Heikkonen, 2023), "A clusterização de dados de séries temporais é uma abordagem comum para encontrar similaridades ou correlações em um espaço de alta dimensionalidade".

1.1 Objetivos

Considerando a contextualização e a problemática apresentadas, nesta seção, apresentaremos os objetivos gerais e específicos que orientaram o desenvolvimento deste trabalho.

1.1.1 Objetivo Geral

O objetivo geral desta monografia é investigar a existência, a intensidade e direção das relações entre o fluxo de investimento estrangeiro, o índice Ibovespa e as ações que o compõem, utilizando técnicas tradicionais e de aprendizado de máquina, em séries temporais.

1.1.2 Objetivos Específicos

Com o intuito de alcançar o objetivo geral deste estudo, buscar-se-á realizar os seguintes objetivos específicos:

- Analisar a correlação e o grau de causalidade entre o fluxo de capital estrangeiro e o índice Ibovespa.
- Analisar a correlação e o grau de causalidade entre o Ibovespa e as ações que fazem parte de sua composição.
- Verificar quais ações específicas apresentam maior sensibilidade ao fluxo estrangeiro direto.
- Descobrir se as correlações se mantêm ou se alteram com o aumento da série temporal.
- Analisar se a relação entre fluxo estrangeiro e Ibovespa, assim como a relação entre o Ibovespa e as ações que fazem parte de sua composição, são fortes o suficiente para prever movimentos futuros.
- Analisar se os resultados do agrupamento superam as técnicas tradicionais e são generalizáveis a ponto de permitir a tomada de decisões de investimento.

1.2 Metodologia da pesquisa

Neste trabalho, investiga-se a relação entre o fluxo de capital estrangeiro, o índice Ibovespa (IBOV) e as ações que o compõem por meio de três etapas principais: (1) análise da correlação entre o fluxo de capital estrangeiro e o IBOV; (2) análise da correlação entre o IBOV e as ações de sua carteira; e (3) análise da correlação entre o fluxo estrangeiro e ações específicas do mercado brasileiro. O objetivo é identificar não apenas a existência de correlações, mas também o grau de causalidade, a fim de verificar se determinados ativos apresentam maior sensibilidade às entradas e saídas de capital estrangeiro.

Em todas as etapas mencionadas, foram utilizados métodos estatísticos tradicionais, como a correlação de Pearson, a regressão linear e o teste de causalidade de Granger. Além disso, aplicaram-se técnicas de aprendizado de máquina não supervisionado, com destaque para o algoritmo de clusterização K-means e a análise do *Principal Component Analysis* (PCA).

Os dados relativos aos investimentos estrangeiros foram extraídos de relatórios divulgados pela B3, enquanto os dados do IBOV e das ações foram obtidos por meio da plataforma Yahoo Finance, utilizando a biblioteca Python *yfinance*. Todos os dados foram tratados com técnicas de normalização *Z-score*, remoção de outliers baseado no *Z-score* e alinhamento temporal, de forma a garantir consistência e comparabilidade entre as séries analisadas.

1.3 Estrutura dos capítulos

Esta monografia é organizada em cinco capítulos.

O Capítulo 1 - visa fornecer ao leitor uma visão geral do trabalho, incluindo o contexto da pesquisa, a justificativa e os objetivos gerais e específicos.

O Capítulo 2 - reúne os principais conceitos, teorias e estudos anteriores que embasam a análise realizada. Este capítulo tem como objetivo oferecer suporte teórico para a compreensão dos temas abordados, como séries temporais, indicadores financeiros, métodos estatísticos e de aprendizado de máquina, além de explicar as ferramentas aplicadas, como a correlação de Pearson, a regressão linear, o teste de causalidade de Granger, o PCA e o K-means.

O Capítulo 3 - Detalha os procedimentos técnicos e metodológicos utilizados para alcançar os objetivos da pesquisa. São descritas as etapas de coleta, organização e tratamento dos dados, bem como os critérios adotados para a aplicação das análises estatísticas e dos métodos computacionais.

No Capítulo 4 - Apresenta os principais resultados obtidos a partir da aplicação das técnicas descritas anteriormente. A interpretação dos dados é feita de maneira crítica, relacionando os achados com os objetivos propostos e com o referencial teórico.

Por fim, o Capítulo 5 - Expõe as conclusões da pesquisa, retomando os objetivos e destacando as contribuições do estudo.

2 REFERENCIAL TEÓRICO

Este capítulo é um componente que fornece a base teórica necessária para o entendimento deste trabalho. Apresentam-se as principais teorias, conceitos e definições existentes sobre as áreas de conhecimento que motivaram a elaboração deste trabalho e que guiaram a execução do percurso metodológico. Neste capítulo, serão explicados os princípios da correlação de Pearson, causalidade de Granger, regressão linear, aprendizado de máquina, com destaque para o *clustering* K-Means, bem como os métodos *Elbow Method* (Método do cotovelo), *Silhouette Score* (Coeficiente de silhueta) e PCA, assim como o método de normalização *Z-score* e remoção de outliers baseada no *Z-score*.

2.1 Correlação de Pearson e Regressão Linear

2.1.1 Correlação de Pearson

A correlação de Pearson é uma medida estatística que avalia a força e a direção da relação linear entre duas variáveis, por meio de um número que vai de –1 a +1. Isto é, quanto mais próximo dos valores extremos (–1 ou +1), maior é a força da correlação. Por outro lado, valores próximos de zero indicam que a correlação é fraca. Em finanças, ela é amplamente utilizada para medir relações entre retornos de ativos, volume de negociação e indicadores econômicos (Imeni; Bao; Nozick, 2024). Logo, se IBOV ou as ações tendem a subir ou cair quando o fluxo estrangeiro aumenta ou diminui, especificamente, isso pode aparecer como uma correlação positiva significativa.

A seguinte fórmula é utilizada para calcular a correlação de Pearson, r

A fórmula do coeficiente de correlação de Pearson entre duas variáveis X e Y, com n observações, é apresentada na Equação 2.1:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

Onde:

- x_i e y_i são os valores individuais das variáveis X e Y;
- $\bar{x}$ e $\bar{y}$ são as médias de X e Y, respectivamente;
- n é o número total de observações;

- O numerador representa a **covariância** entre X e Y;
- O denominador é o produto dos **desvios padrão** de X e Y.

Fonte: Sharma (1996).

Essa fórmula mede o grau de correlação linear entre duas variáveis. O resultado está sempre entre -1 e 1:

- $r = 1$ indica uma correlação linear positiva perfeita;
- $r = -1$ indica uma correlação linear negativa perfeita;
- $r = 0$ indica ausência de correlação linear.

Interpretação:

- Valores de r próximos de 1 indicam forte correlação positiva: quando X aumenta, Y tende a aumentar.
- Valores de r próximos de -1 indicam forte correlação negativa: quando X aumenta, Y tende a diminuir.
- Valores de r próximos de 0 indicam fraca ou nenhuma correlação linear.

2.1.2 Regressão Linear

A Regressão linear é utilizada para descrever a relação linear entre variáveis. Existem dois tipos de regressão linear, a regressão linear simples e a regressão linear múltipla. Nesse trabalho, é utilizada a regressão linear simples, nela é descrita a relação linear entre uma variável independente e sua variável dependente correspondente. Essa relação é expressa como uma linha reta.

A correlação de Pearson e a regressão linear estão intimamente relacionadas. A correlação mede a intensidade e a direção da associação linear entre duas variáveis, enquanto a regressão linear busca modelar essa relação por meio da Equação 2.2:

$$Y = \alpha + \beta X + \varepsilon \tag{2.2}$$

Onde:

- Y é a variável dependente;
- X é a variável independente;

- α é o intercepto (constante);
- β é o coeficiente angular (inclinação da reta);
- ε é o erro aleatório.

Fonte: Montgomery, Peck e Vining (2012).

Em particular, quando as variáveis são padronizadas (isto é, têm média 0 e desvio padrão 1), o coeficiente de regressão β torna-se igual ao coeficiente de correlação de Pearson r.

Dessa forma, a correlação de Pearson pode ser vista como um caso específico da regressão linear, onde o foco está apenas na força e direção da associação, e não na estimativa dos valores de Y a partir de X.

2.2 Causalidade de Granger

A partir do cálculo do coeficiente de correlação de Pearson, é possível verificar a existência de similaridades entre as variáveis analisadas. Contudo, mesmo que se observe um alto grau de correlação, isso não implica, necessariamente, em uma relação de causalidade entre elas (Imeni; Bao; Nozick, 2024). Em outras palavras, uma variação em um ativo não significa, necessariamente, que ela cause uma variação em outro. Essa correlação pode, na verdade, refletir a influência de fatores subjacentes comuns, como condições macroeconômicas, aspectos psicológicos dos investidores ou outros elementos com impacto relevante (Imeni; Bao; Nozick, 2024).

Nesse sentido, a causalidade de Granger serve para verificar se uma série temporal é útil em termos de previsão de outra série temporal, ou seja, se uma variável X precede temporalmente uma variável Y, quando valores defasados de X, além de valores defasados de Y, ajudam a prever o valor presente de Y.

Essa abordagem é relevante neste trabalho, pois permite investigar se o comportamento do fluxo de investidores estrangeiros antecipa (ou é antecipado por) movimentos no índice Ibovespa e das ações que o compõem, assim como a relação de causalidade entre o próprio IBOV e essas ações.

A causalidade pode ser unidirecional — quando apenas uma variável contribui para prever a outra — ou bidirecional, quando ambas possuem capacidade preditiva mútua.

Sejam duas séries temporais X_t e Y_t. Para verificar se X causa Granger Y, comparam-se dois modelos, como é visto na Equação 2.3 e Equação 2.4:

$$Y_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \varepsilon_t \tag{2.3}$$

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=1}^{p} \gamma_j X_{t-j} + u_t \tag{2.4}$$

A hipótese nula do teste de causalidade de Granger é apresentado na Equação 2.5:

$$H_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0 \tag{2.5}$$

Ou seja, as defasagens de X_t não contribuem significativamente para explicar Y_t, implicando que X_t não causa Granger Y_t.

Fonte: Granger (1969).

Para testar essa hipótese, utiliza-se o teste F.

- Se o **valor-p** (*p-value*) for **menor que 0,05**, rejeita-se H_0, indicando que X_t **causa Granger** Y_t.
- Caso contrário, não há evidências estatísticas suficientes para afirmar a existência de causalidade de Granger.

O teste pode ser feito em ambas as direções, verificando se X_t causa Y_t e se Y_t causa X_t, configurando situações de causalidade unidirecional ou bidirecional.

2.3 Aprendizado de máquina

O aprendizado de máquina é uma subárea da inteligência artificial (IA) voltada ao desenvolvimento de modelos computacionais que, a partir de dados, conseguem identificar padrões e realizar estimativas ou auxiliar na tomada de decisões, sem que tenham sido programados explicitamente para executar tarefas específicas. Segundo Rautio (2024), ”O aprendizado de máquina, ao contrário das estatísticas tradicionais, geralmente lida com conjuntos de dados grandes e complexos”. Complementando, (Chollet, 2021 apud Rautio, 2024), destaca que ”O aprendizado de máquina, um ramo da Inteligência Artificial (IA), teve um crescimento notável desde a década de 1990, impulsionado por avanços no poder computacional e pela disponibilidade de conjuntos de dados maiores”.

Os algoritmos de aprendizado de máquina podem ser classificados em duas categorias principais: aprendizado supervisionado e aprendizado não supervisionado.

2.3.1 Aprendizado supervisionado

No aprendizado supervisionado, o modelo é treinado utilizando dados de entrada associados às respectivas saídas esperadas. Esse método permite que o algoritmo identifique padrões subjacentes nos dados para realizar previsões ou classificações. Durante o treinamento, o modelo compara as saídas previstas com os resultados reais e ajusta iterativamente seus parâmetros para minimizar os erros. Essa abordagem é amplamente empregada em tarefas de classificação e regressão, sendo útil em contextos onde existem rótulos claros para os dados.

2.3.2 Aprendizado não supervisionado

Diferentemente do aprendizado supervisionado, o aprendizado não supervisionado busca identificar padrões ou estruturas ocultas nos dados sem a necessidade de rótulos. Essa abordagem não exige dados rotulados, concentrando-se em entender como os sistemas podem autonomamente adquirir representações que refletem a estrutura estatística do conjunto de dados. Como mencionado por Johnston; Jones; Kruge (2019 *apud* Rautio, 2024, p. 9), o aprendizado não supervisionado não está vinculado a saídas explícitas ou avaliações ambientais associadas a cada entrada.

No aprendizado não supervisionado, o objetivo é identificar padrões ou estruturas ocultas dentro dos dados. Ao contrário do aprendizado supervisionado, essa abordagem requer apenas a entrada de dados brutos, ou seja, não exige dados rotulados, concentrando-se em entender como os sistemas podem autonomamente adquirir representações que refletem a estrutura estatística do conjunto de dados.

”Ao contrário do aprendizado supervisionado que engloba o conjunto de problemas de ter um conjunto de dados rotulado que pode ser usado para classificar ou ajustar uma linha de regressão, o aprendizado não supervisionado não envolve saídas explícitas ou avaliações ambientais vinculadas a cada entrada”(Rautio, 2024).

Neste trabalho, a abordagem não supervisionada será utilizada para identificar padrões e correlações. O método selecionado para esta análise é o *clustering*, detalhado na próxima seção.

2.3.3 Clustering

O *clustering* é uma técnica de aprendizado de máquina não supervisionado que organiza dados em *clusters* com base em características similares. Cada *cluster* reúne elementos que possuem maior similaridade entre si, em relação aos pertencentes a outros *clusters*. Essa técnica é amplamente utilizada em diversas áreas, incluindo finanças, onde é aplicada para identificar padrões e tendências ocultas que auxiliam

na elaboração de estratégias e na mitigação de riscos(Imeni; Bao; Nozick, 2024; Marti et al., 2017).

Essa técnica é aplicada por investidores para analisar e desenvolver melhores abordagens de negociação, além de ajudar a construir portfólios diversificados. Com isso, os investidores conseguem diminuir as perdas, preservar o capital e apostar em negociações mais arriscadas sem aumentar o risco total (Ezugwu et al., 2022). De acordo com os estudos em (Marti et al., 2017), soluções baseadas em técnicas de *clustering* e de redes, além de melhorar a precisão econômica, podem servir de apoio para projetar políticas financeiras, detectar riscos, desenvolver indicadores que podem prever crises ou para recuperação econômica. Ainda levando em consideração o trabalho de (Marti et al., 2017), ”As previsões de lucro por ação preparadas com base em dados estatisticamente agrupados (*clusters*) superam as previsões feitas em dados agrupados em critérios industriais tradicionais, bem como previsões preparadas por técnicas de extrapolação mecânica”

De forma geral, as técnicas de *clustering* são divididos em dois principais grupos: O hierárquico e o particional.

2.3.3.1 Algoritmos hierárquicos de agrupamento

O *clustering* hierárquico organiza *clusters* de forma hierárquica, ou seja, os dados são particionados em níveis, criando uma estrutura de árvore conhecida como dendrograma. Os *clusters* podem ser organizados de cima para baixo ou de baixo para cima. A abordagem de baixo para cima é denominada método aglomerativo, enquanto a abordagem de cima para baixo é chamada de método divisivo.

No método aglomerativo, cada ponto de dado é tratado inicialmente como um *cluster* individual. Esses pontos são iterativamente mesclados em *clusters* maiores, formando os vários níveis da hierarquia. A cada iteração, os dois *clusters* mais semelhantes (ou próximos) são combinados em um único *cluster*, até que todos os pontos pertençam a um único *cluster* ou atenda-se ao critério de parada. Já o método divisivo inicia com todos os pontos em um único *cluster*, que é dividido iterativamente em subclusters menores até que cada ponto de dado forme um *cluster* único ou atenda-se ao critério de parada.

A união ou separação de subconjuntos de pontos é baseada na generalização da distância entre pontos individuais para a distância entre subconjuntos de pontos. No *clustering* hierárquico, três métricas fundamentais são empregadas para determinar a distância entre pontos: ligação simples, ligação média e ligação completa (Ezugwu et al., 2022).

2.3.3.2 Algoritmo de agrupamento particional

O agrupamento particional visa dividir os dados em um número pré-definido de *clusters*, onde os dados são organizados em grupos sem qualquer estrutura hierárquica. A ideia é que a divisão seja feita de maneira que a similaridade dentro de cada *cluster* seja alta e a similaridade entre os *clusters* seja baixa. O *clustering* particional é um método de agrupamento em que os dados são divididos em um número fixo de *clusters* não hierárquicos, onde cada ponto de dados pertence a exatamente um *cluster*. Diferentemente do *clustering* hierárquico, que constrói uma estrutura hierárquica de *clusters*, o *clustering* particional divide diretamente os dados em um conjunto pré-definido de *clusters*.

A distância euclidiana ou o critério de erro quadrado é amplamente usado no *clustering* particional, sendo o objetivo geral encontrar a partição que minimiza o erro quadrado para um número fixo de *clusters*. Nos algoritmos particionais, o número de *clusters*, comumente chamado de ”K”, precisa ser escolhido antes da execução do algoritmo, o que pode dificultar a definição do valor ideal de K. Essa escolha é especialmente desafiadora em dados do mundo real, frequentemente caracterizados por alta densidade e dimensionalidade, bem como por *clusters* com formas, tamanhos e densidades variadas (Ezugwu et al., 2022).

Ao analisar os dois métodos, é possível identificar vantagens e desvantagens. A principal limitação dos métodos particionais é a necessidade de definir o número K de *clusters* previamente, exigindo o uso de técnicas auxiliares para determinar esse valor, não fazer isso pode resultar em agrupamentos que orientam mal os processos de tomada de decisão. Por outro lado, os métodos hierárquicos não requerem um número pré-definido de *clusters*, permitindo um trabalho mais dinâmico, adaptável às necessidades do projeto. Contudo, os métodos hierárquicos enfrentam desafios em relação à complexidade computacional e à imprecisão dos critérios de terminação. Além disso, seu tempo de execução pode ser elevado em conjuntos de dados grandes e com alta dimensionalidade, pois requerem memória proporcional ao quadrado do número de grupos na partição inicial (Ezugwu et al., 2022).

Por outro lado, agrupamento particional em especial o K-means tende a ser mais rapido que o hierárquico, sendo adequado também a trabalhar com problemas de *clusters* envolvendo grandes conjuntos de dados. Ainda de acordo com as conclusões de (Ezugwu et al., 2022), os algoritmos particionais são preferíveis para a descoberta de padrões e embora os métodos nesses dois grupos tenham se mostrado muito eficazes e eficientes, ambos são sensíveis a ruídos e *outliers* (valores atípicos ou discrepantes dos valores demais) e geralmente dependem do fornecimento de conhecimento prévio ou informações sobre o número exato de *clusters*.

Neste trabalho, será utilizado o método de *clustering* particional K-means, devido à sua adequação à análise de grandes conjuntos de dados e à descoberta de padrões de agrupamento. Segundo (Ahmed; Seraj; Islam, 2020 apud Rautio, 2024), o algoritmo K-means pode ser considerado um dos algoritmos de agrupamento mais poderosos e utilizados na área de pesquisa.

2.3.4 K-Means

O algoritmo K-means é uma técnica de particionamento baseada em centróides, que separa um conjunto de **N** amostras em **K** *clusters* distintos. Cada *cluster* é denotado como **C*n***, onde ***n*** varia de **1** a **K**. Cada *cluster* possui uma média que representa as amostras dentro dele, chamada de centróide, que é uma medida do ponto central do *cluster*. Inicialmente, é necessário definir o número **K** de *clusters*. Em seguida, **K** pontos são escolhidos aleatoriamente dentro do conjunto de amostras para representar os *clusters* ou centróides iniciais. Cada ponto do conjunto de dados é então atribuído ao *cluster* cujo centróide está mais próximo. O método mais comum para determinar essa proximidade é o cálculo da distância euclidiana entre os pontos, embora também possa ser baseado em outras métricas, como a soma dos critérios de erro quadrado.

Após a atribuição inicial dos pontos, os centróides de cada *cluster* são recalculados como a média dos pontos atribuídos a eles. Esse processo de atribuição de pontos e atualização dos centróides se repete iterativamente até que os centróides não mudem mais. Dessa forma, a similaridade intracluster é gradualmente melhorada a cada iteração.

Para utilizar o algoritmo de particionamento, é fundamental determinar o número K ideal de *clusters* para as amostras disponíveis. O K-means é conhecido por ser sensível a essa definição inicial, e uma escolha inadequada pode levar a resultados subótimos (Imeni; Bao; Nozick, 2024; Ezugwu et al., 2022). Identificar com precisão o número ideal de *clusters* em um conjunto de dados pode ser um processo complexo, geralmente realizado por meio da avaliação de partições de *cluster* obtidas após várias iterações do algoritmo com valores variados de K. (Rautio, 2024).

O número ideal de *clusters* é frequentemente identificado com o auxílio de métodos de validação intracluster, sendo os mais utilizados o ***Elbow Method*** (Método do Cotovelo) e o ***Silhouette Score* (Coeficiente de Silhueta)**.

Uma explicação mais detalhada sobre esses métodos e como eles são utilizados para a definição de K *clusters* será apresentada nas próximas seções.

2.3.5 Pré-processamento para o *Cluster*

Como mencionado anteriormente, algoritmos de *clustering* são sensíveis a ruídos e *outliers*, que são pontos distantes do centróide do *cluster* e acabam sendo forçados a integrar o grupo, distorcendo a forma do *cluster* (Ezugwu et al., 2022). Portanto, antes de utilizar o conjunto de dados para gerar o *clustering*, é fundamental realizar um pré-processamento, com o objetivo de eliminar dados ausentes ou normalizar as variáveis. A normalização garante que todas as variáveis estejam na mesma escala, o que melhora o desempenho e a estabilidade do modelo, além de evitar que valores extremos afetem desproporcionalmente a formação dos *clusters*, resultando em agrupamentos que podem não refletir com precisão as características financeiras reais do conjunto de dados.

A remoção de valores discrepantes (*outliers*) é uma etapa essencial na preparação dos dados para análise de *clustering*, pois algoritmos como o K-means podem ser negativamente impactados pela presença desses valores extremos (Marti et al., 2017). De acordo com o estudo de (Rautio, 2024), valores que se distanciam significativamente dos demais podem influenciar excessivamente a formação dos *clusters*, prejudicando a representatividade dos dados financeiros. Ao eliminar esses *outliers*, os algoritmos conseguem identificar *clusters* mais precisos e representativos, baseados nas tendências gerais dos dados.

Além disso, a remoção de *outliers* é crucial para facilitar comparações entre diferentes entidades. Valores extremos podem distorcer as comparações financeiras, criando diferenças artificiais. Ao remover esses pontos, o conjunto de dados se torna mais uniforme, tornando as comparações mais precisas e confiáveis. O processo de remoção de *outliers* envolve a definição de uma função que identifique e elimine esses valores extremos, sendo aplicada a diversas colunas do conjunto de dados para minimizar o impacto desses valores nos modelos estatísticos.

2.3.6 Validação de clustering

Um aspecto crucial ao trabalhar com técnicas de *clustering* é a avaliação dos agrupamentos resultantes, com o objetivo de identificar o que melhor reflete as características dos dados de entrada. Essa avaliação pode ser realizada por meio de três abordagens principais: validação interna, validação relativa e validação externa.

A validação interna foca na análise dos agrupamentos a partir de características intrínsecas dos *clusters*, sem a necessidade de dados rotulados. Nessa abordagem, a qualidade dos *clusters* é avaliada com base em dois principais critérios: coesão (ou agregação), que reflete a similaridade entre os elementos pertencentes ao mesmo *cluster*, e separação entre os *clusters*, que analisa a distância entre os agrupamentos.

Clusters bem formados devem apresentar alta coesão interna e uma boa separação, ou seja, os elementos dentro de um *cluster* devem ser altamente semelhantes, enquanto os *clusters* distintos devem estar suficientemente afastados uns dos outros. Os métodos mais utilizados para validar *clusters* internamente incluem o *Elbow Method*, o *Silhouette Score*, o *Davies-Bouldin Index* e o *Dunn Index*, que ajudam a mensurar a qualidade dos *clusters* em termos de coesão e separação (Rautio, 2024).

A validação relativa, por sua vez, envolve a modificação dos parâmetros do algoritmo de *clustering*, como o número de *clusters* ou os critérios de distância, para observar como essas alterações impactam a formação dos *clusters*. Essa abordagem permite ajustes no modelo de *clustering* de forma iterativa, com o objetivo de identificar a configuração que melhor se adapta aos dados, e fornece os melhores resultados.

Por fim, a validação externa é frequentemente utilizada para comparar os resultados de *clustering* com rótulos ou informações externas preexistentes, como classes ou categorias conhecidas. O objetivo dessa validação é verificar a precisão e a consistência do agrupamento em relação à ”verdade externa”, ou seja, verificar como os *clusters* gerados se alinham com rótulos previamente conhecidos. Esse tipo de validação é especialmente útil para selecionar o método de *clustering* mais adequado ao problema em questão, uma vez que possibilita a comparação dos resultados obtidos com a realidade observada nos dados rotulados.

No contexto deste trabalho, optou-se por não utilizar rótulos ou informações externas preexistentes para validar o *clustering*. Em vez disso, o foco está totalmente nos métodos de validação interna. Os métodos escolhidos para essa tarefa são o *Elbow Method* (Método do Cotovelo) e *Silhouette Score* (Coeficiente de Silhueta), que são amplamente utilizados para identificar o número ideal de *clusters*, ou seja, o valor de K para o algoritmo de *clustering*.

2.3.7 Elbow Method

De forma simplificada, o algoritmo de *clustering*, nesse caso o K-means, é aplicado com diferentes valores de K. Em seguida, calcula-se a soma das distâncias quadráticas dentro de cada *cluster*, também chamada de inércia ou *Sum of Squared Errors* (SEE). À medida que o valor de K aumenta, a inércia tende a diminuir. No entanto, após um certo ponto, a redução da inércia se torna menos expressiva. Esse ponto de inflexão, onde a taxa de diminuição desacelera, é conhecido como o ”cotovelo”e indica o valor ideal de K.

2.3.8 Silhouette Score

O coeficiente de Silhueta mede o quão bem cada ponto está agrupado com os pontos do seu próprio *cluster*, assim é uma métrica que pode ser usada para avaliar a ”qualidade”dos *clusters* formados, determinando quão distantes e distinguíveis os *clusters* são (Vanhala; Järvi; Heikkonen, 2023).

O valor da pontuação é calculado considerando tanto as distâncias intra-*cluster* quanto inter-cluster. A pontuação de silhueta pode ser usada, por exemplo, em técnicas de agrupamento probabilístico, centróide e hierárquico.

Para cada ponto i, o índice de silhueta $s(i)$ é calculado pela Equação 2.6:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \tag{2.6}$$

onde:

- $a(i)$ é a **distância média** entre o ponto i e todos os outros pontos no **mesmo cluster** (coesão);
- $b(i)$ é a **menor distância média** entre o ponto i e os pontos de **qualquer outro cluster** ao qual i não pertence (separação).

Fonte: Rousseeuw (1987).

Interpretação:

O valor de $s(i)$ está no intervalo $[-1, 1]$, e sua interpretação é:

- $s(i) \approx 1$: O ponto está bem atribuído ao seu cluster.
- $s(i) \approx 0$: O ponto está na fronteira entre dois clusters.
- $s(i) \approx -1$: O ponto pode estar atribuído ao cluster errado.

2.4 PCA (Principal Component Analysis)

A análise de Componentes Principais (PCA) é um dos métodos mais comuns de redução de dimensionalidade, ela transforma variáveis correlacionadas em um novo conjunto de variáveis não correlacionadas (componentes principais), preservando o máximo possível da variância dos dados (Vanhala; Järvi; Heikkonen, 2023).A PCA tem como objetivo encontrar o caminho nos dados onde os pontos variam mais, maximizando a variância. Em outras palavras, quais componentes são os mais dominantes(Rautio, 2024).

O primeiro componente principal (PC1) explica a maior parte da variação nos dado. Já O segundo componente (PC2) explica a maior parte da variância restante, e assim por diante.

Seja X uma matriz de dados com n observações e p variáveis, centralizada em relação à média. O primeiro passo do PCA é calcular a matriz de covariância Σ, dada na Formula 2.7:

$$\Sigma = \frac{1}{n-1} X^\top X \tag{2.7}$$

Em seguida, resolvemos o problema de autovalores com a Formula 2.8:

$$\Sigma v = \lambda v \tag{2.8}$$

onde:

- v é o autovetor (direção do componente principal);
- λ é o autovalor correspondente (variância explicada pelo componente).

Os autovalores são ordenados de forma decrescente, e os autovetores associados formam os eixos principais dos dados. A projeção dos dados nos primeiros k componentes principais permite uma representação mais compacta. Formula 2.9:

$$Z = XV_k \tag{2.9}$$

onde V_k é a matriz formada pelos k primeiros autovetores (componentes principais), e Z é a nova representação dos dados com menor dimensionalidade.

Fonte: Jolliffe e Cadima (2016).

O PCA é especialmente útil em contextos onde os dados possuem muitas variáveis correlacionadas, pois permite manter a estrutura essencial dos dados com menos variáveis, facilitando a visualização e a análise subsequente.

2.5 Normalização e Remoção de Outliers com Z-Score

Neste trabalho todos os dados foram normalizados utilizando a normalização *Z-score*. Existem varios métodos para normalização, um bastante conhecido é a normalização com Min-Max. Ao contrário da normalização min-max, que força os dados para o intervalo [0, 1], o *z-score* mantém a distribuição original, apenas centralizando e escalando.

Além da normalização, este trabalho também aplicou uma remoção de outliers baseada no *Z-score*, com o objetivo de eliminar valores que estivessem estatisticamente distantes da média e que pudessem distorcer análises posteriores.

O *z-score* transforma os dados para que tenham média 0 e desvio padrão 1, usando a fórmula 2.10:

$$z = \frac{x - \mu}{\sigma} \tag{2.10}$$

Onde:

- x é o valor original da variável;
- μ é a média dos dados da variável;
- σ é o desvio padrão dos dados da variável;
- z é o valor padronizado (Z-score).

Fonte: Jolliffe e Cadima (2016).

Interpretação:

O valor de z indica quantos desvios padrão (σ) o valor x está distante da média (μ). Especificamente:

- Se $z = 0$, o valor está exatamente na média;
- Se $z > 0$, o valor está acima da média;
- Se $z < 0$, o valor está abaixo da média.

A técnica de normalização é especialmente útil quando se deseja comparar variáveis com escalas diferentes ou quando se aplicam algoritmos que assumem dados centrados e escalados, como regressão linear, análise de componentes principais (PCA) e algoritmos baseados em distância, como o K-means.

Já na técnica de remoção de outliers, embora utilize a mesma fórmula apresentada anteriormente, o *Z-score* neste contexto tem uma função distinta: identificar valores atípicos (outliers). Para isso, considera-se que qualquer valor cujo *Z-score* absoluto seja superior a um determinado limite, comumente $|z| > 3, estexcessivamentedistantedamdia, sendocla$

3 ESTRATÉGIA METODOLÓGICA

Este capítulo descreve detalhadamente os procedimentos adotados para investigar a influência dos investimentos estrangeiros sobre o Ibovespa e ações brasileiras. Inicialmente, apresentam-se as ferramentas utilizadas no desenvolvimento do modelo, seguidas da descrição das fontes de dados, que incluem os fluxos de capital estrangeiro e os preços do IBOV e das ações, bem como o cálculo dos indicadores. Na sequência, são explicadas as etapas de preparação e normalização dos dados, os critérios de seleção das séries temporais e as diferentes janelas de análise.

3.1 Ferramentas Utilizadas

Para o desenvolvimento deste trabalho e análise dos resultados, foi utilizada a linguagem de programação Python, juntamente com suas bibliotecas especializadas.

3.1.1 Python

Python é uma linguagem de programação de alto nível, de propósito geral, amplamente utilizada na área de ciência de dados. Destaca-se por sua sintaxe simples, legível e de fácil implementação. É considerada uma das linguagens mais populares do mundo, o que se deve, em grande parte, à sua vasta comunidade ativa e à ampla disponibilidade de bibliotecas para as mais diversas finalidades.

As bibliotecas do Python consistem em conjuntos de códigos prontos que podem ser importados e reutilizados, o que permite ao usuário executar tarefas complexas de forma rápida e eficiente, sem a necessidade de desenvolver algoritmos do zero.

Por conta dessas características, o Python foi a linguagem escolhida para o desenvolvimento desse trabalho. Ele oferece bibliotecas que possibilitam desde a coleta e o tratamento de dados, até a aplicação de métodos estatísticos como a correlação de Pearson, regressão linear e causalidade de Granger, além da implementação prática de algoritmos de agrupamento como o K-means. Também conta com ferramentas robustas para visualização de dados, como gráficos e tabelas.

A seguir, são listadas as principais bibliotecas utilizadas:

- **yfinance**: Utilizada para recuperar dados históricos do IBOV e das ações direto da plataforma Yahoo Finance.
- **Pandas**: Ótimo para manipular e organizar os dados em estruturas.

- **Matplotlib e Seaborn**: Criação de gráficos e visualizações para análise dos resultados.
- **scipy.stats**: Usada para estatísticas descritivas e testes, incluindo pearsonr, linregress, zscore.
- **sklearn.cluster**: Para a implementação do algoritmo KMeans.
- **sklearn.metrics**: Para o cálculo do Silhouette Score.
- **sklearn.decomposition**: Para aplicação de PCA (Análise de Componentes Principais).
- **statsmodels.tsa.stattools**: Realização do teste de causalidade de Granger com grangercausalitytests.

3.2 Coleta dos dados

Para a realização dos testes, foram utilizados três conjuntos principais de dados: os investimentos estrangeiros, os preços do índice Ibovespa (IBOV) e os preços de ações brasileiras. A coleta foi realizada com o uso de bibliotecas específicas da linguagem Python, a fim de garantir a atualização, integridade e precisão das informações.

3.2.1 Investimentos Estrangeiros

Os dados referentes aos investimentos estrangeiros foram obtidos por meio de um relatório diário emitido pela B3. Este relatório é disponibilizado em formato PDF, sendo necessário utilizar a biblioteca pdfplumber, que permite a extração estruturada das informações contidas no documento.

Embora o relatório seja emitido diariamente, os dados relativos aos investimentos estrangeiros não são disponibilizados todos os dias. Além disso, quando publicados, geralmente referem-se a transações ocorridas com dois dias de atraso em relação à data de divulgação.

O principal indicador utilizado é a participação percentual dos investimentos, dada por:

$$\text{Participação Percentual} = \left(\frac{\text{Compras} - \text{Vendas}}{\text{Compras} + \text{Vendas}} \right) \times 100$$

Este indicador percentual foi utilizado para investigar a correlação entre o fluxo estrangeiro e o índice IBOV, bem como entre o fluxo estrangeiro e as ações individuais.

3.2.2 IBOV e Ações

Os dados históricos de preços do índice Ibovespa e das ações selecionadas foram coletados por meio da biblioteca *yfinance*, que fornece acesso à *Application Programming Interface* (API) do Yahoo Finance. Para isso, são utilizados os chamados *tickers*, que são códigos de negociação dos ativos financeiros.

- Para o índice Ibovespa, foi utilizado o *ticker*: ^BVSP.
- Para as ações, exemplos de *tickers* utilizados incluem: VALE3.SA, PETR4.SA, ITUB4.SA, BBAS3.SA, entre outros.

Foram coletadas, para cada *ticker*, as datas de negociação e os valores de fechamento.

Após a coleta dos preços de fechamento, foi realizado o cálculo dos retornos percentuais diários, tanto para o IBOV quanto para as ações. Esse cálculo é necessário para transformar os preços absolutos em uma série que represente a variação percentual diária dos ativos, permitindo a análise de desempenho, volatilidade e correlação com outras variáveis. A fórmula utilizada é:

$$R_t = \left(\frac{P_t - P_{t-1}}{P_{t-1}} \right) \times 100$$

Onde:

- R_t representa o retorno percentual no dia t;
- P_t é o preço de fechamento do ativo no dia t;
- P_{t-1} é o preço de fechamento do ativo no dia anterior ($t-1$).

Essa fórmula é utilizada para calcular a variação percentual diária com base nos preços de fechamento. O objetivo é transformar os valores brutos de preço em uma métrica relativa, que permite comparar diferentes ativos entre si e realizar análises estatísticas, como correlação, regressão e clustering.

3.3 Preparação e Normalização

Após a coleta dos dados, é feita a preparação e normalização desses dados para, então, utilizá-los nos métodos de correlação propostos. Essa etapa tem o objetivo de garantir a consistência, a dimensionalidade, a escalabilidade e a comparabilidade dos dados para análises estatísticas e de *clusters*. Dependendo do objetivo, essas etapas podem ser realizadas mais de uma vez e em ordens diferentes.

3.3.1 Alinhamento Temporal

Como mencionado anteriormente, os dados da B3 são divulgados com um atraso de dois dias. Por esse motivo, há datas em que não estão disponíveis informações sobre os investimentos estrangeiros. Além disso, pode haver datas em que o IBOV ou determinadas ações não apresentem registros válidos.

Assim, é necessário alinhar cronologicamente todas as séries, utilizando como referência as datas em que todos os conjuntos de dados possuem valores válidos. Isso garanta que os cálculos estatísticos sejam realizados sobre pares de dados consistentes.

3.3.1.1 Remoção de Outliers e Normalização

Alguns dos métodos utilizados neste trabalho, como regressão linear, causalidade de Granger, K-means e PCA, são sensíveis à presença de *outliers*, pois podem distorcer medidas estatísticas como média, correlação e inclinação de regressões.

Já a normalização dos dados elimina efeitos de escala, evitando a comparação direta entre valores brutos, que podem ser muito altos, e dados percentuais. Exceto a correlação de Pearson, que é insensível a esses valores (embora possa ser normalizada), os métodos de Granger, K-means e PCA são sensíveis.

Neste trabalho, é utilizada a normalização Z-score. Além disso, foi adotado um procedimento de remoção de outliers também baseado no Z-score, no qual pares de valores foram descartados caso apresentassem Z-score superior a um limite predefinido.

3.3.2 Remoção de dados ausentes

Mesmo após o alinhamento temporal, ainda podem existir pares de dados em que um ou ambos apresentem valores ausentes, nulos ou inconsistentes. Após a remoção desses dados, os métodos deixam de ser influenciados por essas lacunas.

3.4 Aplicação dos Métodos

Após as etapas de coleta, preparação e normalização dos indicadores, os dados estão prontos para a aplicação das técnicas estatísticas propostas neste trabalho, como a correlação de Pearson, a causalidade de Granger e K-Means.

Esses processos foram realizados para diferentes janelas de séries temporais. Inicialmente, foram selecionados períodos brutos de 30, 90, 150, 210 e 300 dias corridos. No entanto, após o alinhamento temporal com os dados, preparação e normali-

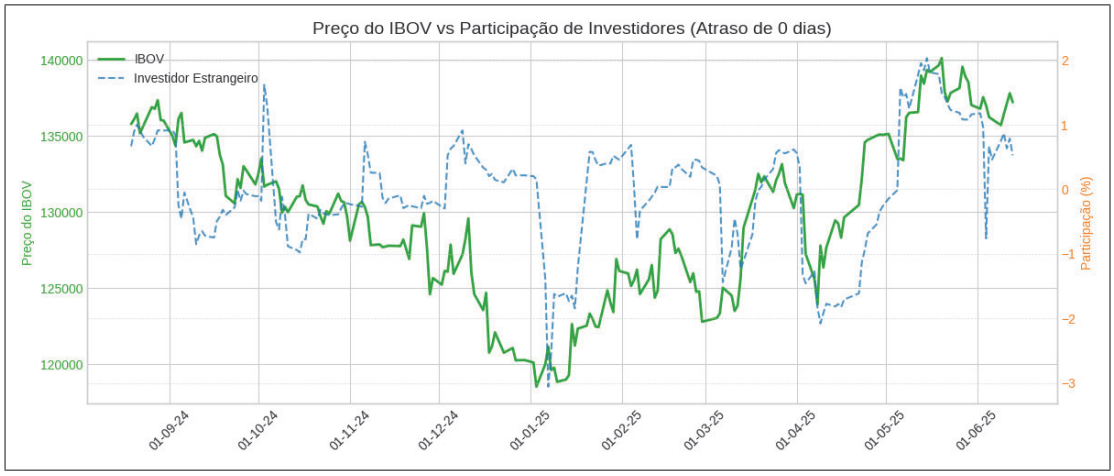
zação, essas janelas foram ajustadas para conter, respectivamente, 22, 62, 101, 138 e 202 observações válidas.

4 RESULTADOS

Após a execução de todo o processo descrito no capítulo anterior, os dados foram aplicados nas análises de correlação de Pearson, regressão linear, causalidade de Granger, K-means e PCA, considerando diferentes janelas de séries temporais, com 22, 62, 101, 138 e 202 pares de amostras. Inicialmente, foi analisada a relação entre o fluxo de investimento estrangeiro e o IBOV. Em seguida, investigou-se a relação entre o IBOV e as ações mencionadas, e, por fim, entre o fluxo de investimento estrangeiro e essas mesmas ações. A partir dessas análises, é possível avaliar a intensidade da relação entre essas variáveis, bem como investigar possíveis relações de causa e efeito. Para isso, foram utilizados três métodos principais: correlação de Pearson, regressão linear e teste de causalidade de Granger.

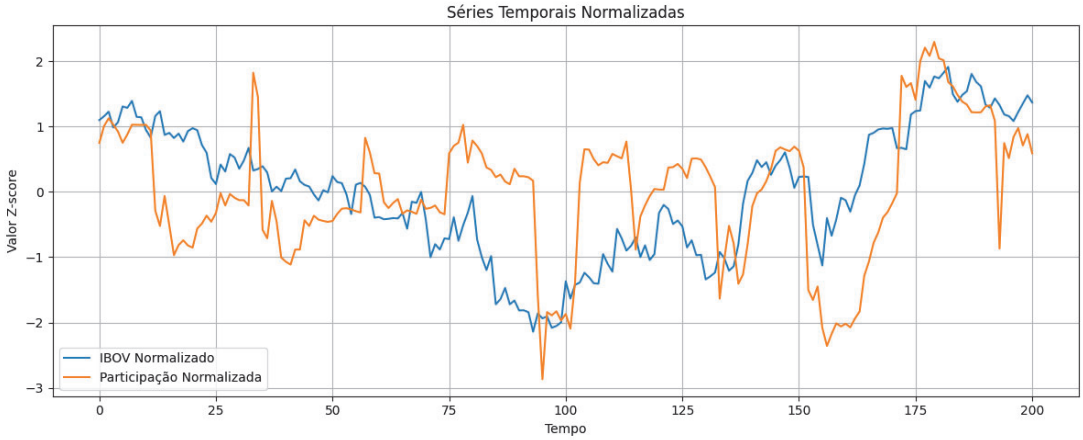
Os dados utilizados nesta pesquisa compreendem o período de 2024-08-19 a 2025-06-14. A (Figura 1) apresenta o gráfico da relação dos valores do IBOV com a participação dos investidores durante o período mencionado. Já a (Figura 2) apresenta esse mesmo gráfico, mas com os dados já normalizados.

Figura 1 – Relação entre preço do IBOV e participação de investidores



Fonte: elaborado pelo autor (2025)

Figura 2 – Séries Temporais Normalizadas



Fonte: elaborado pelo autor (2025)

A (Tabela 1) apresenta os *Tickers* usados nesse trabalho para adquirir os dados das ações. Nela é possível ver o nome da empresa relacionado ao *ticker* e qual o setor atuante.

Tabela 1 – Empresas analisadas na pesquisa

Ticker	Nome da Empresa	Setor
VALE3.SA	Vale S.A.	Mineração
PETR4.SA	Petrobras PN	Petróleo e Gás
ITUB4.SA	Itaú Unibanco	Financeiro
B3SA3.SA	B3 S.A. (Brasil, Bolsa, Balcão)	Financeiro
BBAS3.SA	Banco do Brasil	Financeiro
ABEV3.SA	Ambev S.A.	Bebidas
WEGE3.SA	Weg S.A.	Industrial
PETR3.SA	Petrobras ON	Petróleo e Gás
MGLU3.SA	Magazine Luiza	Varejo
RENT3.SA	Localiza	Transportes
BBDC4.SA	Bradesco PN	Financeiro
BBDC3.SA	Bradesco ON	Financeiro
SANB11.SA	Santander Brasil (Units)	Financeiro
ELET3.SA	Eletrobras ON	Energia Elétrica
ELET6.SA	Eletrobras PNB	Energia Elétrica
ENGI11.SA	Engie Brasil (Units)	Energia Elétrica
GGBR4.SA	Gerdau PN	Siderurgia
CSNA3.SA	CSN (Companhia Siderúrgica Nacional)	Siderurgia
USIM5.SA	Usiminas PNA	Siderurgia
LREN3.SA	Lojas Renner	Varejo
AMER3.SA	Americanas S.A.	Varejo
HAPV3.SA	Hapvida	Saúde
RADL3.SA	Raia Drogasil	Saúde
HYPE3.SA	Hypera Pharma	Saúde
RAIL3.SA	Rumo Logística	Transportes
AZUL4.SA	Azul Linhas Aéreas PN	Transportes
GOLL4.SA	Gol Linhas Aéreas PN	Transportes
CYRE3.SA	Cyrela Brazil Realty	Construção Civil
MRVE3.SA	MRV Engenharia	Construção Civil
EZTC3.SA	Eztec	Construção Civil
POSI3.SA	Positivo Tecnologia	Tecnologia

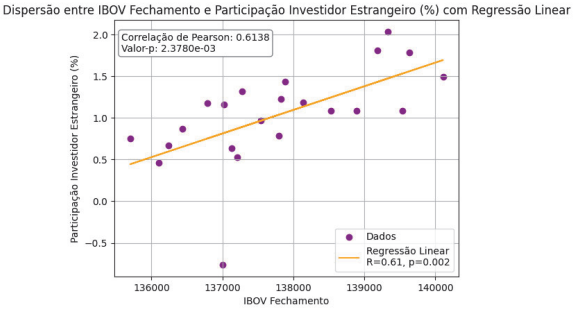
Fonte: elaborado pelo autor (2025)

4.1 Fluxo de Investimentos Estrangeiros e IBOV

4.1.1 Correlação de Pearson e Causalidade de Granger

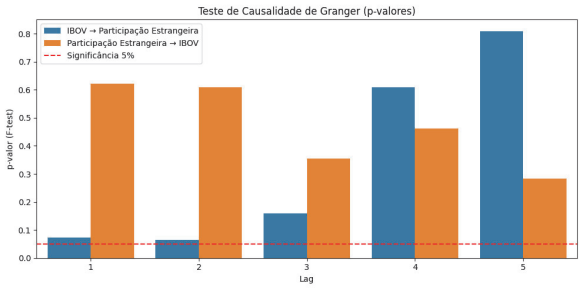
A seguir, são exibidos os gráficos de dispersão com regressão linear e os gráficos de causalidade de Granger para cada uma das séries temporais.

Figura 3 – Gráfico de Dispersão (IBOV - Estrangeiro) - 22 pares



Fonte: elaborado pelo autor (2025)

Figura 4 – Gráfico de Causalidade de Granger (IBOV - Estrangeiro) - 22 pares

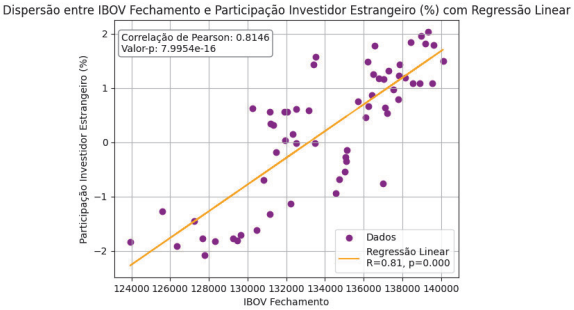


Fonte: elaborado pelo autor (2025)

A (Figura 3) apresenta a correlação de Pearson para 22 pares de amostras, com um coeficiente de correlação r = 0.6138 e valor-p = 2.3780e-03, indicando uma correlação positiva moderada e estatisticamente significativa. A regressão linear reforça essa relação, com R=0.61 e p=0.002.

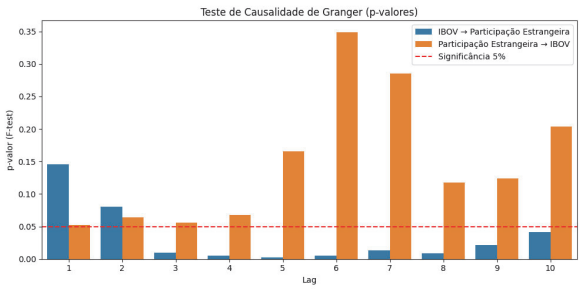
Já a (Figura 4) apresenta o gráfico de causalidade de Granger, que mostra a influência do IBOV sobre a participação estrangeira em 0 defasagens (lags), e a influência da participação estrangeira sobre o IBOV em 0 defasagens. **Não foi identificada causalidade em nenhuma direção**, não sendo possível definir influência temporal.

Figura 5 – Gráfico de Dispersão (IBOV - Estrangeiro) - 62 pares



Fonte: elaborado pelo autor (2025)

Figura 6 – Gráfico de Causalidade de Granger (IBOV - Estrangeiro) - 62 pares

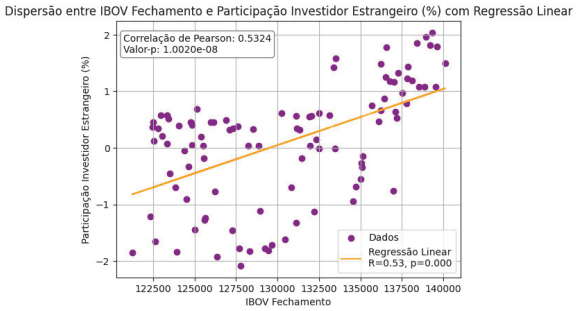


Fonte: elaborado pelo autor (2025)

A (Figura 5) apresenta a correlação de Pearson para 62 pares de amostras, com um coeficiente de correlação r = 0.8146 e valor-p = 7.9954e-16, revelando uma forte relação entre os dados. Também é exibida uma regressão linear, com R=0.81 e p=0.000.

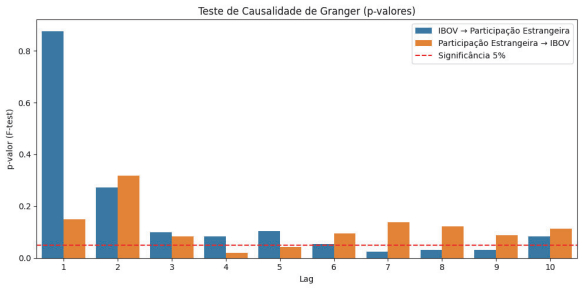
Já a (Figura 6) demonstra o gráfico de causalidade de Granger, que mostra a influência do IBOV sobre a participação estrangeira em 8 defasagens (lags) e a influência da participação estrangeira sobre o IBOV em 0 defasagens. Nesse caso, pode ser identificada **influência do IBOV sobre os investimentos**, mas não há influência inversa. Isso indica que o IBOV pode influenciar futuros investidores estrangeiros.

Figura 7 – Gráfico de Dispersão (IBOV - Estrangeiro) - 101 pares



Fonte: elaborado pelo autor (2025)

Figura 8 – Gráfico de Causalidade de Granger (IBOV - Estrangeiro) - 101 pares

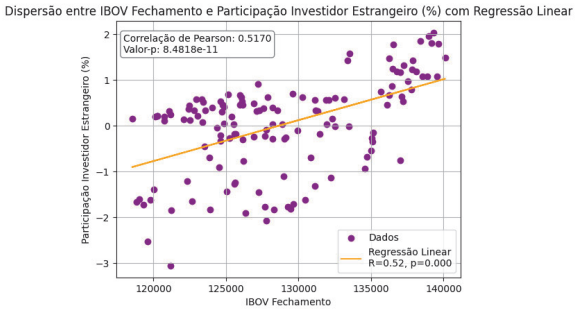


Fonte: elaborado pelo autor (2025)

A (Figura 7) apresenta a correlação de Pearson para 101 pares de amostras, com um coeficiente de correlação r = 0.5324 e valor-p = 1.0020e-08. Também é exibida uma regressão linear, com R=0.53 e p=0.000. Um valor menor que nas amostragens anteriores, mas ainda moderado e significativo.

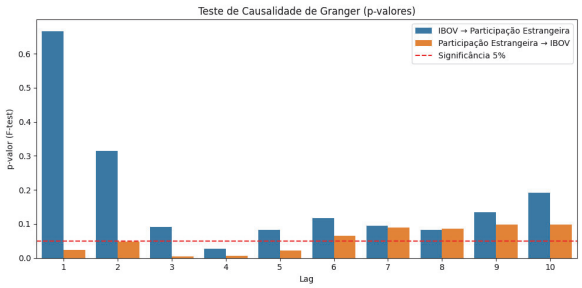
Já a (Figura 8) demonstra o gráfico de causalidade de Granger, que mostra a influência do IBOV sobre a participação estrangeira em 3 defasagens (lags) e a influência da participação estrangeira sobre o IBOV em 2 defasagens. Nesse caso, pode ser identificada uma **causalidade mútua** e **dinâmica**.

Figura 9 – Gráfico de Dispersão (IBOV - Estrangeiro) - 138 pares



Fonte: elaborado pelo autor (2025)

Figura 10 – Gráfico de Causalidade de Granger (IBOV - Estrangeiro) - 138 pares

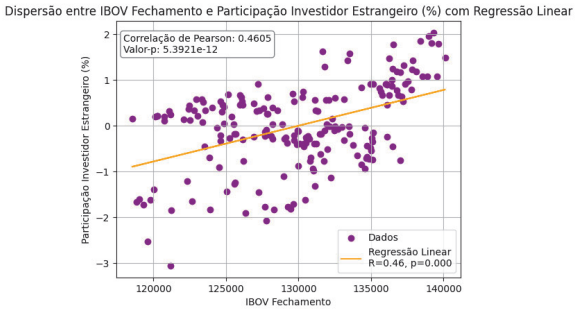


Fonte: elaborado pelo autor (2025)

A (Figura 9) apresenta a correlação de Pearson para 138 pares de amostras, com um coeficiente de correlação r = 0.5170 e valor-p = 8.4818e-11. Também é exibida uma regressão linear, com R=0.52 e p=0.000. Exibe correlação moderada.

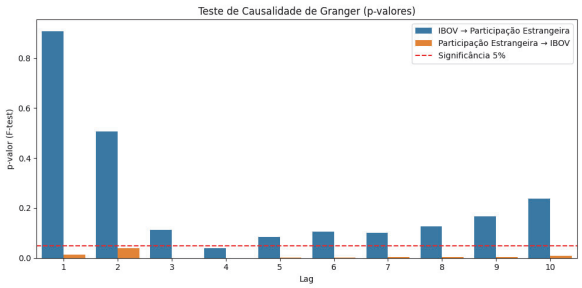
Já a (Figura 10) demonstra o gráfico de causalidade de Granger, que mostra a influência do IBOV sobre a participação estrangeira em 1 defasagem (lag) e a influência da participação estrangeira sobre o IBOV em 5 defasagens. Nesse caso, a causalidade **ainda é mútua**, mas a **participação estrangeira influencia mais o IBOV**.

Figura 11 – Gráfico de Dispersão (IBOV - Estrangeiro) - 202 pares



Fonte: elaborado pelo autor (2025)

Figura 12 – Gráfico de Causalidade de Granger (IBOV - Estrangeiro) - 202 pares



Fonte: elaborado pelo autor (2025)

A (Figura 11) apresenta a correlação de Pearson para 202 pares de amostras, com um coeficiente de correlação r = 0.4605 e valor-p = 5.3921e-12. Também é exibida uma regressão linear, com R=0.46 e p=0.000. Um resultado abaixo dos anteriores, mas com uma correlação moderada.

Já a (Figura 12) demonstra o gráfico de causalidade de Granger, que mostra a influência do IBOV sobre a participação estrangeira em 1 defasagem (lag) e a influência da participação estrangeira sobre o IBOV em 10 defasagens. Nesse caso, existe

apenas um lag de IBOV -> Estrangeiro, e na direção contrária existem dez lags, o que demonstra uma grande influência de investimentos sobre IBOV e com **efeito prolongado**.

A (Tabela 2) apresenta o resumo dos resultados da correlação de Pearson, assim como da regressão linear entre o fluxo de investimento estrangeiro com IBOV. É possível identificar uma queda no valor de correlação ao longo do aumento dos pares utilizados.

Já a (Tabela 3) apresenta os resultados da causalidade de Granger entre o IBOV e o fluxo de investimento estrangeiro. É possível perceber que inicialmente não existe causalidade em nenhuma direção, e ao ponto em que o número de pares aumenta, a causalidade passa a ser unilateral na direção IBOV -> Fluxo, depois bidirecional, e por fim unilateral com a direção tendendo de Fluxo -> IBOV.

Tabela 2 – Resultados: Correlação e Regressão (IBOV e Estrangeiro)

Pares	Correlação Pearson	Valor-p (Pearson)	Regressão Linear
22	0.6138	2.3780e-03	R=0.61, p=0.002
62	0.8146	7.9954e-16	R=0.81, p=0.000
101	0.5324	1.0020e-08	R=0.53, p=0.000
138	0.5170	8.4818e-11	R=0.52, p=0.000
202	0.4605	5.3921e-12	R=0.46, p=0.000

Fonte: elaborado pelo autor (2025)

Tabela 3 – Resultados: Causalidade de Granger e Lags (IBOV e Estrangeiro)

Pares	Granger (IBOV → Fluxo)	Lags (IBOV → Fluxo)	Granger (Fluxo → IBOV)	Lags (Fluxo → IBOV)
22	NÃO	-	NÃO	-
62	SIM	3,4,5,6,7,8,9,10	NÃO	-
101	SIM	7,8,9	SIM	4,5
138	SIM	4	SIM	1,2,3,4,5
202	SIM	4	SIM	1,2,3,4,5,6,7,8,9,10

Fonte: elaborado pelo autor (2025)

Ao comparar os resultados das séries temporais, observa-se que a correlação diminui conforme o número de pares aumenta. Isso é esperado, pois com amostras maiores, a variabilidade tende a aumentar. Ainda assim, todas as correlações encontradas são estatisticamente significativas.

Já na análise de causalidade de Granger, observa-se evidência de causalidade apenas a partir de 62 pares, chegando até a apresentar relações bidirecionais. No

entanto, ao final, observa-se uma forte evidência de que os **investidores influenciam o IBOV com maiores defasagens**.

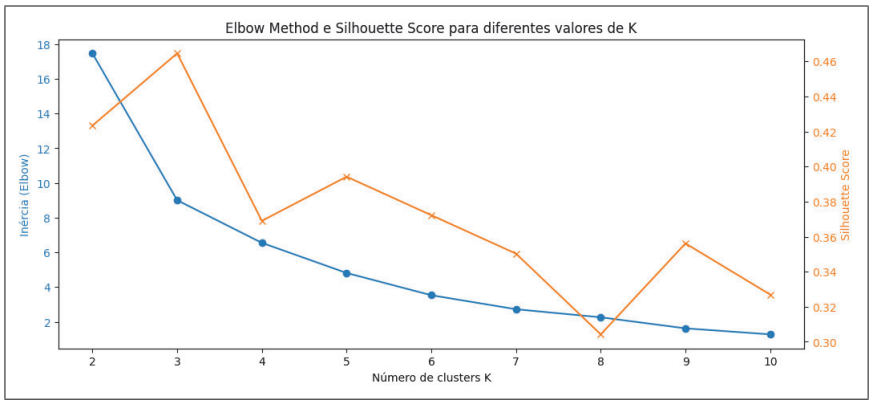
Dessa forma, pode-se sugerir que o IBOV impacta as decisões dos investidores estrangeiros de maneira rápida, enquanto os investidores afetam o IBOV de forma mais prolongada.

4.1.2 K-Means e PCA

A seguir, são exibidos o *Elbow Method* e *Silhoutte Score*, *cluster* e o PCA para cada um das séries temporais.

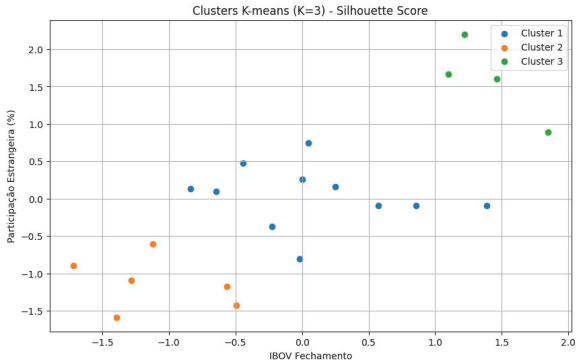
4.1.2.1 Série Temporal - 22 pares

Figura 13 – Elbow Method e Silhoutte Score - K (IBOV - Estrangeiro) (22 pares)



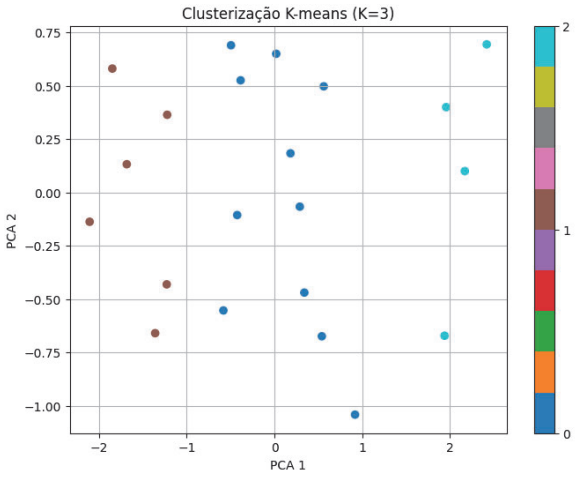
Fonte: elaborado pelo autor (2025)

Figura 14 – Cluster k-Means (IBOV - Estrangeiro) - 22 pares



Fonte: elaborado pelo autor (2025)

Figura 15 – PCA (IBOV - Estrangeiro) - 22 pares



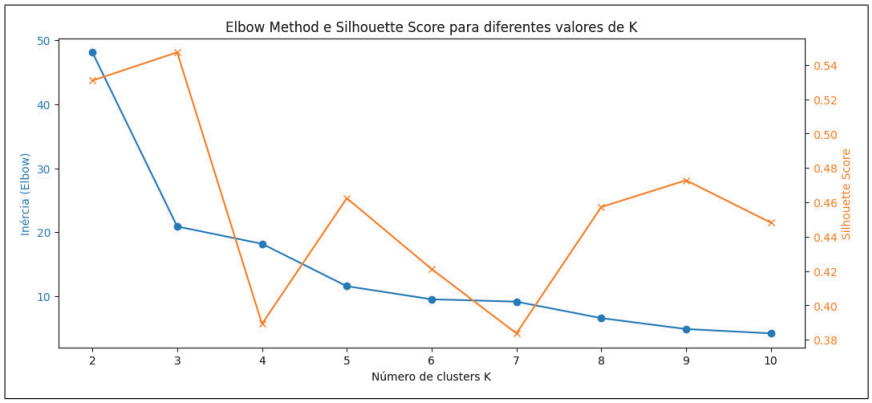
Fonte: elaborado pelo autor (2025)

A (Figura 13) apresenta um gráfico com os resultados do *Elbow Method* e *Silhouette Score* para 22 pares. Após a análise, é possível verificar que ambos apontam o melhor valor de K = 3.

Na (Figura 14) apresentam-se *clusters* pouco visíveis e com sobreposição. **Não há padrão claro**. Esse comportamento se repete no PCA, conforme mostrado na (Figura 15).

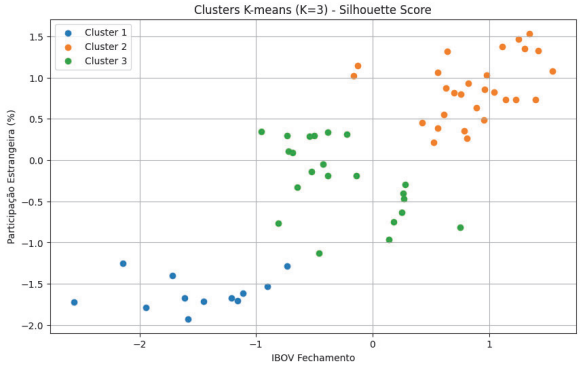
4.1.2.2 Série Temporal - 62 pares

Figura 16 – Elbow Method e Silhoutte Score - K (IBOV - Estrangeiro) (62 pares)



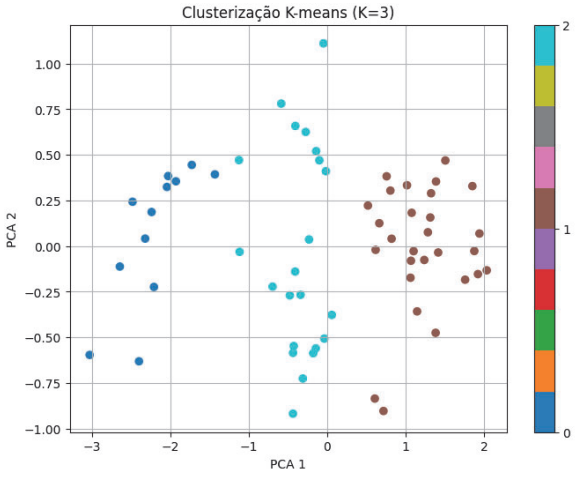
Fonte: elaborado pelo autor (2025)

Figura 17 – Cluster k-Means (IBOV - Estrangeiro) - 62 pares



Fonte: elaborado pelo autor (2025)

Figura 18 – PCA (IBOV - Estrangeiro) - 62 pares



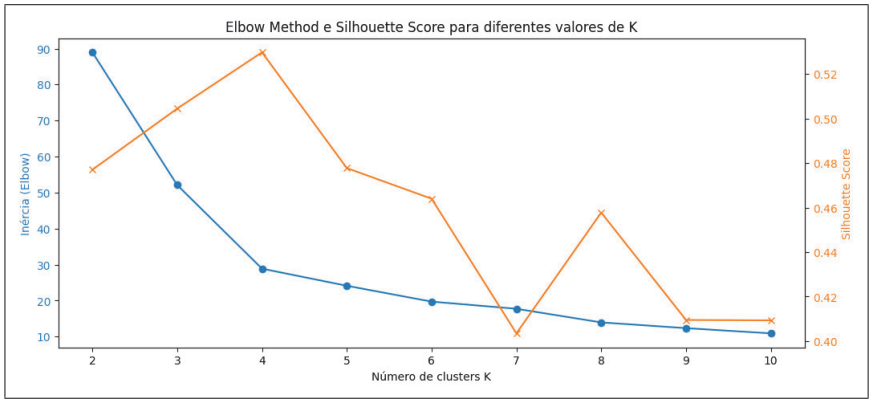
Fonte: elaborado pelo autor (2025)

A (Figura 16) apresenta um gráfico com os resultados de *Elbow Method* e *Silhouette Score* para o valor de 62 pares. Após a análise, é possível verificar que ambos apontam o melhor valor de K = 3.

Nas (Figura 17) e (Figura 18) observa-se a formação de três aglomerados distintos. Nesse caso, já é possível identificar uma tendência: quanto maior o IBOV, maior é a participação estrangeira no agrupamento laranja, enquanto o agrupamento azul apresenta baixa participação e baixo IBOV, podendo indicar momentos de baixa e fuga no investimento estrangeiro.

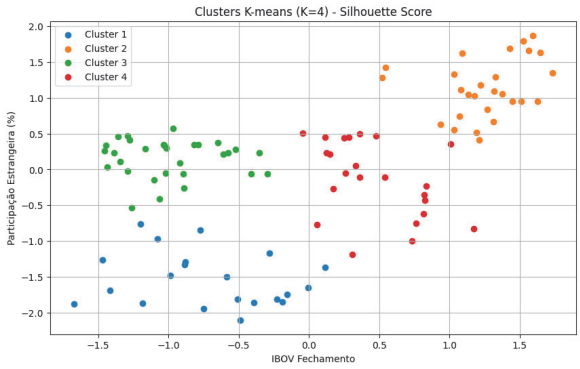
4.1.2.3 Série Temporal - 101 pares

Figura 19 – Elbow Method e Silhoutte Score - K (IBOV - Estrangeiro) (101 pares)



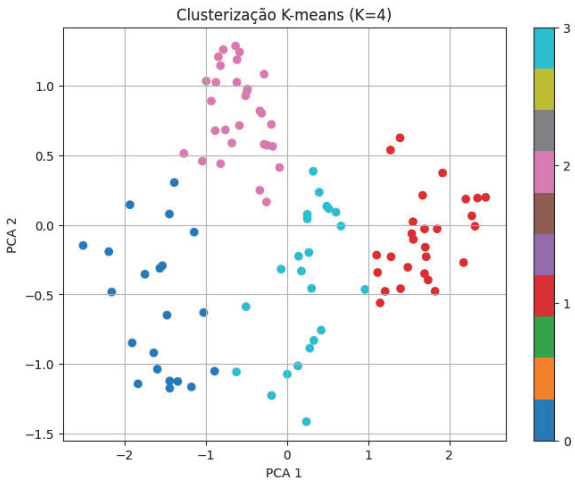
Fonte: elaborado pelo autor (2025)

Figura 20 – Cluster k-Means (IBOV - Estrangeiro) - 101 pares



Fonte: elaborado pelo autor (2025)

Figura 21 – PCA (IBOV - Estrangeiro) - 101 pares



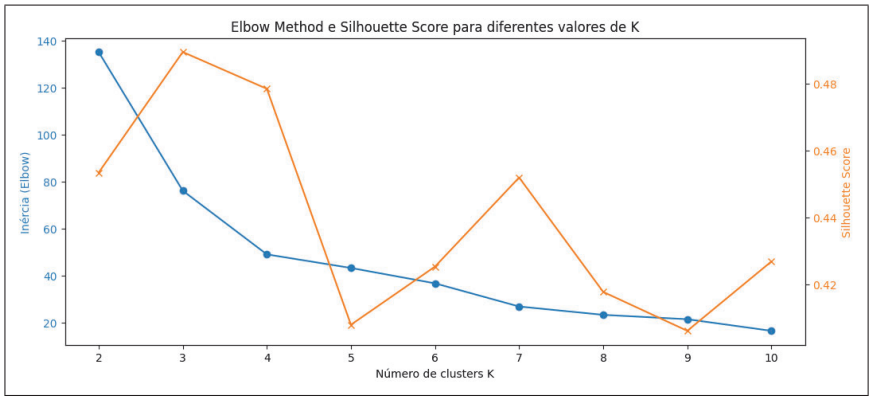
Fonte: elaborado pelo autor (2025)

A (Figura 19) apresenta um gráfico com os resultados de *Elbow Method* e *Silhouette Score* para o valor de 101 pares. Após a análise, é possível verificar que ambos apontam o melhor valor de K = 4.

Nas (Figura 20) e (Figura 21) apresenta-se mais um agrupamento, em comparação com os anteriores, e a separação é mais clara. Esse novo agrupamento é intermediário.

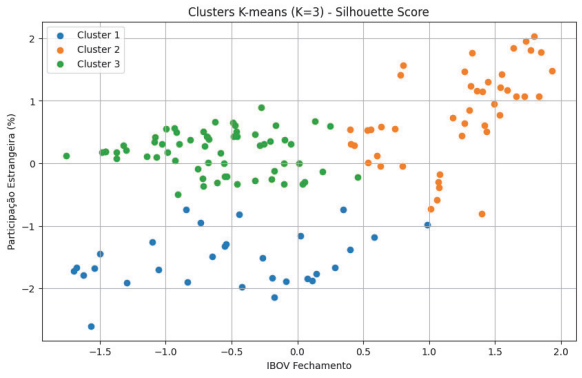
4.1.2.4 Série Temporal - 138 pares

Figura 22 – Elbow Method e Silhoutte Score - K (IBOV - Estrangeiro) (138 pares)



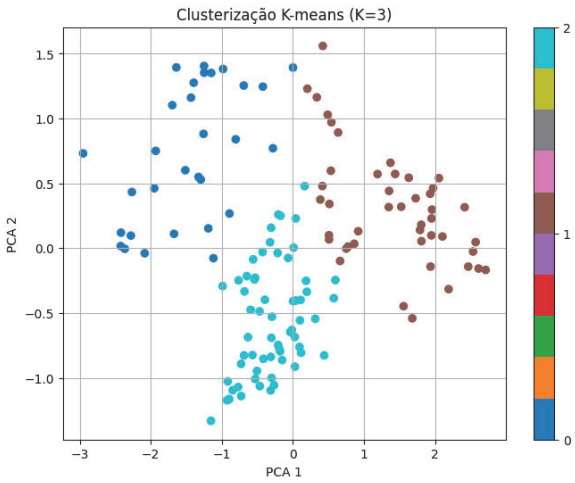
Fonte: elaborado pelo autor (2025)

Figura 23 – Cluster k-Means (IBOV - Estrangeiro) - 138 pares



Fonte: elaborado pelo autor (2025)

Figura 24 – PCA (IBOV - Estrangeiro) - 138 pares



Fonte: elaborado pelo autor (2025)

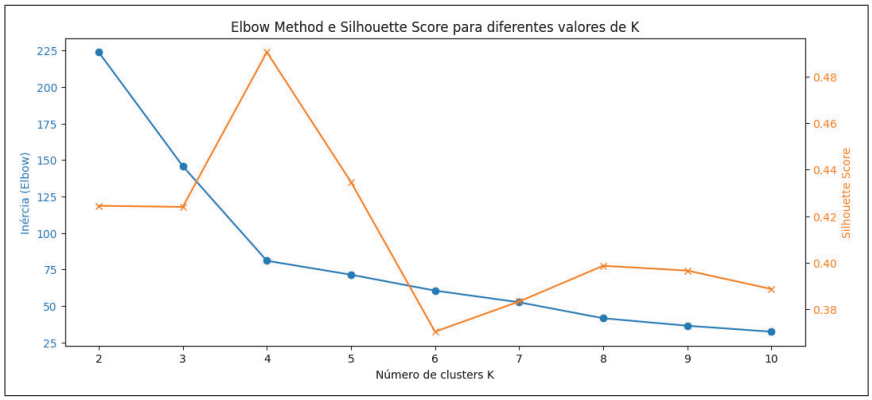
A (Figura 22) apresenta um gráfico com os resultados de *Elbow Method* e *Silhouette Score* para o valor de 138 pares. Após a análise, é possível verificar que ambos apontam o melhor valor de K = 3.

Na (Figura 23), os *clusters* seguem os padrões das séries anteriores, mas o agrupamento verde contém mais dados. Como está no centro, pode indicar neutrali-

dade no mercado. Da mesma forma, a (Figura 24) apresenta o PCA com essas mesmas caracteristicas.

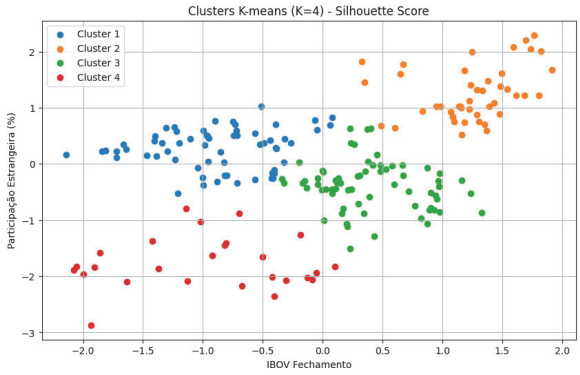
4.1.2.5 Série Temporal - 200 pares

Figura 25 – Elbow Method e Silhoutte Score - K (IBOV - Estrangeiro) (200 pares)



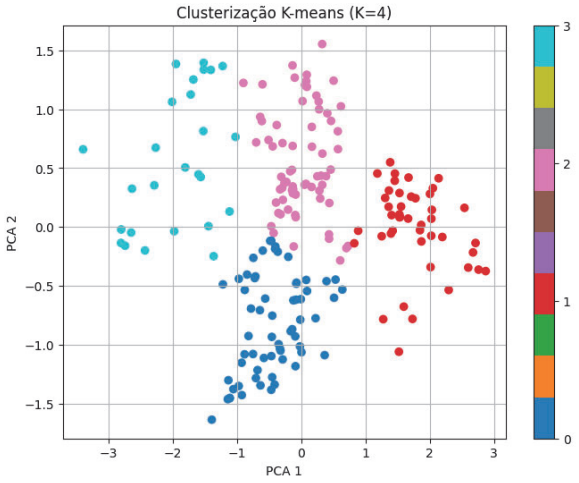
Fonte: elaborado pelo autor (2025)

Figura 26 – Cluster k-Means (IBOV - Estrangeiro) - 200 pares



Fonte: elaborado pelo autor (2025)

Figura 27 – PCA (IBOV - Estrangeiro) - 200 pares



Fonte: elaborado pelo autor (2025)

A (Figura 25) apresenta um gráfico com os resultados de *Elbow Method* e *Silhouette Score* para o valor de 138 pares. Após a análise, é possível verificar que ambos apontam o melhor valor de K = 4.

Já na (Figura 26), os *clusters* apresentam maior separação e mais robustez. É possível verificar um agrupamento com alta participação estrangeira e alto IBOV, e outro com baixa participação e IBOV negativo.

Na (Figura 27), é possível verificar pontos mais compactos e centralizados.

Logo, com poucos pontos, os *clusters* são pouco informativos. No entanto, à medida que o número de pontos aumenta, eles se tornam mais bem definidos e começam a revelar padrões. Da mesma forma, com o aumento das amostras, o PCA passa a extrair com mais precisão as variações relevantes.

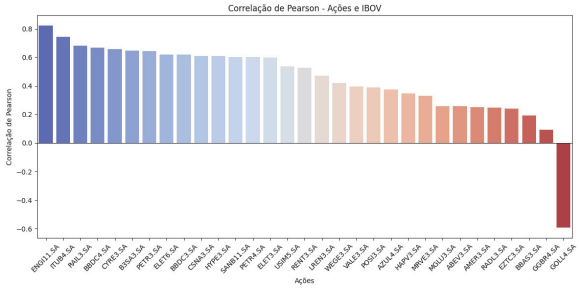
Em quase todas as visualizações, observa-se um agrupamento de pontos com alta participação estrangeira e valores elevados do IBOV, sugerindo uma relação de influência entre ambos e uma correlação positiva.

Dessa forma, fica evidente que o uso do K-means mostrou-se eficiente na segmentação do comportamento entre o IBOV e a participação estrangeira.

4.2 IBOV e Ações

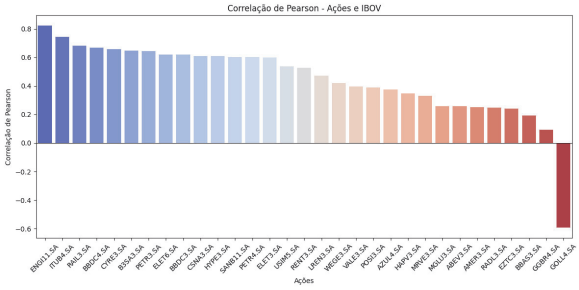
A seguir, são exibidos os gráficos de correlação de Pearson entre o IBOV e as ações em cada um das séries temporais.

Figura 28 – Correlação de Pearson - IBOV e Ações (22 pares)



Fonte: elaborado pelo autor (2025)

Figura 29 – Correlação de Pearson - IBOV e Ações (62 pares)

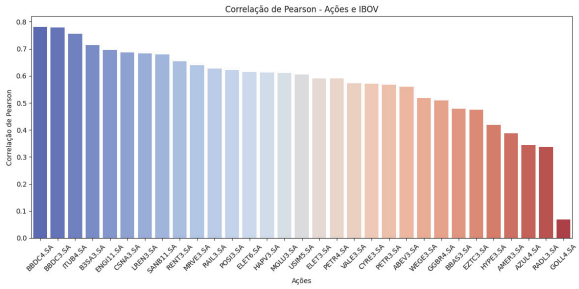


Fonte: elaborado pelo autor (2025)

A (Figura 28) apresenta uma alta volatilidade nas correlações, que oscilam bastante. Várias ações mostram momentos de descorrelação, ou até mesmo correlação negativa.

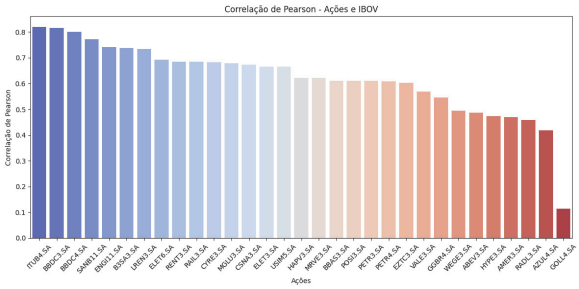
Já na (Figura 29), a correlação se mostra mais estável, embora ainda apresente momentos de oscilação.

Figura 30 – Correlação de Pearson - IBOV e Ações (101 pares)



Fonte: elaborado pelo autor (2025)

Figura 31 – Correlação de Pearson - IBOV e Ações (138 pares)

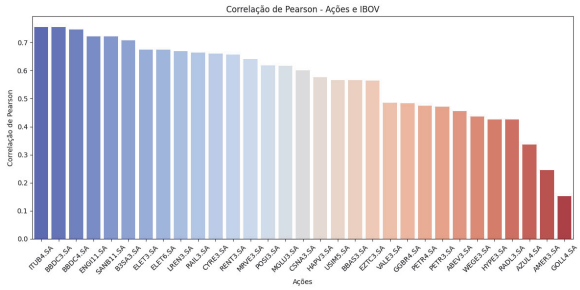


Fonte: elaborado pelo autor (2025)

Na (Figura 30) a correlação tende a se manter mais consistentemente positiva.A correlação se torna mais estável e tende a oscilar entre 0.4 e 0.8, e indica uma relação mais consistente com o IBOV no médio prazo.

Já na (Figura 31), observa-se uma Correlação positiva, alta e estável. Isso mostra que a estrutura do índice IBOV é coesa no longo prazo: as ações, em sua média ponderada, andam juntas com o índice. A ação parece fortemente ligada ao índice no longo prazo.

Figura 32 – Correlação de Pearson - IBOV e Ações (202 pares)



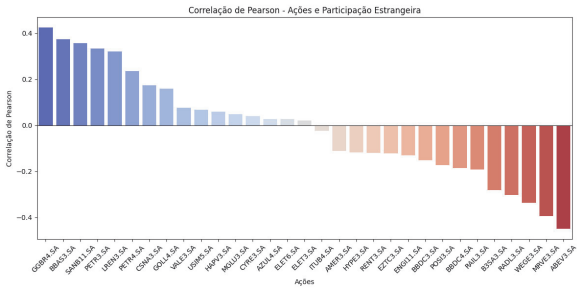
Fonte: elaborado pelo autor (2025)

Na (Figura 32), assim como na (Figura 31), observa-se uma Correlação positiva, alta e estável. Indica que a ação segue uma tendência estrutural fortemente relacionada ao IBOV.

4.3 Fluxo de Investimentos Estrangeiros e Ações

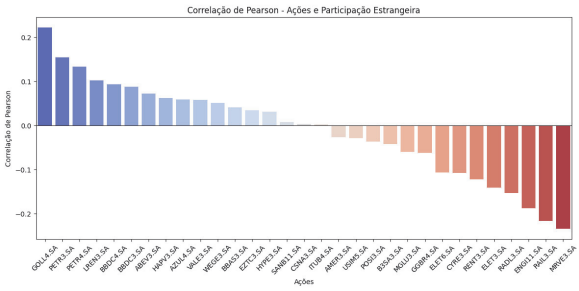
A seguir, são exibidos os gráficos de correlação de Pearson entre o fluxo de investimentos estrangeiros e as ações em cada um das séries temporais.

Figura 33 – Correlação de Pearson - Ações e Participação Estrangeira (22 pares)



Fonte: elaborado pelo autor (2025)

Figura 34 – Correlação de Pearson - Ações e Participação Estrangeira (62 pares)

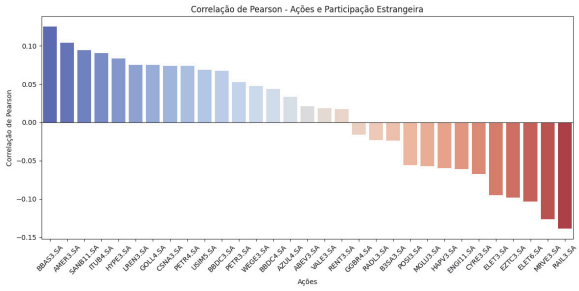


Fonte: elaborado pelo autor (2025)

Na (Figura 33) as correlações variam de +0,40 a -0,45, com algumas ações apresentando correlação positiva moderada e outras com forte correlação negativa. Essa correlação negativa pode indicar saída de capital estrangeiro ou um comportamento oposto ao fluxo de investimento.

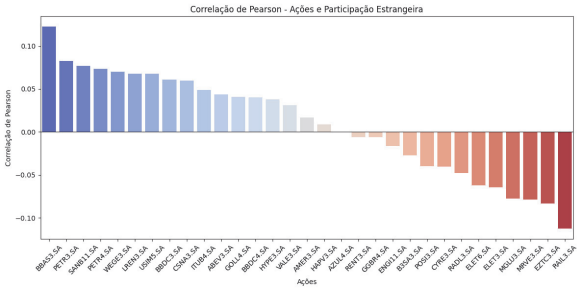
Já na (Figura 34) as correlações variam +0,22 até –0,25.

Figura 35 – Correlação de Pearson - Ações e Participação Estrangeira (101 pares)



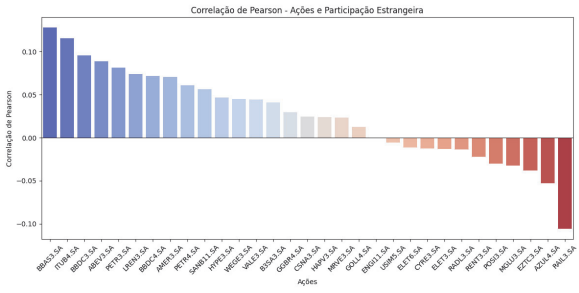
Fonte: elaborado pelo autor (2025)

Figura 36 – Correlação de Pearson - Ações e Participação Estrangeira (138 pares)



Fonte: elaborado pelo autor (2025)

Figura 37 – Correlação de Pearson - Ações e Participação Estrangeira (202 pares)



Fonte: elaborado pelo autor (2025)

Na (Figura 35) as correlações variam de +0,12 até –0,14. E na (Figura 36), as correlações variam de +0,12 até –0,11. Ambas apontam uma correlação próximas de 0, indicando fraca ou nenhuma correlação linear.

Na (Figura 37) as correlações variam de +0,12 até –0,11. E apresentam os mesmos resultados que as amostras com 101 e 138 pares.

Nota-se que a amplitude das correlações diminui à medida que o número de pares aumenta. O maior valor registrado foi de aproximadamente +0,40, considerado uma correlação moderada. Esse valor aparece na primeira série temporal, com janela mais curta. Portanto, os investimentos estrangeiros podem exercer pouca influência em ações individuais de forma direta.

5 CONCLUSÕES

A análise das séries temporais entre o IBOV e os investimentos estrangeiros utilizou técnicas como correlação de Pearson, regressão linear, teste de causalidade de Granger, PCA e K-means para investigar relações e possíveis causalidades.

Os resultados indicam que, à medida que o número de pares de dados aumenta, a força da correlação de Pearson tende a diminuir. Essa queda pode estar associada ao crescimento da variabilidade e do ruído presentes em amostras maiores.

Por outro lado, o teste de Granger demonstrou que, com a ampliação da amostra, surge uma influência significativa do IBOV sobre o capital estrangeiro. À medida que a amostra se expande, essa influência passa a ser de mão dupla, sugerindo uma relação de interdependência que se desenvolve ao longo do tempo.

De maneira geral, a correlação de Pearson identificou uma associação positiva e estatisticamente significativa entre o fluxo de capital estrangeiro e o IBOV, podendo assumir caráter unidirecional ou bidirecional conforme o número de observações disponíveis.

Dessa forma, pode-se sugerir que o IBOV impacta as decisões dos investidores estrangeiros de maneira rápida, enquanto os investidores afetam o IBOV de forma mais prolongada.

Além disso, o K-means e o PCA revelaram padrões mais claros com o aumento dos dados, evidenciando agrupamentos com alta participação estrangeira e IBOV elevado, reforçando a existência de uma relação positiva entre ambos.Em quase todas as visualizações, observa-se um agrupamento de pontos com alta participação estrangeira e valores elevados do IBOV, sugerindo uma relação de influência entre ambos.

Dessa forma, fica evidente que o uso do K-means mostrou-se eficiente na segmentação do comportamento entre o IBOV e a participação estrangeira

Em relação à correlação entre o IBOV e as ações individuais, a análise indicou que, no curto prazo, os papéis podem apresentar movimentos independentes do índice, revelando certa aderência, mas ainda com variações. No entanto, a longo prazo, observa-se uma correlação mais consolidada, evidenciando que a estrutura do IBOV é coesa: em média ponderada, as ações tendem a acompanhar o desempenho do índice.

Por outro lado, a correlação entre o fluxo de capital estrangeiro e as ações apresentou valores mais baixos. No curto prazo, os coeficientes variaram entre +0,40 e -0,45, enquanto, ao longo do tempo, os valores oscilaram entre +0,12 e -0,11. Observa-

se, portanto, que a amplitude dessas correlações diminui à medida que a quantidade de pares aumenta. Isso sugere que os investimentos estrangeiros exercem influência limitada sobre o comportamento de ações individuais.

Esses resultados oferecem uma contribuição relevante para a compreensão do comportamento do mercado acionário brasileiro frente à atuação de investidores estrangeiros, podendo servir de base para o desenvolvimento de estratégias analíticas ou modelos preditivos voltados à tomada de decisão no campo financeiro.

5.1 Trabalhos Futuros

Este trabalho focou na análise de correlação entre a participação de investidores estrangeiros e sua relação com o comportamento do IBOV e de ações específicas no mercado brasileiro. Contudo, existem diversos caminhos possíveis para aprofundamento e continuidade da pesquisa.

Uma primeira possibilidade seria aumentar as janelas de amostragem, tanto em termos temporais quanto no número de ações. Isso permitiria uma avaliação mais robusta e sensível a diferentes ciclos econômicos e movimentos do mercado. Também permitiria entender se os padrões encontrados se mantêm ao longo de diferentes contextos de mercado.

Além disso, sugere-se relacionar os resultados obtidos com eventos econômicos relevantes ocorridos no período analisado. Fatores como crises financeiras, alterações na taxa de juros, decisões de política monetária, podem ajudar a explicar variações nos fluxos de investimento e na dinâmica entre os indicadores.

Adicionalmente, seria interessante analisar os dados por setor de atuação das empresas listadas na B3. Ao dividir as ações por segmentos da economia (como setor bancário, varejo, energia, etc.), é possível observar se o comportamento em relação ao IBOV ou ao fluxo de capital estrangeiro varia de acordo com o setor. Isso pode revelar padrões específicos ou sensibilidades setoriais que passam despercebidas em uma análise mais agregada.

Quanto à metodologia, o presente estudo utilizou a correlação de Pearson e causalidade de Granger como principais medidas estatísticas. Em pesquisas futuras, propõe-se a utilização de outras medidas de correlação, como por exemplo, a correlação de Spearman, que é mais robusta a dados não lineares e à presença de *outliers*, podendo oferecer uma nova perspectiva sobre os dados analisados.

Por fim, há espaço para aplicar outras abordagens estatísticas, como a análise de cointegração, além de técnicas computacionais mais avançadas, como algoritmos de aprendizado de máquina, que podem trazer novas descobertas sobre as relações

entre os participantes do mercado.

REFERÊNCIAS

AHMED, M.; SERAJ, R.; ISLAM, S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 2020. Citado na página 25.

CASTRO, W. S. *Estudo de correlação e causalidade nos movimentos de preços entre o índice Ibovespa e os contratos e mini contratos futuros de Ibovespa e de Dólar*. 2022. Trabalho de Conclusão de Curso, Universidade Estadual do Piauí. Citado na página 14.

CHOLLET, F. *Deep Learning with Python*. 2. ed. [S.l.]: Manning Publications, 2021. Citado na página 21.

EZUGWU, A. E. et al. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Information Sciences*, 2022. Citado 4 vezes nas páginas 23, 24, 25 e 26.

GRANGER, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, The Econometric Society, v. 37, n. 3, p. 424–438, 1969. Citado na página 21.

IMENI, M.; BAO, Z.; NOZICK, V. Multiscale partial correlation analysis of tehran stock market indices: Clustering and inter-index relationships. *Journal of Operational and Strategic Analytics*, 2024. Citado 6 vezes nas páginas 14, 15, 18, 20, 23 e 25.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 374, n. 2065, p. 20150202, 2016. Citado 2 vezes nas páginas 29 e 30.

LEAL, L. V. P. *Modelo de configuração do DeepAR para predição de ações correlacionadas*. 2024. Trabalho de Conclusão de Curso, Universidade Estadual do Piauí. Citado na página 15.

MARTI, G. et al. A review of two decades of correlations, hierarchies, networks and clustering in financial markets. *Physica A: Statistical Mechanics and its Applications*, 2017. Citado 3 vezes nas páginas 14, 23 e 26.

MENG, F. et al. Stock price co-movement prediction based on stock market technique indicators. *Information Sciences*, 2024. Citado na página 14.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to Linear Regression Analysis*. 5. ed. Hoboken, NJ: John Wiley & Sons, 2012. ISBN 978-1-118-60338-9. Citado na página 20.

RAUTIO, T. Comparative analysis of clustering techniques for stock selection in finnish stock markets using common financial metrics. *Theseus*, 2024. Citado 6 vezes nas páginas 21, 22, 25, 26, 27 e 28.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, n. C, p. 53–65, 1987. Citado na página 28.

SHARMA, S. *Applied Multivariate Techniques*. New York: John Wiley & Sons, 1996. ISBN 978-0-471-31064-0. Citado na página 19.

VANHALA, J.; JäRVI, A.; HEIKKONEN, J. Knowledge-based recommender system for stocks using clustering and nearest neighbors. *UTUPub – University of Turku Repository*, 2023. Citado 2 vezes nas páginas 15 e 28.