Research, Society and Development, v. 10, n. 2, e22910212421, 2021
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v10i2.12421

Clusterização de Ativos e suas relações com variáveis macroeconômicas e índices financeiros

Clustering of Assets and its relationship with macroeconomic variables and the financial indexes

Clustering de Activos y su relación con variables macroeconómicas e índices financieros

Recebido: 27/01/2021 | Revisado: 02/02/2021 | Aceito: 05/02/2021 | Publicado: 13/02/2021

Daiane Rodrigues dos Santos
ORCID: https://orcid.org/0000-0001-9215-2260
Universidade Cândido Mendes, Brasil
E-mail: daianesantoseco@gmail.com
Tuany Esthefany Barcellos de Carvalho Silva
ORCID: https://orcid.org/0000-0003-4543-9116
Pontifícia Universidade Católica, Brasil
E-mail: tuanybarcellos@id.uff.br
Campo Elias Suárez Villagrán
ORCID: https://orcid.org/0000-0002-4232-6717
Instituto Nacional de Matemática Pura e Aplicada, Brasil
E-mail: camplise@impa.com
Tiago Costa Ribeiro
ORCID: https://orcid.org/0000-0002-6990-3908
Faculdade IBMEC, Brasil
E-mail: tiagor86@gmail.com

Resumo
O advento do mercado financeiro é um dos acontecimentos mais fascinantes do nosso tempo. Ao longo dos anos, pesquisadores e investidores se interessaram em desenvolver ferramentas para auxiliar na tomada de decisões referentes a alocação de capital. o presente artigo propõe a clusterização como uma métrica para separar um conjunto de ativos, através de um método de agrupamento que maximiza a semelhança entre grupos e minimiza a semelhança entre diferentes grupos com a finalidade de atenuar o risco do portfólio. Adicionalmente, utilizamos regressões lineares múltiplas para evidenciar se os ativos pertencentes aos clusters respondem de forma similar a algumas variáveis macroeconômicas e índices financeiros. para o período analisado - janeiro de 2019 a janeiro de 2020 - obtivemos 8 diferentes clusteres de ativos com um mínimo de 1 ativo (1,32% do total de ativos) e máximo de 30 ativos (42,86% do total de ativos). No que tange as relações com as variáveis selecionadas, o índice de mercado ANBIMA (IMAB) e o índice SMLL (*smalll caps*) são as variáveis que mais se relacionam com os clusteres e as variáveis IPCA e Ibovespa são as que menos apresentaram significância na aplicação econométrica proposta neste artigo.
Palavras-chave*:* Ativos financeiros; Cluster; Decisões financeiras.

Abstract
The advent of the financial market is one of the most fascinating events of our time. Over the years, researchers and investors have been interested in developing tools to assist in making decisions regarding capital allocation. This work proposes clustering as a metric to classify a set of assets, through a grouping method that maximizes similarity between elements of the same group and minimizes similarity between different groups in order to mitigate portfolio risk. Additionally, we use multiple linear regressions (MLR) to show whether the assets belonging to the clusters respond in a similar way to some macroeconomic variables and financial indexes. For the analyzed period -January 2019 to January 2020 - we obtained 8 different clusters of assets with a minimum of 1 asset (1.32 % of total assets) and a maximum of 30 assets (42.86% of total assets). With respect to the relationships with the selected variables, the ANBIMA market index (IMAB) and the smll index (*smalll caps*) are the most closely variables related to the clusters whereas that the IPCA and Ibovespa indexes are the least significant variables in the econometric application proposed in this paper.
Keywords: Financial assets; Cluster; Financial decisions.

Resumen
La llegada del mercado financiero es uno de los acontecimientos más fascinantes de nuestro tiempo. A lo largo de los años, los investigadores e inversores se han interesado en desarrollar herramientas para ayudar a tomar decisiones sobre la asignación de capital. Este artículo propone el agrupamiento como métrica para separar un conjunto de activos, utilizando un método de agrupamiento que maximiza la similitud entre grupos y minimiza la similitud entre diferentes grupos para mitigar el riesgo de cartera. Además, utilizamos regresiones lineales múltiples para mostrar si los activos

que pertenecen a los clústeres responden de manera similar a algunas variables macroeconómicas e índices financieros. para el período analizado - enero de 2019 a enero de 2020 - obtuvimos 8 grupos de activos diferentes con un mínimo de 1 activo (1,32% del activo total) y un máximo de 30 activos (42,86% del activo total). En cuanto a las relaciones con las variables seleccionadas, el índice de mercado ANBIMA (IMAB) y el índice SMLL (smalll caps) son las variables que más se relacionan con los clusters y las variables IPCA e Ibovespa son las menos significativas en la aplicación econométrica propuesta en este artículo.
Palabras clave: Activos financieros; Racimo; Decisiones financieras.

1. Introdução

O advento do mercado financeiro é um dos acontecimentos mais fascinantes do nosso tempo. Ele teve impacto significativo em muitas áreas como negócios, educação, empregos, tecnologia e, portanto, na economia. Ao longo dos anos, pesquisadores e investidores se interessaram em desenvolver ferramentas para auxiliar na tomada de decisões referentes a alocação de capital nas inúmeras possibilidades de ativos ofertadas com a globalização financeira.

Tomar decisões de alocação de recursos financeiros e gerir portifólios de maneira satisfatória são tarefas fatigantes, principalmente devido à aleatoriedade no mercado e vieses frequentemente vistos no comportamento humano relacionado a investimento e tomada de decisões irracionais. A análise de cluster serve como um método para descobrir quais ativos são diferentes entre si, auxiliando no processo de tomada de decisões.

São diversos os fatores que afetam a decisão de alocação de ativos, fatores como objetivos pessoais, nível de tolerância ao risco, horizonte de investimento, desastres raros, fatores transacionais e custos fixos da participação no mercado de ações, são alguns deles. Ademais, as decisões são influenciadas por processos cognitivos. Estudos recentes salientam que os agentes tendem a ser otimistas demais sobre suas perspectivas de vida, e esse otimismo afeta diretamente suas decisões financeiras. O excesso de confiança inclui superestimação e excesso de precisão. Diante do exposto ferramentas quantitativas inerentes a análise de volatilidade (risco) são de extrema importância para a decisão eficiente de alocação dada a aleatoriedade do mercado financeiro.

O risco é mitigado quando o gestor da carteira de investimentos escolhe ativos que tendem a se comportar de maneiras diferentes entre si em reação a um determinado fator. Por exemplo, duas ações do mesmo setor tendem a se mover juntas na presença de fatores que estão associados ao setor como um todo. Por exemplo, espera-se que a inclusão de um imposto, ou incentivo, a uma determinada indústria afete num mesmo sentido as ações incluídas nesse setor. Analogamente, podemos esperar que ações brasileiras se movam juntas entre si e de modo desassociado das ações americanas, na presença de algum fator que afete exclusivamente o mercado brasileiro. A métrica mais comum para se estimar o nível e o sentido das interações entre os ativos é a correlação. Dentro de uma visão de portfólio, as correlações deveriam ser avaliadas uma a uma entre todos os ativos, o que geralmente dificulta uma visualização do comportamento esperado para a carteira como um todo.

O presente artigo propõe a clusterização como uma métrica para separar um conjunto de ativos (ativos que compõem o Ibovespa), através de um método de agrupamento que maximiza a semelhança entre grupos e minimiza a semelhança entre os grupos com a finalidade de atenuar o risco do portfólio. A utilização da métrica supratranscrita permite descobrir combinações de ativos que podem compor um portfólio mais diversificado e com menor risco. A métrica permite ainda, avaliar a robustez das relações entre os clusters e diversos fatores de risco ao longo do tempo.

Após a clusterização dos ativos utilizamos regressões lineares para testar se os ativos pertencentes aos clusters respondem de forma similar as variações de algumas variáveis, variáveis macroeconômicas e índices financeiros. As variáveis testadas foram: O IPCA, o Índice de Mercado ANBIMA (IMAB[1]), a taxa de câmbio (Dólar), o Índice SMLL (*small caps*) e o Ibovespa.

[1] IMA-B: formado por títulos públicos indexados à inflação medida pelo IPCA (Índice Nacional de Preços ao Consumidor Amplo), que são as NTN-Bs (Notas do Tesouro Nacional – Série B ou Tesouro IPCA+ com Juros Semestrais)

2. Investimentos em Renda Variável e Carteiras de Investimento

No mercado financeiro existem diversas classes de ativos de diferentes segmentos para alocação de recursos dos investidores. De acordo com Lanzarini *et al.* (2011), dentre as opções existentes no mercado financeiro, o mercado de capitais, que abriga em si o segmento renda variável, possibilita de aplicação em múltiplas formas de investimento, ou classes de ativos, dentre elas a aplicação em ações negociadas na bolsa de valores.

A diversificação é uma estratégia amplamente utilizada, com a finalidade de reduzir o risco de um portfólio até um nível idiossincrático. Ou seja, o objetivo da diversificação é reduzir a medida de risco utilizada, por exemplo a volatilidade da carteira, através da inclusão de ativos não correlacionados entre si, de forma que a volatilidade da carteira é substancialmente inferior a média ponderada da volatilidade dos ativos.

Elaborar uma carteira devidamente balanceada a ponto de otimizar o retorno esperado por unidade de risco utilizada não é uma tarefa trivial, e pode ser avaliada sob diferentes aspectos através de diferentes alternativas.

A realidade econômica, sem dúvida, confirma a existência da inter-relação entre retorno e risco, esse fato que podemos chamar de fato estilizado, pode ser considerado como a razão para inúmeros estudos acadêmicos e de mercado em relação a diversificação de carteiras de ativos. Este artigo, como supracitado, apresenta uma forma de análise para auxiliar o decisor nas escolhas do seu portifólio de renda variável.

3. Análise de Agrupamentos (clusters)

Análise de agrupamento, ou clusterização, consiste em técnicas computacionais que permitem a separação de objetos em grupos. Essa análise é um procedimento de Estatística Multivariada que objetiva particionar os elementos em dois ou mais clusters considerando a similaridade deles de acordo com critérios pré-estabelecidos. Tais critérios, de acordo com Papenbrock (2011), normalmente são baseados em uma função de dissimilaridade que recebe dois objetos retornando a distância entre eles. Após a implementação de uma métrica de qualidade os grupos devem apresentar uma alta homogeneidade interna e heterogeneidade externa, ou seja, os elementos de um conjunto devem ser mutuamente similares e diferentes dos elementos de outros conjuntos (Linden, 2009).

A clusterização pode ser vista como ferramenta auxiliar, a métrica em questão pode ser utilizada como um conjunto de procedimentos para organizar séries temporais com base em dados de similaridade ou dissimilaridade entre as mesmas. É o encaixe de um espaço de alta dimensão em uma estrutura semelhante a uma árvore, representada em dendrogramas. A dissimilaridade entre objetos é medida por uma matriz de distância cujos componentes assemelham-se à distância entre dois pontos. A técnica supracitada pode ser descrita como um processo de duas etapas: (i) a escolha de uma medida de distância e (ii) a escolha do algoritmo de cluster. Essas duas etapas juntas definem todo o resultado do agrupamento.

As séries temporais de retorno de ativos financeiros concentram-se na dissimilaridade entre as evoluções de tempo síncronas de um conjunto de ativos. A matriz de distâncias entre esses ativos será a entrada do algoritmo hierárquico de cluster que usa uma regra de ligação para determinar uma estrutura hierárquica. Após o índice de proximidade ter sido definido e calculada a matriz de distância, o agrupamento hierárquico pode ser realizado por um algoritmo de agrupamento adequado.

3.1 Medidas de Dissimilaridade

Segundo Linden (2009), similaridade entre os elementos é uma medida empírica de correspondência, ou semelhança, entre os objetos que serão agrupados. Os métodos de agrupamento podem ser descritos por uma matriz contendo uma medida de dissimilaridade ou de proximidade entre cada par de objetos, onde cada entrada $\boldsymbol{p_{ij}}$ na matriz é um valor numérico que

demonstra quão próximos os objetos $\boldsymbol{i}$ e $\boldsymbol{j}$ são. Os coeficientes de dissimilaridade apresentados são funções $\boldsymbol{d}$: $\boldsymbol{\Gamma X \Gamma \Rightarrow \Re}$, onde $\boldsymbol{\Gamma}$ representa o conjunto de objetos de interesse. Estas funções permitem a transformação da matriz de dados,

$$\Gamma = \begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{1l} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \quad (1)$$

Em uma matriz de distâncias,

$$\mathrm{d} = \begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix} \quad (2)$$

Sendo, $\boldsymbol{d(i,j)}$ a distância calculada entre os elementos $\boldsymbol{i}$ e $\boldsymbol{j}$.

As funções de dissimilaridade precisam seguir alguns critérios, sendo estes:

$$\boldsymbol{d(i,j) \geq 0, \forall\, i,j \in \Gamma} \quad (3)$$

$$\boldsymbol{d(i,j) = d(j,i), \forall\, i,j \in \Gamma} \quad (4)$$

$$\boldsymbol{d(i,j) + d(i,k) \geq d(i,k), \forall\, i,j,k \in \Gamma} \quad (5)$$

Após atender as propriedades listadas acima se a métrica também possuir a propriedade $\boldsymbol{d(ax,ay) = |a|d(x,y)}$, então ela é denominada norma. Existem muitas métricas de dissimilaridade, neste trabalho a métrica aplicada foi a distância euclidiana, que é dada pela seguinte equação:

$$\boldsymbol{d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{ip} - x_{jp}|^2)}} \quad (6)$$

3.2 Heurísticas de agrupamento

Para a construção dos clusters existem duas técnicas de agrupamento, conhecidas como método hierárquico que consiste em identificar agrupamentos e o provável número $\boldsymbol{g}$ de grupos, por uma série de fusões sucessivas, ou uma série de sucessivas divisões, tendo seus resultados observados no dendograma, que ilustra as fusões ou divisões feitas em níveis sucessivos. O outro método é conhecido como não-hierárquico, onde o número $\boldsymbol{g}$ de grupos é pré-estabelecido, esta técnica consiste em encontrar diretamente uma partição de $\boldsymbol{n}$ itens em $\boldsymbol{k}$ clusters, por dois requisitos como, semelhança interna e isolamento dos clusters formados. Neste trabalho utiliza-se a técnica não-hierárquica K-*Means.*

3.2.1 Análise Fatorial (principais componentes)

Para determinar um $\boldsymbol{k}$ inicial para aplicação da técnica não-hierárquica k-médias, usou-se a análise fatorial e de principais componentes. A análise fatorial proporciona a descrição da variabilidade de variáveis correlacionadas observadas em um menor número de variáveis não observadas, que são linearmente relacionadas com as variáveis originais. Modela-se as

variáveis observadas como uma combinação linear dos fatores comuns somado a um erro aleatório,

$$\begin{aligned} \boldsymbol{Z_1} &= \boldsymbol{l_{11}\,F_1} + \boldsymbol{l_{12}\,F_2} + \cdots + \boldsymbol{l_{1n}F_n} + \boldsymbol{\varepsilon_1} \\ \boldsymbol{Z_2} &= \boldsymbol{l_{21}\,F_1} + \boldsymbol{l_{22}\,F_2} + \cdots + \boldsymbol{l_{2n}F_n} + \boldsymbol{\varepsilon_2} \\ &\vdots \\ \boldsymbol{Z_p} &= \boldsymbol{l_{p1}\,F_1} + \boldsymbol{l_{p2}\,F_2} + \cdots + \boldsymbol{l_{pn}F_n} + \boldsymbol{\varepsilon_p} \end{aligned} \qquad (7)$$

Então,

$\boldsymbol{Z_i} = \frac{\boldsymbol{X_i - \mu_i}}{\boldsymbol{\sigma_i}}$: é a variável padronizada da equação

$\boldsymbol{X_i}$: variável original com média $\boldsymbol{\mu_i}$ e variância $\boldsymbol{\sigma_i^2}$

$\boldsymbol{\varepsilon_i}$: i-ésimo erro aleatório, sendo $\boldsymbol{i = 1, \ldots, p}$

$\boldsymbol{F_j}$: j-ésimo faotor comum, sendo $\boldsymbol{j = 1, \ldots, n}$

$\boldsymbol{l_{ij}}$: coeficiente da i-ésima variável padronizada $\boldsymbol{Z_i}$ no j-ésimo fator $\boldsymbol{F_j}$

A Análise Fatorial assume a existência de um modelo estatístico que utiliza técnicas de regressão para testar hipóteses e está relacionada com a análise de componentes principais (MINGOTI, 2005).

A análise de componentes principais (ACP) é uma técnica multivariada de modelagem da estrutura de covariância, tendo como principal objetivo conseguir explicar a estrutura de covariância e variância de um vetor aleatório, composto por $\boldsymbol{n}$ variáveis aleatórias, por meio da combinação linear das variáveis originais, chamadas de componentes principais (Hongyu, *et al*, 2008). Como está análise busca explicar a maior parte da variação total existente nas variáveis, é adequado para extrair a maior proporção da variância com o menor número de fatores, ou seja, através desta análise pode-se definir um valor para $\boldsymbol{k}$ podendo usá-lo na aplicação do método não-hierárquico *K-Means.*

3.2.2 Método *K-Means*

O K-Means é uma heurística de agrupamento não hierárquico que tem como objetivo minimizar a distância dos objetos a um conjunto de $\boldsymbol{k}$ centros. A distância entre um ponto $\boldsymbol{p_i}$ e um grupo de clusters é dada por $\boldsymbol{d(p_i, \chi)}$, definida como a distância do ponto ao centro mais próximo dele. A função a ser minimizada então, é dada por:

$$\boldsymbol{d(P, \chi) = \frac{1}{n}\sum_{i=1}^{n} d(p_i, \chi)^2} \qquad (8)$$

O algoritmo depende de um parâmetro $\boldsymbol{k}$ = número de clusters, definido pelo usuário, neste trabalho o parâmetro $\boldsymbol{k}$ foi definido através da análise fatorial e de componentes principais. O método aplicado não-hierárquico possibilita a definição prévia do número de clusters, em cada estágio, novos clusters podem ser formados por divisão ou junção de clusters inicialmente definidos, sem a necessidade das observações de dendogramas, os algoritmos são iterativos e têm uma maior capacidade de análise do conjunto de dados e cada item é alocado para um cluster que tem um centroide mais próximo (média).

4. Resultados

Para as análises e elaboração dos clusters foram coletados, através da base de dados da B3 e ANBIMA, 70 ativos em um período diário de 02 de janeiro de 2019 à 31 de janeiro de 2020. A Tabela 1 apresenta o número total de cluster com seus respectivos ativos, que foram agrupados através da análise fatorial e análise de cluster.

Tabela 1. Ativos que compõem os Clusters.

Cluster	Ativos
1	AZUL4, BPAC11, CYRE3, GOLL4, MRVE3, SMLS3
2	ELET3, ELET6
3	BRML3, CCRO3, CMIG4, HGTX3, COGN3, CVCB3, ECOR3, NTCO3, IGTA3, RENT3, LREN3, MULT3, PETR3, PETR4, QUAL3, SBSP3, UGPA3, YDUQ3
4	BTOW3, LAME4, MGLU3, VVAR3
5	ABEV3, B3SA3, BBSE3, BBDC3, BBDC4, BBAS3, BRKM5, CRFB3, CSAN3, EMBR3, ENBR3, EGIE3, EQTL3, FLRY3, HYPE3, IRBR3, ITSA4, ITUB4, KLBN11, BRDT3, RADL3, RAIL3, SANB11, SULA11, SUZB3, TAEE11, VIVT4, TIMP3, TOTS3, WEGE3
6	BRFS3, JBSS3, MRFG3
7	BRAP4, GOAU4, GGBR4, CSNA3, USIM5, VALE3
8	CIEL3

Fonte: Autores.

A Tabela 2 apresenta as estatísticas descritivas: os valores de mínimo, média, máximo e a variância dentro de cada cluster em um período diário de janeiro de 2019 a janeiro 2020, nota-se que o cluster 5 apresentou uma menor variância comparado aos demais, já o cluster 8 composto somente por um ativo o CIEL3, foi o que mostrou maior variabilidade no período observado, sendo o único a apresentar retorno médio negativo.

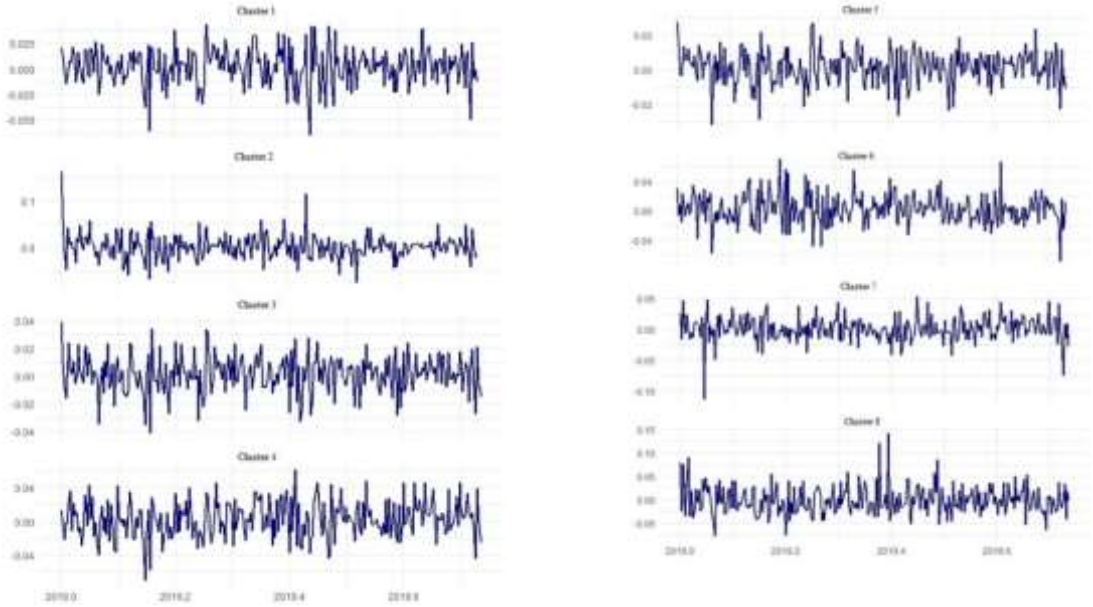
Tabela 2. Estatísticas descritivas dos clusters.

Cluster	Mínimo	Média	Máximo	Variância
1	-0,0651	0,0021	0,0434	0,00031
2	-0,0753	0,0017	0,1619	0,00069
3	-0,0416	0,0014	0,0388	0,00017
4	-0,0699	0,0027	0,0611	0,00046
5	-0,0321	0,0012	0,0274	9.422e-05
6	-0,6957	0,0023	0,0712	0,00041
7	-0,1141	0,0008	0,0525	0,00038
8	-0,0771	-0,0006	0,1426	0,00093

Fonte: Autores com base nos dados retirados da B3.

A Figura 1 apresenta as séries de cada cluster individualmente, assim como na Tabela 2 observa-se os pontos de mínimo e máximo, nota-se que todos os clusters possuem uma série estável em torno da média, no cluster 5 é possível identificar uma menor variabilidade no período analisado, já nos clusters 7 e 8 observa-se os menores retornos médios dentre os demais, o cluster 8 possui a maior variância atingindo seu ponto de máximo em outubro de 2019, e mostra média negativa, o cluster 7 possui a menor média entre os retornos positivos, a maior média é apresentada pelo cluster 4.

Figura 1. Séries contendo a média dos retornos dos ativos por clusters.



Fonte: Autores com base nos dados retirados da B3.

Para analisar o índice de performance diária dos ativos a fim de definir um valor para $\boldsymbol{k}$ que possibilite a utilização da técnita de k-*means* para a análise de agrupamento, aplicou-se a análise fatorial e análise de componentes principais, foram observados 70 ativos. Para verificar se o uso da técnica de análise fatorial é apropriado, aplicou-se o teste de esferacidade Bartlett e Kaiser-Meyer-Olkin (KMO).

O objetivo do teste de esfericidade de Bartlett consiste em verificar se todas as correlações dentro da matriz de correlações são significativas, logo as hipótesese a serem testadas são:

$$\begin{cases} \boldsymbol{H_0}\textbf{: As variâncias dos grupos são iguais.} \\ \boldsymbol{H_1}\textbf{: As variâncias dos grupos são diferentes.} \end{cases}$$

O resultado do teste apresentou $p - valor < 2.2e - 16$, logo, considerando um nível de significância de 5% temos evidências para rejeitar a hipótese nula $\boldsymbol{H_0}$, ou seja, as variâncias dos ativos comparados são diferentes, sendo assim segundo teste de Bartlett o uso da análise fatorial é apropriado.

Aplicou-se o teste KMO (Kaiser-Meyer-Olkin) para avaliar a adequação do tamanho da amostra, o resultado deste teste varia entre 0 e 1, sendo aceitável para análise fatorial resultados acima de 0,5. Neste teste obteve-se KMO = 0,92, logo, a amostra é adequada para análise fatorial.

Feitos os testes, aplicou-se análise fatorial e por meio da diagonalização de matrizes simétricas positivas semi-definidas obteve-se os componentes principais. O primeiro componente principal responde por cerca de 31% da variância total dos dados padronizados, ao passo que se tomarmos os oito primeiros componentes a proporção é cerca de 70% da variância total. Esses fatores representam o número mínimo de causas que condicionam um máximo de variabilidade existente, ou seja, a análise fatorial baseia-se em 8 fatores, logo para a análise de agrupamento e aplicação da técnica k-médias será utilizado $\boldsymbol{K = 8}$.

Os mercados de ações são afetados por muitos fatores altamente interrelacionados que incluem fatores econômicos, políticos, variáveis psicológicas e específicas da empresa. A análise técnica e fundamentalista são as duas principais abordagens para analisar os mercados financeiros. Para investir em ações e alcançar retornos com baixos riscos, os investidores usaram essas duas principais abordagens para tomar decisões nos mercados financeiros.

Para a análise de agrupamento optou-se pelo método não – hierárquico aplicando a técnica de *k-means*, a escolha dessa abordagem se deu pelo tamanho da amostra observada, pois para conjuntos de dados considerados grandes este método apresenta resultados mais significativos, para sua utilização o valor de $\boldsymbol{k}$ precisa ser previamente estabelecido, feito o uso da análise fatorial este valor foi definido por $\boldsymbol{k = 8}$, ou seja, os ativos observados serão agrupados em oito clusters por semelhança interna dos mesmos. Este método de partição permite mensurar a proximidade entre os grupos de ativos, utilizando a distância euclidiana existente entre os centroides destes grupos. Após a realização das análises e implementação dos métodos pré-estabelecidos, pode-se observar na Tabela 3 e na Figura 2 a divisão dos ativos agrupados em cada cluster, nota-se que o cluster 5 (azul celeste) possui cerca de 42,86% dos ativos observados, ou seja, 30 desses ativos possuem características semelhantes e são altamente correlacionados, já o cluster 8 (rosa) apresenta somente um ativo, logo, o mesmo não mostrou similaridade com nenhum outro ativo dentro da amostra no período selecionado. Pode-se observar os clusters com seus respectivos ativos na Figura 4 de acordo com a Tabela 1 supra apresentada.

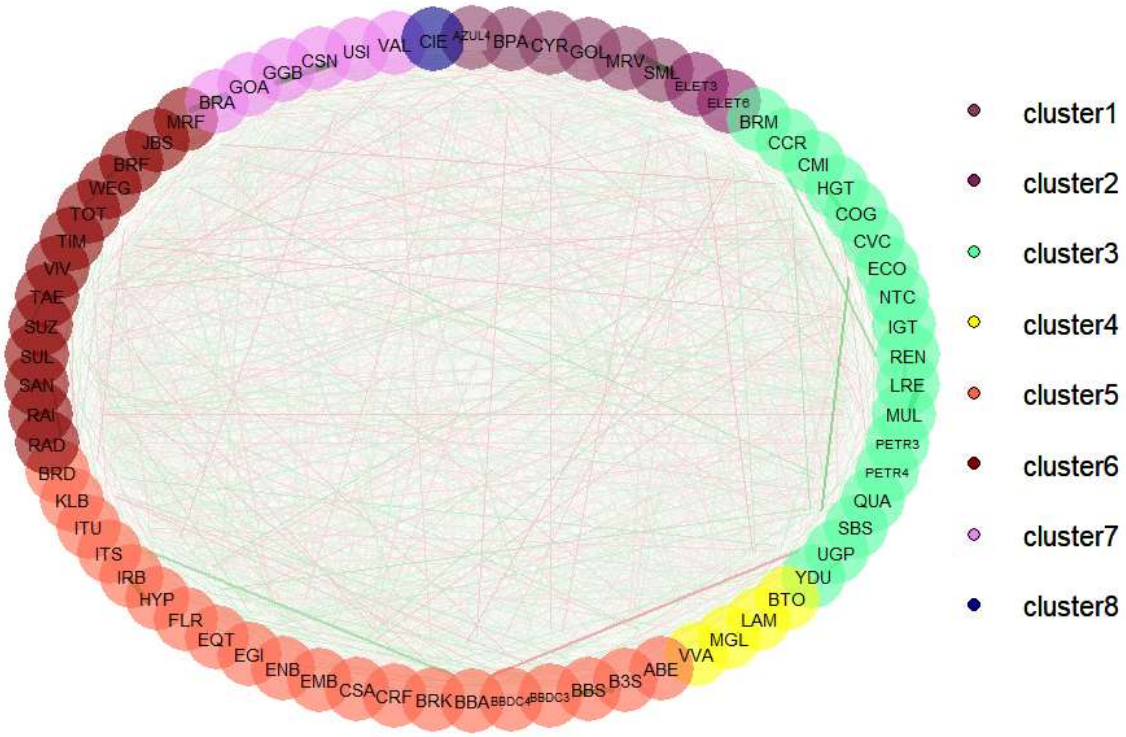
Tabela 3. Quantidade de ativos que compõem os Clusters.

Cluster	Quantidade de ativos	Porcentagem de participação
1	6	8,6%
2	2	2,9%
3	18	25,72%
4	4	5,7%
5	30	42,86%
6	3	4,3%
7	6	8,6%
8	1	1,32%

Fonte: Autores com base nos dados retirados da B3.

Os ativos são conectados, na Figura 2, por uma reta ou verde, ou vermelha. As retas verdes correspondem a correlações positivas e as vermelhas correlações negativas. Vale ressaltar que quando mais espessa (escura) a reta for mais forte é a correlação entre os ativos. Como pode ser visto, as ações da Azul Linhas Aéreas Brasileiras e da Gol Linhas Aéreas apresentam correlação forte e positiva, similarmente os ativos LAME4 e BTOW3, ECOR3 e CCRO3 também apresentam o mesmo padrão de correlação, acentuada e positiva.

Figura 2. clusters contendo os ativos.



Fonte: Autores com base nos dados retirados da B3.

Após a clusterização dos ativos selecionados (Tabela 1), utilizamos regressões para testar se os ativos pertencentes aos clusters respondem de forma similar a algumas variáveis macroeconômicas e índices financeiros (O IPCA, o Índice de Mercado ANBIMA - IMAB, A taxa de câmbio - Dólar o Índice SMLL – *small caps* e o Ibovespa. O resultado é apresentado nas tabelas a seguir)

Tabela 4. Cluster 1: avaliação da dependência dos retornos dos ativos que compõem o cluster 1 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 1					
P-valores referentes a significância dos betas da regressão					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
AZUL4	-	1,3E-07	1,2E-04	2,2E-16	-
BPAC11	-	3,6E-03	6,1E-03	3,8E-09	-
CYRE3	-	2,2E-16	-	2,2E-16	-
GOLL4	3,7E-02	2,3E-10	1,9E-04	1,9E-04	-
MRVE3	-	9,2E-07	-	2,2E-16	-
Percentual para a aderência das variáveis macroeconômicas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	1	6	3	6	0
Percentual	17%	100%	50%	100%	0%

Fonte: Autores com base nos dados retirados da B3.

Como pode ser observado na Tabela 4, todos os 6 ativos que compõem o cluster 1 apresentaram p-valor inferior a 0,05 (nível de significância de 5%), ou seja, rejeitamos a hipótese de que o beta é igual a zero para as variáveis explicativas IMAB e SMLL. Três das seis variáveis mostraram significância para o Dólar e apenas 1 para o IPCA. Este resultado nos mostra que a variação dos retornos dos ativos que compõem o cluster podem ser escritos como uma função linear das variáveis supracitas. Ressalta-se que uma modelagem mais minuciosa pode ser feita para a verificação do grau de impacto das variáveis eocnomicos selecionadas. Cabe destacar que as regressões aqui apretentadas atenteram tiveram bons níveis de aderênica (R^2 superiro a 0.65, Estatistica F inferior a 0,05 e resídos normalizados).

Na Tabela 5 é possível observar que todos os ativos da empresa Centrais Eletricas Brasileiras SA (ELET3 e ELET6) que compõem o cluster 2, apresentaram p-valor inferior a 0,05, ou seja, rejeitamos a hipótese de que o beta é igual a zero para as variáveis explicativas IMAB e SMLL. Ou seja, a variação dos retornos dos ativos ELET3 e ELET6 podem ser escritos como uma função linear das variações do ínice de títulos públicos indexados à inflação medida pelo IPCA e do índice Small Cap (SMLL) que reflete os ativos das empresas de menor capitalização na B3.

Tabela 5. Cluster 2: avaliação da dependência dos retornos dos ativos que compõem o cluster 2 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 2					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
ELET3	-	9,39E-03	-	1,19E-02	4,23E-02
ELET6	-	4,91E-03	-	1,30E-02	3,74E-02
Percentual para a aderência das variáveis macroeconômicas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	0	2	0	2	0
Percentual	0%	100%	0%	100%	0%

Fonte: Autores com base nos dados retirados da B3.

De acordo com os resultados apresentados na Tabela 6, todos os ativos que compõem o cluster 3 apresentaram p-valor inferior a 0,05, ou seja, rejeitamos a hipótese de que o beta é igual a zero para a variável explicativa SMLL, ao passo que apensas, 17%, 22% e 78% das variações das cotações das ações que compõem o cluster respondem as variáveis Dolar, IPCA e IMAB respectivamente.

Tabela 6. Cluster 3: avaliação da dependência dos retornos dos ativos que compõem o cluster 3 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 3					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
BRML3	-	3,47E-12	-	2,20E-16	-
CCRO3	-	1,32E-09	-	2,43E-03	-
CMIG4	-	3,74E-08	-	2,15E-04	-
HGTX3	-	1,63E-09	-	2,20E-16	-
COGN3	-	4,37E-10	-	2,20E-16	-
CVCB3	-	1,42E-09	-	2,20E-16	-
ECOR3	-		-	1.41e-11	-
NTCO3	-		-	1.61e-09	-
IGTA3	-	2,00E-16	-	2,00E-16	-
RENT3	-	3,37E-10	-	2,20E-16	-
LREN3	9,99E-03	4,19E-12	-	2,20E-16	-
MULT3	1,92E-02	2,20E-16	-	2,20E-16	-
PETR3	-	2,00E-16	1,24E-02	2-E16	2-E16
PETR4	1,49E-02	2,00E-16	1,04E-03	2,00E-16	2,00E-16
QUAL3	-	-	9,82E-03	1,92E-14	-
SBSP3	5,03E-02	8,03E-07	-	2,20E-16	3,85E-02
UGPA3	-	-	-	2,20E-16	-
YDUQ3	-	1,56E-06	-	2,20E-16	-
Percentual para a aderência das variáveis macroeconômicas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	4	14	3	18	0
Percentual	22%	78%	17%	100%	0%

Fonte: Autores com base nos dados retirados da B3.

Como pode ser observado na Tabela 7, todos os ativos que compõem o cluster 4 apresentaram p-valor inferior a 0,05, ou seja, rejeitamos a hipótese de que o beta é igual a zero para o IMAB. Enquanto 75%, 25% e 25% das variações das cotações das ações que compõem o cluster respondem as variáveis IPCA, Ibovespa e SMLL, respectivamente.

Tabela 7. Cluster 4: avaliação da dependência dos retornos dos ativos que compõem o cluster 4 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 4					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
BTOW3	-	1,08E-08	-	2,30E-16	4,00E-02
LAME4	-	1,85E-02	-	-	-
MGLU3	4,03E-02	1,68E-13	-	2,20E-16	-
VVAR3	-	4,94E-08	-	2,20E-16	-
Percentual para a aderência das variáveis macroeconômicas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	1	4	0	3	1
Percentual	25%	100%	0%	75%	25%

Fonte: Autores com base nos dados retirados da B3.

Com relação aos clusteres 5 e 6, podemos observar nas Tabelas 8 e 9 que a variabilidade da única variável econômica que influenciou a variabilidade no retorno de todos os ativos do cluster foi a SMLL, ou seja, a performance econômica dos clusteres é bastante sensível a esta variável.

Tabela 8. Cluster 5: avaliação da dependência dos retornos dos ativos que compõem o cluster 5 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 5					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
ABEV3	-	1,13E-03		3,48E-10	6,62E-05
B3SA3	-	2,00E-16		2,00E-16	2,00E+16
BBSE3	-	2,66E-06		5,92E-14	2,37E-04
BBDC3	-	2,00E-16	9,17E-03	2,00E-16	2,00E-16
BBDC4	-	2,00E-16	2,34E-03	2,00E-16	2,00E-16
BBAS3	-	2,00E-16	3,96E-02	2,00E-16	2,00E-16
BRKM5	-	2,00E-16	-	2,00E-16	2,00E-16
CRFB3	-	2,10E-09	-	2,20E-16	4,45E-03
CSAN3	1,82E-02	2,00E-11	-	2,20E-16	-
EMBR3	-	-	-	6,71E-10	-
ENBR3	5,79E-03	1,13E-08	-	2,50E-16	-
EGIE3	-	2,70E-12	-	2,20E-16	-
EQTL3	4,47E-02	1,24E-15	-	2,20E-16	-
FLRY3	-	-	-	2,20E-16	-
HYPE3	-	5,90E-04	-	-	-
IRBR3	-	1,39E-03	-	8,75E-05	8,23E-04
ITSA4	-	2,00E-16	2,20E-16	2,00E-16	2,00E-16
ITUB4	-	2,00E-16	9,28E-02	2,00E-16	2,00E-16
KLBN11	-	1,20E-02	-	7,94E-10	2,38E-05
BRDT3	-	5,31E-08	-	4,51E-14	-
RADL3	-	6,55E-05	-	3,93E-06	-

RAIL3	3,69E-04	9,38E-08	-	4,46E-14	-
SANB11	-	2,00E-16	-	2,00E-16	2,00E-16
SULA11	-	-	-	1,10E-09	-
SUZB3	-	-	-	3,10E-02	2,72E-03
TAEE11	-	8,62E-13	-	2,20E-16	-
VIVT4	-	2,26E-08	-	2,65E-13	6,31E-04
TIMP3	-	1,46E-07	-	1,01E-08	1,65E-05
TOTS3	-	4,07E-02	-	6,61E-10	4,35E-02
WEGE3	-	4,07E-02	-	6,61E-10	4,35E-02
Percentual para a aderência das variáveis macroeconômicas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	4	26	5	30	18
Percentual	13%	87%	17%	100%	60%

Fonte: Autores com base nos dados retirados da B3.

Tabela 9. Cluster 6: avaliação da dependência dos retornos dos ativos que compõem o cluster 6 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 6					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
BRFS3	-	-	-	9,53E-05	-
JBSS3	-	-	-	5,65E-05	-
MRFG3	-	-	-	4,12E-05	-
Percentual para a aderência das variáveis selecionadas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	0	0	0	3	0
Percentual	0%	0%	0%	100%	0%

Fonte: Autores com base nos dados retirados da B3.

Na Tabela 10, observa-se que todos os 6 ativos que compõem o cluster 7 apresentaram p-valor inferior a 0,05 (nível de significância de 5%), ou seja, rejeitamos a hipótese de que o beta é igual a zero para as variáveis explicativas Ibovespa e SMLL. No Cluster 8, formado apenas pelo ativo Cielo SA, CIEL3, as variações do IMAB, Ibovespa e SMLL podem ser consideradas como variáveis dependentes.

Tabela 10. Cluster 7: avaliação da dependência dos retornos dos ativos que compõem o cluster 7 a variáveis econômicas e índices financeiros (variáveis explicativas).

Cluster 7					
P-valores referentes as significâncias dos betas da regressão múltipla					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
BRAP4	2,25E-02	1,21E-03	-	5,02E-10	1,69E-08
GOAU4	8,85E-03	8,87E-08	-	2,20E-16	2,46E-10
GGBR4	4,92E-02	1,20E-05	-	2,20E-16	1,78E-10
CSNA3	-		-	8,94E-08	2,01E-08
USIM5	1,32E-03	3,73E-06	-	2,20E-16	1,33E-07
VALE3	3,25E-02	2,09E-03	-	1,21E-08	2,02E-11
Percentual para a aderência das variáveis selecionadas					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Número de ativos	5	5	0	6	6
Percentual	83%	83%	0%	100%	100%

Fonte: Autores com base nos dados retirados da B3.

5. Considerações Finais

O presente trabalho teve como objetivo apresentar o método de clusterização para auxiliar investidores na tomada de decisão. A metodologia implementada pode ser usada para realizar uma escolha mais assertiva de portfólios de investimento, pois agrupando os ativos conforme suas similaridades é possível analisar de forma mais clara seus retornos, dessa forma pode-se gerar mais informações para a tomada de decisão do gestor.

Para as análises e elaboração desta pesquisa foram selecionados 70 ativos em um período diário de 02 de janeiro de 2019 à 31 de janeiro de 2020. No presente artigo aplicou-se a metodologia de análise de agrupamento utilizando a técnica não hierárquica de *K-means*, com isto obteve-se 8 clusters com ativos agrupados segundo a semelhança de seus retornos diários. O cluster 4 é composto por 4 ativos e juntos eles possuem o maior retorno médio, o cluster 5 é o que contém a maior quantidade de ativos e apresenta a menor variabilidade comparado aos demais, o cluster 8 possui apenas o ativo CIEL3, mostrando alta variabilidade no período analisado, sendo o único a apresentar retorno médio negativo. Com este estudo foi possivel evidenciar a eficacia do método de agrupamento como uma ferramento para o auxilio na tomada de decisão e na elaboração de novos portfólios, podendo diversifica-los de acordo com as características de cada investidor.

No que tange as relações com as variáveis selecionadas, o índice de mercado AMBIMA (IMAB) e o índice SMLL (small caps) são as variáveis que mais se relacionam com os clusteres e as variáveis IPCA e Ibovespa são as que menos apresentaram significância na aplicação econométrica proposta neste artigo. Esta informação pode auxiliar nas decisões financeiras e mitigar riscos, visto que os investidores podem escolher ativos e/ou clusteres que estejam relacionados com variáveis distintas. Objetiva-se para trabalhos futuros, continuar a análise de agrupamento utilizando outros métodos, iniciando pelo método de análise fatorial confirmatória. Objetiva-se para trabalhos futuros, continuar a análise de agrupamento utilizando outros métodos, iniciando pelo método de análise fatorial confirmatória.

Referências

Anderson, T. An introduction to multivariate statistical analysis. John Wiley & Sons, 675.1984.

Bussab, W., Miazaki, E., & Andrade, D. Introdução à análise de agrupamentos. *Associação Brasileira de Estatística*, p. 105. 1990.

Doni, M. Análise de cluster: métodos hierárquicos e de particionamento. Universidade Presbiteriana Mackenzie. 2004.

Halkini, M., Batistakis, Y., & Vazirgiannis, M. On Clustering Validation Techniques, 2001.

Hongyu, K., Sandanielo, V., & Martins, G. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. *E&S - Engineering and Science*, p. 1-5. 2015.

Lanzarini, J., Queiroz, F., Queiroz, J., Vasconcellos, N., & Hekis, R. A popularização do mercado de ações brasileiro: as mudanças recentes na bolsa de valores. XXXI *Encontro Nacional de Engenharia de Produção*. 2011.

Linden, R. Técnicas de Agrupamento. *Revista de Sistemas de Informação da FSMA*, 18-36(4).

Mingoti, S. Análise de Dados Através de Métodos de Estatística Multivariada: uma abordagem aplicada. Belo Horizonte: *Editora UFMG*. 2005.

Nievola, J. Análise de Agrupamento. *PPGIa*, PUCPR. 2006.

Palma, L. Agrupamento de dados: k- médias. *Universidade Federal do Recôncavo da Bahia Centro de Ciências Exatas e Tecnológicas*.2018.

Papenbrock, J. Asset Clusters and Asset Networks in Financial Risk Management and Portfolio Optimization**.** *Tese de doutorado em economia da Faculdade de Economia do Instituto de Tecnologia Karlsruhe*. 2011.

Quintal, G. Análise de clusters aplicada ao Sucesso/Insucesso em Matemática. *Universidade da Madeira Departamento de Matemática e Engenharias*.2006.
Reis, E. Estatística multivariada aplicada. Lisboa: *Edições Silabo*, p. 342. 1997.

Rocha, T., Peres, S. M., Biscaro, H., Madeo, R., &Boscarioli, C. Tutorial sobre Fuzzuc-Means e Fuzzy Learning Vector Quantization: Abordagens Híbridas para Tarefas de Agrupamentos e Classificação. *Revista de Informática Teórica e Aplicada*, 9(1).

Sarajane M., & Clodoaldo A. Técnicas de Agrupamento (Clustering). 2015.

Silva, T. Método Estatístico de Análise de Cluster Aplicado aos dados de uma Associação de Proteção Veicular. Universidade Federal de Minas Gerais Especialização em Estatística – Ênfase em Mercado e Indústria. 2013.

Totti, R., Vencovsky, R., & Batista, L. Utilização de métodos de agrupamentos hierárquicos em acessos de Paspalum (Graminea (Poaceae)). 2001**.**

Zaiane, O., Oliveira, S. Geometric data transformation for privacy preserving Clustering. Edmonton, Alberta, Canada, 2003.