Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Fernando Biscaia Veiga

Algoritmos de Agrupamento Aplicado nas Empresas Listadas na B3

Curitiba
2022

Fernando Biscaia Veiga

Algoritmos de Agrupamento Aplicado nas Empresas Listadas na B3

Monografia apresentada ao Programa de Especialização em Data Science e Big Data daSetor de Ciências Exatas, Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Fernando de Pol Mayer

Curitiba

2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Algoritmos de Agrupamento Aplicado nas Empresas Listadas na B3

Fernando Biscaia Veiga[1]
Prof. Dr. Fernando de Pol Mayer[2]

Resumo

A utilização da bolsa de valores como ferramenta previdenciária se mostra muito útil considerando os riscos da previdência social em decorrência do aumento da proporção da população idosa em nosso país.
Para isso, o desenvolvimento didático de critérios para maior assertividade quanto as ações a serem investidas se faz necessário.
Primeiramente através do entendimento dos indicadores fundamentalistas com o intuito de avaliar a situação e as projeções das empresas. Além disso, avaliar a correlação entre as empresas onde se pretende aplicar os recursos, visando a diversificação da carteira de investimentos de forma a amenizar os riscos.
Propõe-se a utilização de algoritmos de agrupamento para avaliar a similaridade das mesmas.
Ao final, unir os estudos de correlação ao estudo dos indicadores fundamentalistas para geração de insumos para tomada de decisão sobre onde alocar os recursos.

Palavras-chave: Previdência social, Indicadores fundamentalistas, Algoritmos de agrupamento.

Abstract

The use of the stock exchange as a social security tool is very useful considering the risks of social security due to the increase in the proportion of the elderly population in our country.
For this, the didactic development of criteria for greater assertiveness regarding the actions to be invested is necessary.
First, through the understanding of fundamentalists indicators in order to assess the situation and projections of companies. In addition, to evaluate the correlation between the companies where the resources are intended to be invested, aiming at the diversification of the investment portfolio in order to mitigate the risks.
It is proposed to use clustering algorithms to assess their similarity.
In the end, unite correlation studies with the study of fundamentalist indicators to generate inputs for decision making on where to allocate resources.

Keywrods: Social Security, Fundamentalists indicators, Clustering algorithms.

1 Introdução

A forma como atualmente funciona a Previdência Social Brasileira se mostra não sustentável a longo prazo e põe em risco a aposentadoria das atuais e futuras gerações. Isto porque, no formato atual a população economicamente ativa sustenta os já aposentados.
Ao início deste formato, a pirâmide etária da população possuía uma baixa proporção de pessoas idosas. A maior parcela da população se mantinha ativa sustentando uma pequena fatia da mesma, de pessoas já aposentadas.
Porém, as projeções já para as próximas décadas indicam a inversão da pirâmide etária. Cada vez mais, a alta proporção de idosos se tornará um desafio para a Previdência Social.
Isto fez com que a busca por maneiras de garantir a aposentadoria sem a necessidade da Previdência Social popularizasse os investimentos na bolsa de valores Brasileira, tendo aumentado consideravelmente o número de investidores nos últimos anos.
Contudo, para que a aplicação de recursos na bolsa de valores seja eficaz enquanto ferramenta previdenciária é necessário que as aplicações sejam baseadas no estudo das empresas listadas em bolsa. Primeiramente através do estudo individual das empresas as quais pretende-se investir, afim de verificar a qualidade e sustentabilidade das mesmas. Para isso, o estudo dos indicadores fundamentalistas da empresa são essenciais para que o investidor possa fazer um retrato da empresa e decidir se vale a pena o investimento no ativo.
Outro ponto a ser considerado são as oscilações naturais no mercado financeiro, tornando prudente a diversificação dos investimentos como forma de amenizar o risco. E neste ponto, cabe a avaliação da correlação entre as empresas. Investir em empresas com comportamento similares traria pouco benefício em relação ao risco, visto que momentos de crise em uma

[1]Fernando Biscaia Veiga, fernandobiscaiav@gmail.com.
[3] Prof. Dr. Wagner Hugo Bonat - DEST/UFPR.

empresa poderia refletir nas demais da carteira.
Neste sentido, a aplicação de um algoritmo de agrupamento aplicado a variação das cotações históricas das empresas listadas na B3, como forma de avaliar a correlação das mesmas, avaliado conjuntamente aos indicadores fundamentalistas, pode ser utilizado como insumo para tomada de decisão quanto a quais empresas investir.

2 Materiais e Métodos

Uma carteira de investimentos deve contemplar empresas com bons fundamentos não correlacionadas.
Os indicadores fundamentalistas, que são métricas financeiras calculadas a partir de documentos como o Balanço Patrimonial e Demonstrativo do Resultado do Exercício, das empresas listadas na bolsa de valores brasileira são disponibilizadas em diferentes páginas, podendo ser facilmente obtidas e utilizadas para verificar a situação financeira de uma empresa, assim como, suas perspectivas para médio e longo prazo.
Já a aplicação de algoritmo de agrupamento para as cotações históricas destas mesmas empresas, permite a criação de clusters entre as empresas mais correlacionadas.
A junção destas duas análises podem ser utilizadas como insumo para montagem de uma carteira de investimentos segura e bem diversificada, visto a possibilidade de ranquear as empresas com os melhores fundamentos e, entre estas, selecionar empresas de diferentes clusters.
Todas as análises deste estudo foram realizadas com o software R. [1]

2.1 Base de dados

Os indicadores fundamentalistas foram obtidos através do site "fundamentus" via Web Scraping. [2]
A base de dados utilizada para a aplicação do algoritmo de agrupamento refere-se as cotações históricas dos tickers das empresas listadas na bolsa de valores brasileira, extraída no R através do pacote "BatchGetSymbols". [3]
A base contém os valores diários de abertura, fechamento, maior e menor valor de cada ticker.
Para a aplicação do algoritmo foram mantidas as variáveis descritas na Tabela 1.
O período utilizado para o estudo contempla de fevereiro de 2021 a fevereiro de 2022. A opção inicial era a de avaliar o histórico de cinco anos, porém alguns fatores influenciaram na redução do período.
Primeiramente, o início da pandemia afetou demais a bolsa de valores, fazendo com que todas as empresas apresentassem um comportamento muito atípico.
Além disso, várias empresas listadas possuem o IPO (oferta pública inicial) recente, dessa forma, não havendo histórico anterior.
Cabe ressaltar que várias empresas da bolsa possuem mais de um ticker de negociação, porém separadas entre ordinárias, preferenciais e units. Para estes casos, foram feitas as exclusões de tickers da análise, mantendo apenas um por empresa.

Tabela 1: Dicionário do conjunto de dados de cotações.

Variável	Descrição
ref.date	Data de referência
ticker	Código de identificação da empresa
price.close	Valor de fechamento da cotação

A Tabela 2 apresenta a descrição da base de indicadores fundamentalistas.

Tabela 2: Dicionário da base de dados de indicadores.

Variável	Descrição
P_L	Preço/Lucro
P_VP	Preço/Valor patrimonial
P_EBIT	Preço /EBIT
PSR	Preço/Receita líquida
P_Ativos	Preço/Ativos totais
P_Cap_Giro	Preço/Capital de giro
P_Ativo_Circ_Liq	Preço/Ativo circulante líquido
Div_Yield	Dividend Yield
EV_EBTIDA	Valor da empresa/EBTIDA
EV_EBIT	Valor da empresa/EBIT
Cres_Rec_5anos	Crescimento receita 5 anos
LPA	Lucro por ação
VPA	Valor patrimonial por ação
Margem_Bruta	Lucro bruto/Receita líquida
Marg_Ebit	EBIT/Receita líquida
Marg_Liquida	Lucro líquido/Receita líquida
Ebit_Ativo	EBIT/Ativos totais
ROIC	Retorno sobre o capital investido
ROE	Retorno sobre patrimônio líquido
Liquidez_Corrente	Ativo circulante/passivo circ.
Div_Bruta_Patrimonio	Dívida bruta total
Giro_Ativos	Receita líquida/Ativos totais

2.2 Limpeza e preparo dos dados

Com relação a base de indicadores fundamentalistas, nenhum tratamento se fez necessário após sua extração.
Essa base trata-se de uma fotografia da empresa no momento da extração e deve ser atualizada continuamente para avaliação da situação financeira das empresas.
Sobre a base de cotações históricas, alguns tratamentos se fizeram necessário para o estudo. A aplicação do algoritmo foi realizada na variação diária da cotação

em relação ao dia anterior. Portanto, a base considera o percentual de variação do preço de fechamento em relação preço de fechamento do dia anterior.
Uma das razões em se optar por trabalhar com a variação percentual das cotações foi para que o algoritmo agrupasse as empresas com relação a similaridade das variações diárias, ao invés de agrupar através da faixa de preço de negociação.
Um exemplo disso pode ser observado na Figura 1, onde duas empresas do ramo de transmissão de energia elétrica, conhecidamente correlacionadas, seriam agrupadas em diferentes clusters por serem negociadas em diferentes faixas de preço.

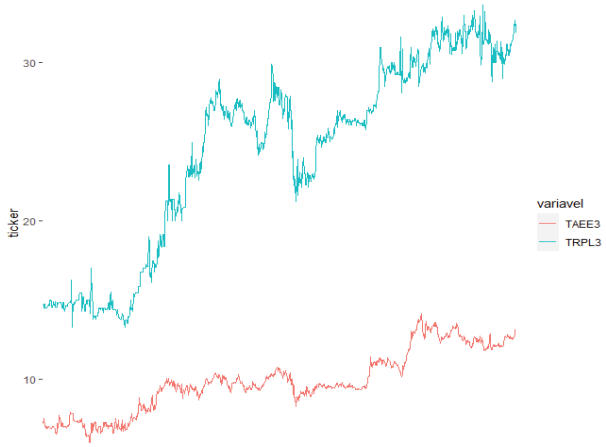

Figura 1: Gráfico dos valores de fechamento das cotações históricas dos tickers TAEE3 e TRPL3.

O agrupamento por faixa de preço não faria sentido, visto que as empresas listadas na B3 são negociadas nas mais diversas faixas de preço, não sendo isto um indicativo de qualidade.
Ao considerar a variação dos valores de cotações, casos como o apresentado na Figura 1 tendem a serem alocados no mesmo agrupamento, visto que as variações diárias se acompanham. A Figura 2 apresenta a série apresentada na Figura 1 considerando as variações percentuais diárias.

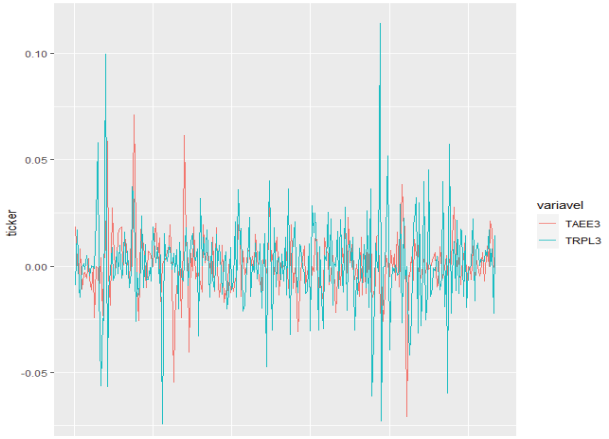

Figura 2: Gráfico da variação diária das cotações históricas dos tickers TAEE3 e TRPL3.

Outra razão para o uso da variação diária das cotações foi para tirar a dependência temporal das séries. Para verificar a não dependência das mesmas, avaliou-se a autocorrelação das séries históricas das quatro empresas de maior peso na composição do índice Bovespa, conforme Figura 3.

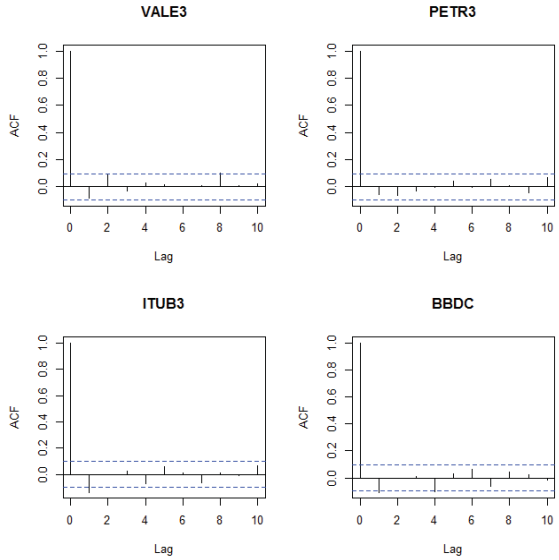

Figura 3: Gráfico de autocorrelação das quatro empresas de maior peso na composição do índice Bovespa.

2.3 Modelos empregados

Após os tratamentos prévios da base de dados e a verificação da não correlação temporal dos principais tickers avaliados, pôde-se iniciar as análises para criação de clusters.
Para a criação desses clusters, aplicou-se algoritmos de agrupamento, que objetivam identificar padrões existentes em conjuntos de dados.
Neste caso, a intenção é verificar similaridades entre as variações nas cotações das ações listadas na bolsa de valores.
Para tanto, foram testados dois algoritmos: agrupamento hierárquico e k-means.

2.3.1 Agrupamento hierárquico

Os algoritmos de agrupamento hierárquico formam-se a partir de uma matriz de similaridade. [4]
Estes algoritmos diferem-se nos métodos para determinar a distância entre os agrupamentos e, no modo como essa distância é estimada.
Para este caso, foi utilizado o método de Ward com distância euclideana.
No método de Ward os grupos são formados pela maximização da homogeneidade dentro dos grupos.

A distância euclideana se ajusta bem ao método de Ward e é a menor distância entre dois pontos no R^n, definido por:

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2}$$

Onde:

- x_{ik} = variação da cotação da ação i
- x_{jk} = variação da cotação da ação j

2.3.2 K-means

O algoritmo k-means gera agrupamentos baseado em análises e comparações dos valores numéricos dos dados. [4]
Utiliza o valor médio dos dados para fixar k centróides, de maneira aleatória, um para cada cluster e, associa cada indivíduo ao centróide mais próximo.
A quantidade de centróides é definido após a verificação do número ótimo de clusters.

3 Resultados e Discussões

Para ambos os algoritmos, o procedimento seguido foi semelhante. Após o tratamento da base de dados foi verificado o número ótimo de clusters através do método silhouette, aplicado o algoritmo e verificado os agrupamentos.
O método silhouette serve como subsídio para a determinação do número ideal de clusters, calculando a função de custo, soma dos quadrados das distâncias internas dos clusters, onde determina-se como número ótimo no momento em que a adição de um novo cluster pouco altera a função de custo.

3.1 Resultado agrupamento hierárquico

Para a aplicação deste algoritmo, o método silhouette retornou como sendo 10 o número ótimo de clusters, conforme Figura 4.

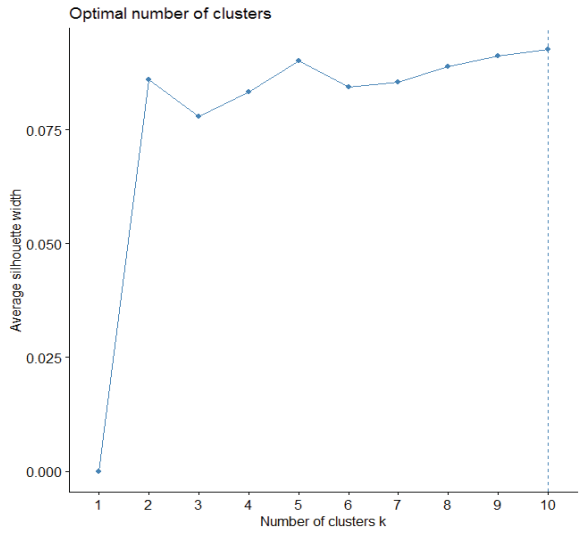

Figura 4: Número ótimo de clusters Agrupamento Hierárquico.

Porém, pode-se observar que apesar de indicar a utilização de 10 clusters, a diferença a partir de 5 agrupamentos não aparenta ser tão siginificativa. Para tanto foram aplicados o algoritmo para os dois números de agrupamentos
As divisões dos agrupamentos com 10 clusters, pode ser observada no dendograma apresentado na Figura 5. Um ponto a ser observado é que cinco entre os dez agrupamentos aparentam a captura de outliers, visto conterem um ou dois tickers.
Nestes casos, em algum momento da série histórica avaliada algum fator influenciou significativamente no valor da cotação, seja uma super valorização ou desvalorização repentina.
Por serem movimentos muito atípicos no comportamento das negociações dos tickers na bolsa de valores o algoritmo os manteve em agrupamentos isolados.

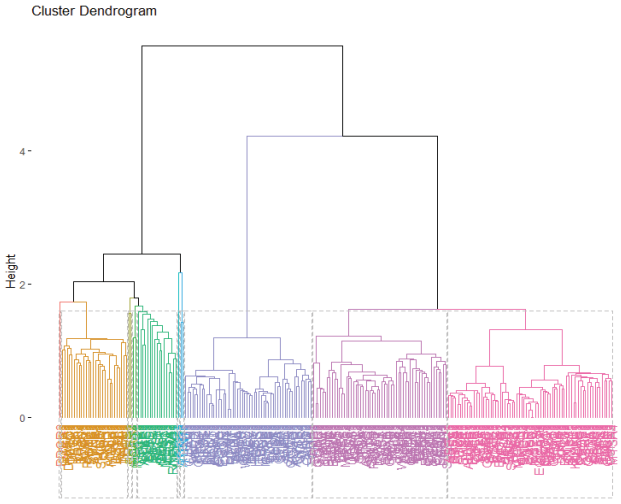

Figura 5: Dendograma Agrupamento Hierárquico 10 clusters.

Este ponto fica mais evidente ao avaliar o agrupamento na projeção dos componentes principais, apresentado na Figura 6.

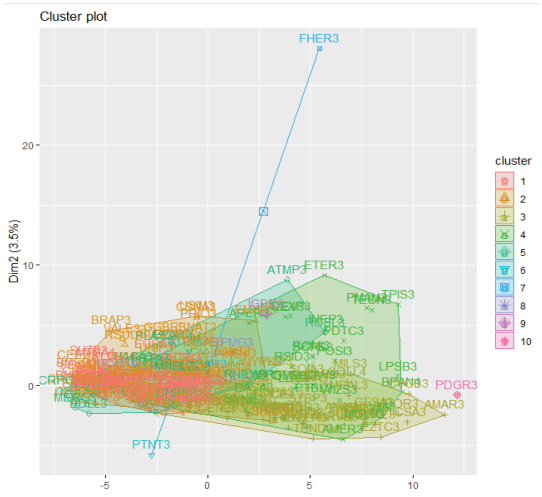

Figura 6: Projeção Componentes Principais Agrupamento Hierárquico 10 clusters.

O ticker FHER3 se apresenta isoladamente, sendo este um ticker que passou por um momento de forte alavancagem, resultando em uma repentina supervalorização, para logo em seguida descer seu patamar de negociação.
Essas variações atípicas tornaram esse ticker um provável outlier, dessa forma não apresentando uma similaridade em relação as demais empresas, ficando isolado em um cluster.
Em prosseguimento as análises, foi então testado o agrupamento com cinco clusters. O dendograma é apresentado na Figura 7.

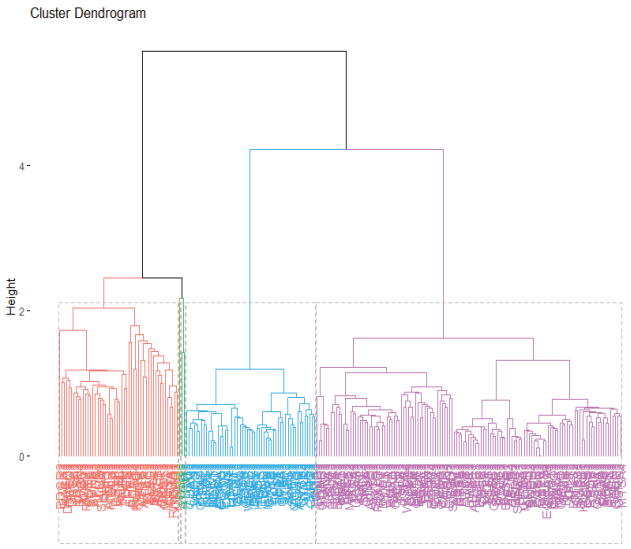

Figura 7: Dendograma Agrupamento Hierárquico 5 clusters.

Considerando cinco clusters, em dois deles foram populados com possíveis outliers, sendo um cluster com um ticker e outro com dois tickers.
A projeção dos componentes principais pode ser observado na Figura 8.

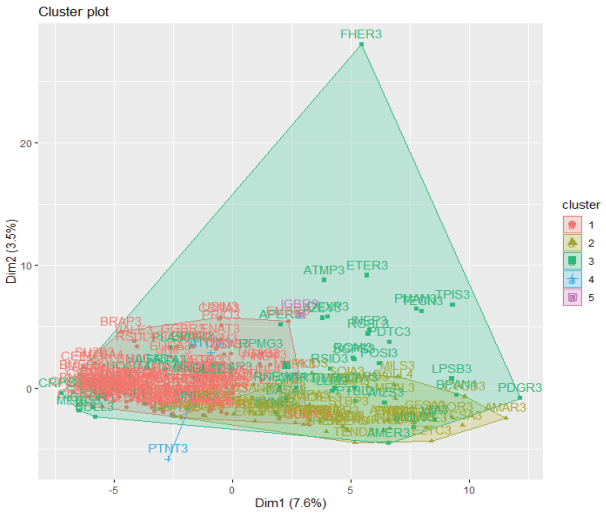

Figura 8: Projeção Componentes Principais Agrupamento Hierárquico 5 clusters.

Observa-se que o ticker FHER3 que havia ficado isolado em um agrupamento anteriormente, foi agora agrupado a outros tickers.
Desta vez os tickers que se mantiveram isolados foram os seguintes: ATOM3, IGBR3 e PTNT3.
Como ultimo experimento, foram então retirados esses três tickers da análise. A Figura 9 apresenta o dendograma.

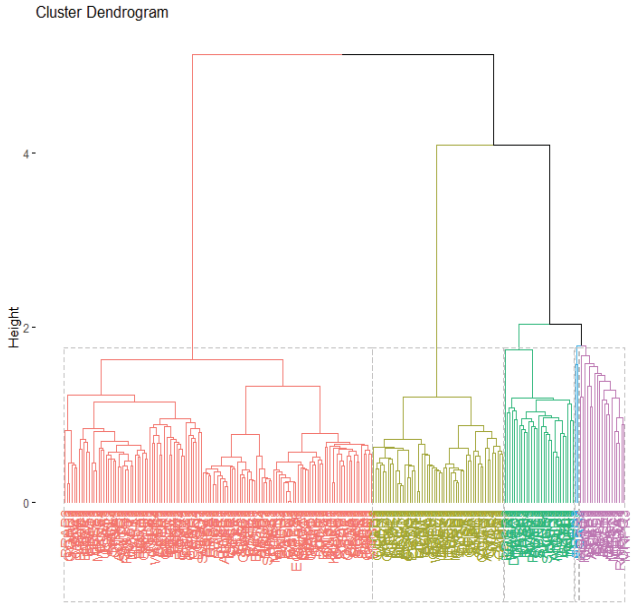

Figura 9: Dendograma Agrupamento Hierárquico 5 clusters, sem possíveis outliers.

Com a retirada dos possíveis outliers, apenas um cluster captou possíveis outliers, contendo dois tickers.
A projeção dos componentes principais é apresentado na Figura 10.

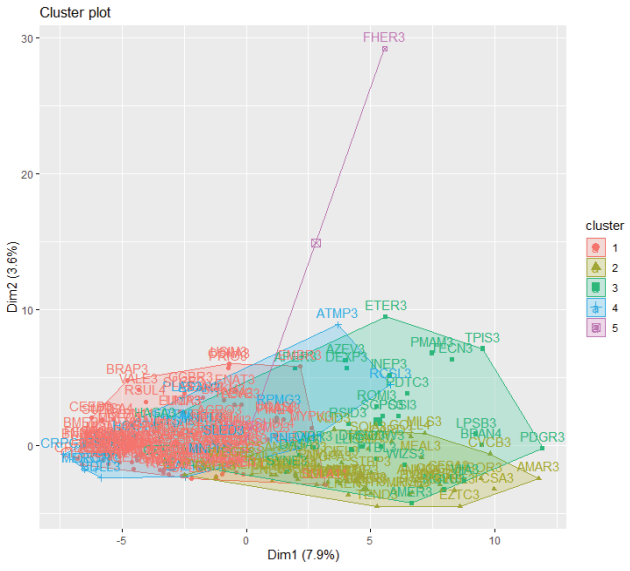

Figura 10: Projeção Componentes Principais Agrupamento Hierárquico 5 clusters, sem possíveis outliers.

3.2 Resultado K-means

Para o algoritmo de K-means, a aplicação do método de silhouette resultou como sendo 7 o número ótimo de clusters, conforme Figura 11.

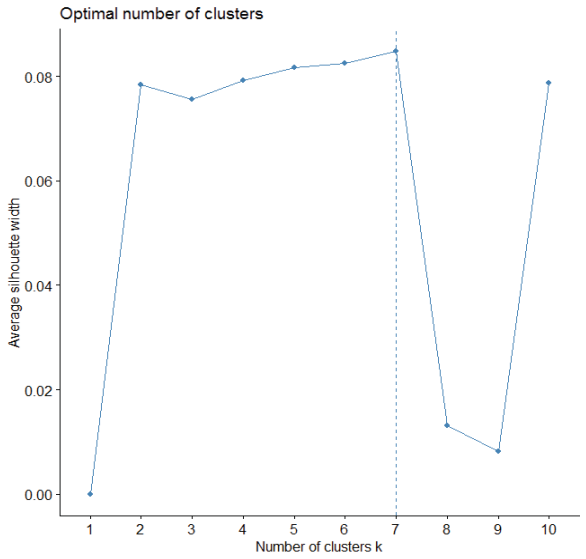

Figura 11: Número ótimo de clusters K-means.

O algoritmo foi então aplicado considerando 7 centróides.
Diferentemente da aplicação do algoritmo de agrupamento hierárquico, a aplicação do algoritmo K-means concentrou a maior parte dos tickers em apenas dois clusters, conforme observado na Figura 12.

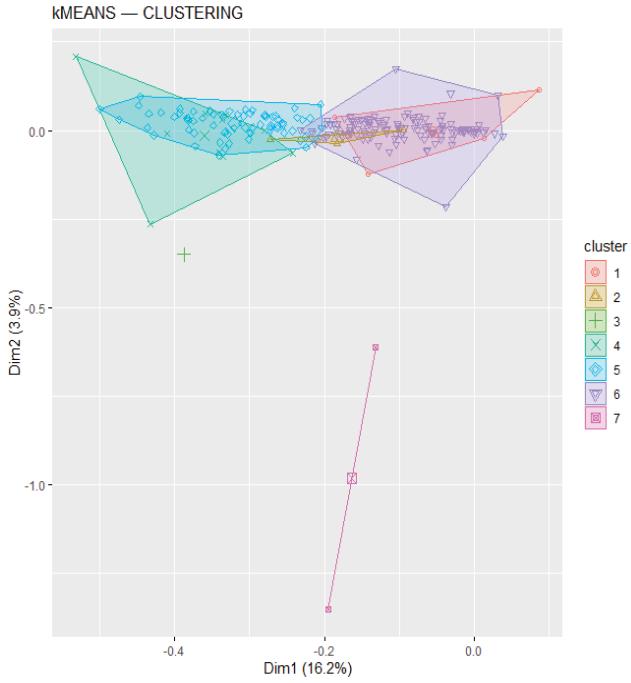

Figura 12: Projeção Componentes Principais K-means 7 clusters.

Este resultado acaba por limitar seu uso como insumo para a montagem de uma carteira investimentos diversificada, visto que a maioria dos clusters contém ações isoladas de empresas que apresentaram comportamentos atípicos, ou seja, possíveis outliers.

3.3 União das Análises

O passo final do estudo consiste na união entre a análise fundamentalista da empresa e a análise de agrupamentos.
Na análise dos indicadores fundamentalistas, apresentados na Tabela 2, deve-se observar o retrato financeiro da empresa, suas expectativas e avaliar se a mesma se encaixa no perfil de investimento desejado.
E através da análise de agrupamentos, filtrar entre as empresas com os fundamentos desejados, tickers de diferentes clusters, visando a diversificação dos ativos, de forma a minimizar o risco dos investimentos.
A análise conjunta pode servir como insumo para elaboração de uma carteira de investimentos diversificada e segura, considerando diferentes estratégias.
Como exemplo, buscar empresas com o preço da ação descontada, considerando o indicador fundamentalista P/VP (Preço sobre o Valor Patrimonial) abaixo de um. Podendo também acrescentar outros filtros na consulta, como por exemplo, entre as empresas descontadas, empresas com dividend yield (percentual de lucro distribuído ao acionista em relação ao valor da cotação) superior a algum determinado ponto de corte, em diferentes clusters, caso o interesse principal seja o retorno via dividendos, conforme apresentado na Figura 13.

	ticker	P_VP	Div_Yield	cluster
1	CIEL3	0.93	4.1	2
2	CLSC3	0.85	6.0	1
3	CPLE3	0.82	9.1	1
4	CSMG3	0.77	4.9	1
5	EVEN3	0.62	5.0	2
6	GGBR3	0.89	14.3	1
7	GOAU3	0.77	22.3	1
8	GUAR3	0.87	5.4	2
9	MYPK3	0.51	10.5	1
10	POMO3	0.75	4.6	1
11	SAPR3	0.75	5.2	1
12	USIM3	0.57	11.2	1

Figura 13: Tabela de insumos filtrada com empresas com valor descontado e Div Yield superior a 4%.

Outras estratégias podem ser abordadas com a análise

conjunta dos indicadores, mas sempre buscando a dilui-ção do risco investindo em empresas não correlaciona-das.

4 Conclusões

Para ambos os algoritmos, alguns resultados esperados se confirmaram, como por exemplo, empresas do mesmo ramo, onde espera-se a correlação das mesmas, sendo agrupadas no mesmo cluster.
A similaridade de variações entre as empresas nem sempre é intuitiva e, neste sentido, o algoritmo de agrupamento hierárquico se mostrou eficaz para a captação dessas correlações, apesar de ter gerado agrupamentos contendo outliers.
Já os agrupamentos gerados a partir do algoritmo K-means se mostrou pouco eficiente no objetivo de avaliar a similaridade entre as empresas, visto que pouco conseguiu segregar as variações diárias das cotações.
É importante ressaltar que este estudo deve server apenas como um insumo para definição das empresas a serem feitas as aplicações. Outros pontos devem ser observados para a tomada de decisão.

Agradecimentos

Agradeço ao Professor Dr. Fernando de Pol Mayer pela orientação no desenvolvimento deste estudo, aos demais professores da especialização por todo o conhecimento compartilhado e a minha esposa Joelma pelo incentivo em todo o período de especialização.

Referências

[1] R Development Core Team. (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, url: *http://www.R-project.org*.

[2] *Indicadores Fundamentalistas*, Fundamentus, Acessado 29 de março de 2022 Disponível em: https://www.fundamentus.com.br/detalhes.php?papel=.

[3] Marcelo Perlin (2022), *BatchGetSymbols: Downloads and Organizes Financial Data for Multiple Tickers. R package version 2.6.4.*, https://cran.r-project.org/web/packages/BatchGetSymbols/index.html

[4] *Clustering*, Scikit Learn, Acessado 29 de março de 2022 Disponível em: https://scikit-learn.org/stable/modules/clustering.html#clustering