American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

http://asrjetsjournal.org/

Clustering of Assets: An Alternative to Assist in Financial Decisions

Daiane Rodrigues dos Santos[a]*, Tuany Esthefany Barcellos de Carvalho Silva[b], Campo Elias Suárez Villagrán[c], Tiago Costa Ribeiro[d]

[a]*Universidade Cândido Mendes, Rio de Janeiro, Brazil*
[b]*Pontifícia Universidade Católica, Rio de Janeiro, Brazil*
[c]*Instituto de Matemática Pura e Aplicada, Rio de Janeiro, Brazil*
[d]*IBMEC, Rio de Janeiro, Brazil*
[a]*Email: economista.daiane@gmail.com,* [b]*Email: tuanybarcellos@id.uff.br,* [c] *Email: camplise@gmail.com,*
[d]*Email: tiagor86@gmail.com*

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 article proposes clustering as a metric to separate a set of assets, through a grouping method that maximizes the similarity inside groups with the purpose to reduce the risk of the portfolio. For the period analyzed - january 2019 to january 2020 – we have got 8 different assets clusters with a minimum of 1 (1.32% of total assets) and a maximum of 30 assets (42.86% of total assets).

Palavras-chave: financial assets; cluster; financial decisions.

1. Introduction

The advent of the financial market is one of the most fascinating events of our time. It has had a significant impact on many areas such as business, education, jobs, technology and therefore on the economy. Over the years, researchers and investors have been interested in developing tools to assist in decision making regarding the allocation of capital in the countless asset possibilities offered by financial globalization.

---

* Corresponding author.

Making decisions to allocate financial resources and manage portfolios in a satisfactory manner are stressful tasks due to the randomness in the market and biases of human behavior related to investment and irrational decision-making. Cluster analysis serves as a method to discover which assets are different from each other, assisting in the decision-making process [6]. There are several factors that affect the decision to allocate assets, such as personal objectives, risk tolerance level, investment horizon, rare disasters, transactional factors, and fixed costs of stock market participation. In addition, decisions are influenced by cognitive processes. Recent studies point out that agents tend to be too optimistic about their life perspectives, this directly affects their financial decisions. Overconfidence includes overestimation and over accuracy. Considering this, the quantitative tools of volatility (risk) analysis are extremely important for an efficient allocation decision given the randomness of the financial market [11]. Risk is mitigated when the investment portfolio manager chooses assets that tend to behave differently in reaction to a given factor. For example, two actions of the same sector tend to move together in the presence of factors that are associated. The inclusion of a tax, or incentive, to a certain industry is expected to affect in the same direction the actions included in that sector. Similarly, we can expect Brazilian stocks to move together and disassociate from American stocks, in the presence of some factor that affects exclusively the Brazilian market. The most common metric for estimating the level and direction of interactions between assets is correlation. Within a portfolio view, correlations should be evaluated exclusively among all assets, which generally makes it difficult to see the expected behavior for the portfolio. This article proposes clustering as a metric to separate a few assets, which make up the Ibovespa, through a grouping method that maximizes and minimizes the similarity between the groups to mitigate portfolio risk. The use of the metric described above allows us to discover combinations of assets that can compose a more diversified portfolio and with less risk. The metric also allows us to evaluate the robustness of the relationships between clusters and various risk factors over time.

2. Investments in equity income and investment portfolios

In the financial market there are several asset classes from different segments for the allocation of investors' resources. According to [6], among the options available in the financial market, the capital market, which houses the variable income segment, allows investment in multiple forms, or asset classes, such as investment in shares traded on the stock exchange. Diversification is a widely used strategy to reduce the risk of a portfolio to a specific level. In other words, the objective of diversification is to reduce the risk measure used. For example, the volatility of the portfolio, through the inclusion of uncorrelated assets, so that the volatility of the portfolio is substantially lower than the weighted average volatility of the assets. To elaborate a portfolio properly balanced to the point of optimizing the expected return per unit of risk used is not a trivial task, and can be evaluated in different aspects through different alternatives. The economic reality confirms the existence of the interrelation between return and risk. This stylized fact can be considered as the reason for numerous academic and market studies regarding asset portfolio diversification. This article, as mentioned above, presents a form of analysis to assist in the choices of your equity portfolio.

3. Cluster Analysis

Grouping analysis, or clustering, consists of computational techniques that allow the separation of objects into

groups. This analysis is a Multivariate Statistics procedure that aims to partition the elements into two or more clusters considering their similarity according to pre-established parameters. Such parameters, according to [11], are usually based on a dissimilarity function that receives two objects returning the distance between them. After the implementation of a quality metric the groups should present a high internal homogeneity and external heterogeneity. That is, the elements of a set must be mutually similar and different from the elements of other sets [7]. Clustering can be seen as an auxiliary tool. The metric in question can be used as a set of procedures to organize time series based on data of similarity or dissimilarity between them. It is the fitting of a high dimension space in a structure similar to a tree, represented in dendrograms. The dissimilarity between objects is measured by a distance matrix whose components resemble the distance between two points. This technique can be described as a two-step process: (i) the choice of a distance measurement and (ii) the choice of the cluster algorithm. These two steps together define the entire result of the cluster [1]. Financial asset return time series focus on the dissimilarity between synchronous time evolutions of a group of assets. The matrix of distances between these assets will be the entry of the hierarchical cluster algorithm that uses a linking rule to determine a hierarchical structure. After the proximity index has been defined, and the distance matrix calculated, the hierarchical clustering can be performed by an appropriate clustering algorithm.

3.1. Dissimilarity Measures

According to [7], similarity between the elements is an empirical measure of correspondence, or similarity, between the objects that will be grouped. Grouping methods can be described by a matrix containing a measure of dissimilarity or closeness between each pair of objects, where each entry $\boldsymbol{p_{ij}}$ in the matrix is a numerical value that demonstrates how close the objects $\boldsymbol{i}$ and $\boldsymbol{j}$ are. The dissimilarity coefficients presented are functions $\boldsymbol{d}: \boldsymbol{\Gamma}\ \boldsymbol{X}\ \boldsymbol{\Gamma} \Rightarrow \mathfrak{R}$, where $\boldsymbol{\Gamma}$ represents the set of interest objects. These functions allow the transformation of the data matrix,

$$\Gamma = \begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{1l} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix} \tag{1}$$

In a distance matrix,

$$\mathbf{d} = \begin{bmatrix} \mathbf{0} & & & & \\ \boldsymbol{d}(\mathbf{2},\mathbf{1}) & \mathbf{0} & & & \\ \boldsymbol{d}(\mathbf{3},\mathbf{1}) & \boldsymbol{d}(\mathbf{3},\mathbf{2}) & \mathbf{0} & & \\ \vdots & \vdots & \vdots & & \\ \boldsymbol{d}(\boldsymbol{n},\mathbf{1}) & \boldsymbol{d}(\boldsymbol{n},\mathbf{2}) & \dots & \dots & \mathbf{0} \end{bmatrix} \tag{2}$$

Being, $\boldsymbol{d}(\boldsymbol{i},\boldsymbol{j})$ the calculated distance between the elements $\boldsymbol{i}$ and $\boldsymbol{j}$.

The dissimilarity functions need to follow some criteria:

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

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

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

After meeting the properties listed above, if the metric also has property $\boldsymbol{d(ax,ay) = |a|d(x,y)}$, then it is called the norm. There are many metrics of dissimilarity, in this work the applied metric was the Euclidean distance, which is given by the following equation:

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

3.2. Grouping Heuristics

For the construction of clusters there are two techniques of grouping, known as the hierarchical method which consists of identifying clusters and the probable number $\boldsymbol{g}$ of groups, by a series of successive mergers, or a series of successive divisions, having its results observed in the dendrogram, which illustrates the mergers or divisions made in successive levels. The other method is known as non-hierarchical, where the number $\boldsymbol{g}$ of groups is pre-established. This technique consists in directly finding a partition of $\boldsymbol{n}$ items in clusters, by two requirements as, internal similarity and isolation of the $\boldsymbol{k}$ clusters formed. In this work the non-hierarchical K-Means technique is used [9].

3.2.1. Factorial Analysis (main components)

To determine an initial $\boldsymbol{k}$ for application of the non-hierarchical k-media technique, factor and principal component analysis was used. Factorial analysis provides the description of the variability of correlated variables observed in a smaller number of unobserved variables, which are linearly related to the original variables. The observed variables are modeled as a linear combination of common factors added to a random error,

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

Then,

$Z_i = \frac{X_i - \mu_i}{\sigma_i}$: is the standardized variable of the equation

X_i: original variable with average μ_i and variance σ_i^2

ε_i: i-th random error, being $i = 1, \ldots, p$

F_j: j-th common engine, being $j = 1, \dots, n$

l_{ij}: coefficient of the i-th standardized variable Z_i in the j-th factor F_j

Factorial Analysis assumes the existence of a statistical model that uses regression techniques to test hypotheses and is related to the analysis of principal components [8]. Principal component analysis (PCA) is a multivariate technique for modeling the covariance structure. Its main objective is to explain the covariance and variance structure of a random vector composed of random variables by linearly combining the original variables, called principal components [5]. As this analysis seeks to explain most of the total variation, it is adequate to extract the largest proportion of the variance with the smallest number of factors. Therefore, through this analysis it is possible to define a value for $\boldsymbol{k}$ and use it in the application of the non-hierarchical method K-Means.

3.2.2. K-Means Method

K-Means is a nonhierarchical grouping heuristic that aims to minimize the distance of objects to a set of $\boldsymbol{k}$ centers. The distance between a point $\boldsymbol{p_i}$ and a cluster is given by $\boldsymbol{d(p_i, \chi)}$, defined as the distance from the point to the nearest center. The function to be minimized is then given by:

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

The algorithm depends on a parameter $\boldsymbol{k}$ = number of clusters, defined by the user. In this work parameter $\boldsymbol{k}$ was defined through factor and main components analysis. The non-hierarchical method applied allows the previous definition of the number of clusters. At each stage, new clusters can be formed by dividing or joining clusters initially defined, without the need for dendrogram observations. The algorithms are iterative and have a greater capacity for data set analysis, and each item is allocated to a cluster that has a closer (average) centroid [10].

4. Results

Table 1: Cluster Assets

Cluster	Assets
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

Source: own elaboration

For the analyses and preparation of the clusters, 70 assets were collected, through the B3 and ANBIMA database, in a daily period from January 02, 2019 to January 31, 2020. Table 1 presents the total number of clusters with their respective assets, which were grouped through factor and cluster analysis.

Table 2 presents the descriptive statistics: the minimum, average, maximum and variance values within each cluster in a daily period from January 2019 to January 2020. Note that cluster 5 showed a lower variance compared to the others, while cluster 8, composed only of the asset CIEL3, was the one that showed the greatest variability in the observed period, being the only one to show a negative average return.

Table 2: Descriptive cluster statistics

Cluster	Minimum	Average	Maximum	Variance
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

Source: own preparation based on B3 data

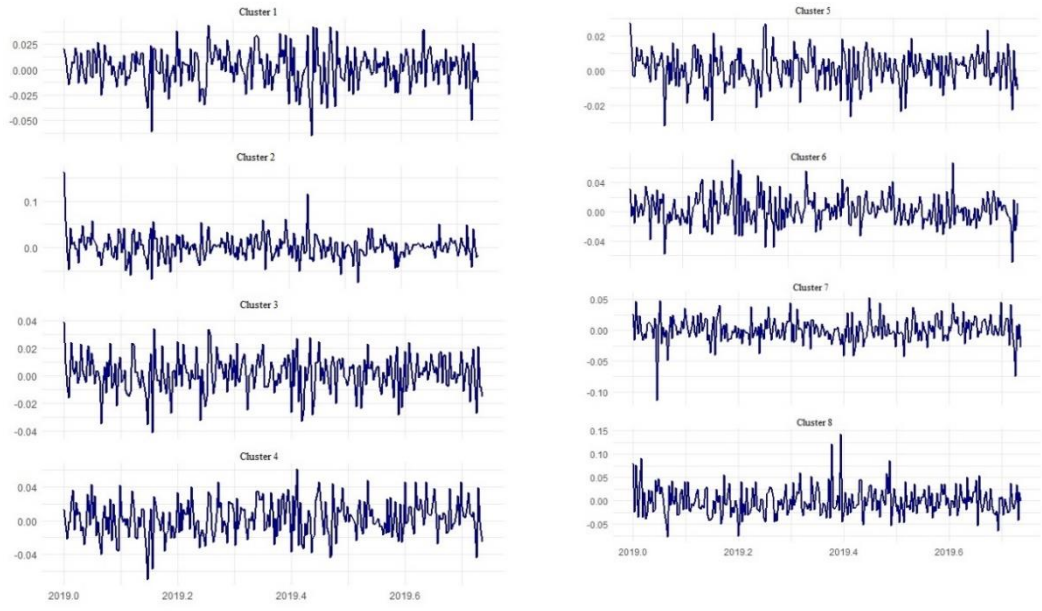

Figure 1: Series containing the average asset returns by cluster

Source: own preparation based on B3 data

Figure 1 shows the series of each cluster individually, and Table 2 shows the minimum and maximum points. Note that all clusters have a stable series around the average. In cluster 5 it is possible to identify a lower variability in the analyzed period, while in clusters 7 and 8 it is possible to observe the lowest average returns among the others. Cluster 8 has the highest variance, reaching its maximum point in October 2019, and shows a negative average. Cluster 7 has the lowest average among the positive returns. The highest average is presented by cluster 4.

To analyze the daily performance index of the assets in order to define a value that enables the use of the k-means technique for the cluster analysis, factor analysis and principal component analysis were applied. 70 assets were observed. To verify whether the use of the factor analysis technique is appropriate, the Bartlett and Kaiser-Meyer-Olkin (KMO) test of sphericity was applied. The purpose of the Bartlett's test of sphericity is to check whether all correlations within the matrix are significant. The hypotheses to be tested are:

$$\begin{cases} H_0: The\ variances\ of\ the\ groups\ are\ equal \\ H_1: The\ variances\ of\ the\ groups\ are\ different \end{cases}$$

The test result showed p-value < 2.2e-16, so, considering a significance level of 5%, we have evidence to reject the null hypothesis $\boldsymbol{H_0}$. That is, the variances of the compared assets are different, so according to Bartlett's test the use of the photo analysis is appropriate. The KMO (Kaiser-Meyer-Olkin) test was applied to assess the adequacy of the sample size. The result of this test varies between 0 and 1 and results above 0.5 are acceptable for factor analysis. In this test we obtained KMO = 0.92, so the sample is suitable for factor analysis. Once the tests were done, factor analysis was applied and, by means of the diagonalization of positive symmetric semi-definite matrices, the principal components were obtained. The first principal component accounts for about 31% of the total variance of the standardized data, while if we take the first eight components the proportion is about 70% of the total variance. These factors represent the minimum number of causes that condition a maximum of existing variability. That is, the factor analysis is based on 8 factors, so for the cluster analysis and application of the k-means technique $\boldsymbol{K = 8}$ will be used. Stock markets are affected by many highly interrelated factors that include economic, political, psychological and company-specific variables. Technical and fundamental analysis are the two main approaches to analyze financial markets. To invest in stocks and achieve returns with low risks, investors have used these two main approaches to make decisions in the financial markets. For the cluster analysis it was chosen the non-hierarchical method applying the k-means technique. The choice of this approach was due to the size of the observed sample, because for data sets, considered large, this method presents more significant results. For its use, the value of $\boldsymbol{k}$ must be previously established. Once the factor analysis was used, this value was defined by $\boldsymbol{k = 8}$, that is, the observed assets will be grouped into eight clusters by their internal similarity. This partition method allows us to measure the proximity between the groups of assets, using the Euclidean distance between the centroids of these groups. After performing the analysis and implementing the pre-established methods, it can be seen in Table 3 and Figure 2 the division of assets grouped into each cluster, it is noted that cluster 5 (sky blue) has about 42.86% of the observed assets, that is, 30 of these assets have similar characteristics and are highly correlated, while cluster 8 (pink) has only one asset, so it showed no similarity with any other asset within the sample in the selected period. One can observe the clusters with their respective assets in Figure 4 according to Table 1 presented above.

Table 3: Quantity of assets that composes the Clusters

Cluster	Quantity of assets	Participation Percentage
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%

Source: own preparation based on B3 data

The assets are connected in Figure 2 by a green or red straight line. The green lines correspond to positive correlations and the red ones to negative correlations. It is worth noting that the thicker (darker) the line is, the stronger the correlation between the assets. As can be seen, the shares of Azul Linhas Aéreas Brasileiras and Gol Linhas Aéreas present strong and positive correlation, similarly the assets LAME4 and BTOW3, ECOR3 and CCRO3 also present the same correlation pattern, strong and positive.

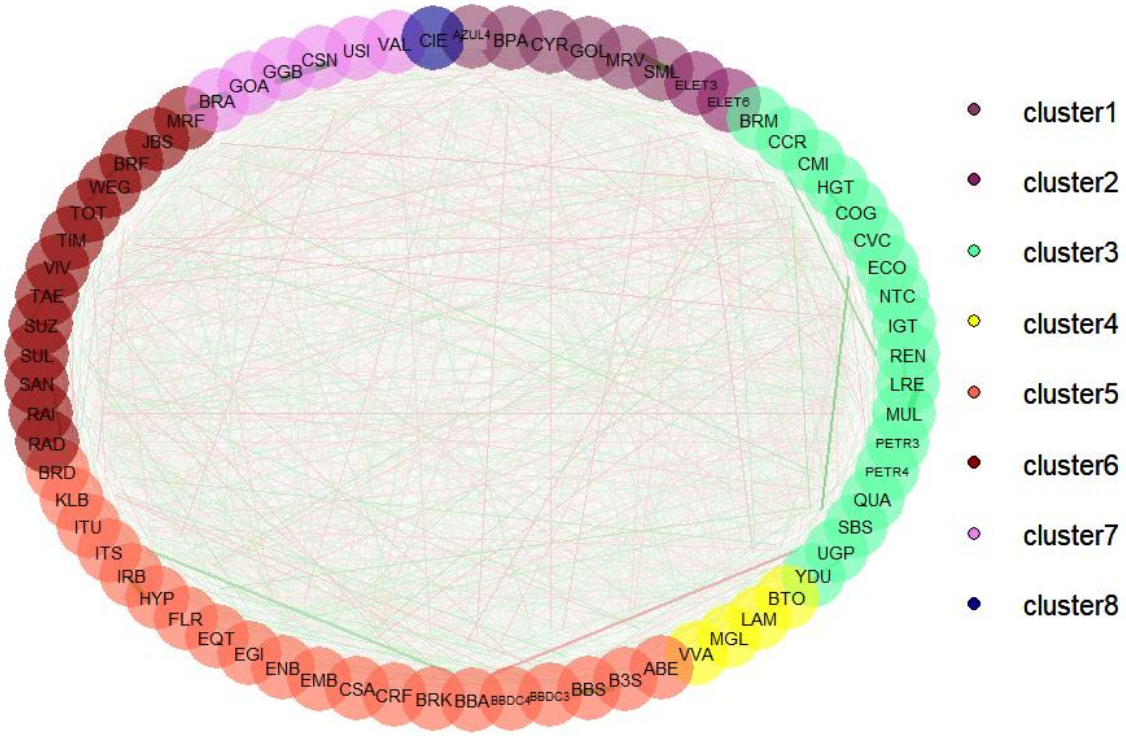

Figure 2: Clusters containing the assets.

Source: own preparation based on B3 data

After the clustering of the selected assets (Table 1), we used regressions to test whether the assets belonging to the clusters respond similarly to some macroeconomic variables and financial indices (The IPCA, the ANBIMA Market Index - IMAB, the exchange rate - Dollar the SMLL Index - small caps and the Ibovespa (the result is shown in the tables below).

Table 4: Cluster 1: assessment of the dependence of the returns on the assets that make up cluster 1 on economic variables and financial indexes (explanatory variables)

Cluster 1					
P-values referring to the significance of the betas of the multiple regression					
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	-
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	1	6	3	6	0
Percent	17%	100%	50%	100%	0%

Source: own preparation based on B3 data

As can be seen in Table 4, all 6 assets that make up cluster 1 had a p-value of less than 0.05 (5% significance level), that is, we reject the hypothesis that the beta is equal to zero for the explanatory variables IMAB and SMLL. Three of the six variables showed significance for the Dollar and only 1 for the IPCA. This result shows us that the variation in the returns of the assets that make up the cluster can be written as a linear function of the aforementioned variables. It is noteworthy that a more detailed modeling can be done to verify the degree of impact of the selected economic variables. It is worth noting that the regressions presented here attempted to have good levels of adherence (R^2 higher than 0.65, Statistics F below 0.05 and normalized residues). In Table 5 it is possible to observe that all the assets of the company Centrais Eletricas Brasileiras SA (ELET3 and ELET6) that make up cluster 2, presented a p-value of less than 0.05, that is, we reject the hypothesis that the beta is equal to zero for the explanatory variables IMAB and SMLL. In other words, the variation in the returns on assets ELET3 and ELET6 can be written as a linear function of the changes in the index of government bonds indexed to inflation measured by the IPCA and the Small Cap index (SMLL) that reflects the assets of the companies with the lowest capitalization in the country B3.

Table 5: Cluster 2: assessment of the dependence of the returns on the assets that make up cluster 2 on economic variables and financial indexes (explanatory variables)

Cluster 2					
P-values referring to the significance of the betas of the multiple regression					
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
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	0	2	0	2	0
Percent	0 %	100%	0%	100%	0%

Source: own preparation based on B3 data

Table 6: Cluster 3: Assessment of the dependence of the returns on the assets that make up cluster 3 on economic variables and financial indexes (explanatory variables)

Cluster 3					
P-values referring to the significance of the betas of the multiple regression					
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	-
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	4	14	3	18	0
Percent	22%	78%	17%	100%	0%

Source: own preparation based on B3 data

According to the results presented in Table 6, all assets that make up cluster 3 had a p-value of less than 0.05, that is, we reject the hypothesis that the beta is equal to zero for the explanatory variable SMLL, whereas whereas only 17%, 22% and 78% of the variations in the prices of the shares that make up the cluster answer the variables Dolar, IPCA and IMAB respectively.

As can be seen in Table 7, all assets that make up cluster 4 had a p-value of less than 0.05, that is, we reject the hypothesis that the beta is equal to zero for the IMAB. While 75%, 25% and 25% of the variations in the prices of the shares that make up the cluster respond to the variables IPCA, Ibovespa and SMLL, respectively.

Table 7: Cluster 4: assessment of the dependence of the returns on the assets that make up cluster 4 on economic variables and financial indexes (explanatory variables)

Cluster 4					
P-values referring to the significance of the betas of the multiple regression					
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	-
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	1	4	0	3	1
Percent	25%	100%	0%	75%	25%

Source: own preparation based on B3 data

Regarding clusters 5 and 6, we can see in Tables 8 and 9 that the variability of the only economic variable that influenced the variability in the return of all assets in the cluster was SMLL, that is, the economic performance of the clusters is quite sensitive to this variable.

Table 8: Cluster 5: assessment of the dependence of the returns of the assets that make up cluster 5 on economic variables and financial indexes (explanatory variables)

Cluster 5					
P-values referring to the significance of the betas of the multiple regression					
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
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	4	26	5	30	18
Percent	13%	87%	17%	100%	60%

Source: own preparation based on B3 data

Table 9: Cluster 6: assessment of the dependence of the returns of the assets that make up cluster 6 on economic variables and financial indexes (explanatory variables)

Cluster 6					
P-values referring to the significance of the betas of the multiple regression					
Ativos	**IPCA**	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
BRFS3	-	-	-	9,53E-05	-
JBSS3	-	-	-	5,65E-05	-
MRFG3	-	-	-	4,12E-05	-
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	0	0	0	3	0
Percent	0%	0%	0%	100%	0%

Source: own preparation based on B3 data

Table 10 shows that all 6 assets that make up cluster 7 had a p-value of less than 0.05 (significance level of 5%), that is, we reject the hypothesis that the beta is equal to zero for the explanatory variables Ibovespa and SMLL. In Cluster 8, formed only by the asset Cielo SA, CIEL3, the variations of IMAB, Ibovespa and SMLL can be considered as dependent variables.

Table 10: Cluster 7: assessment of the dependence of the returns on the assets that make up cluster 7 on economic variables and financial indexes (explanatory variables)

Cluster 7					
P-values referring to the significance of the betas of the multiple regression					
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
Percentage for adherence of selected variables					
	IPCA	**IMAB**	**Dólar**	**SMLL**	**Ibovespa**
Number of assets	5	5	0	6	6
Percent	83%	83%	0%	100%	100%

Source: own preparation based on B3 data

5. Concluding Remarks

This paper aimed to present the clustering method to assist investors in making decisions. The implemented methodology can be used to make a more assertive choice of investment portfolios, because by grouping the assets according to their similarities it is possible to analyze more clearly their returns, thus generating more information for the manager's decision making. For the analysis and preparation of this research, 70 assets were selected in a daily period from January 02, 2019 to January 31, 2020. In this article we applied the methodology of cluster analysis using the non-hierarchical K-means technique and obtained 8 clusters with assets grouped according to the similarity of their daily returns. Cluster 4 consists of 4 assets and together they have the highest average return, cluster 5 is the one that contains the largest number of assets and has the lowest variability compared to the others, cluster 8 has only the asset CIEL3, showing high variability in the analyzed period, being the only one to show a negative average return. With this study it was possible to demonstrate the effectiveness of the grouping method as a tool to assist in decision making and in developing new portfolios, being able to diversify them according to the characteristics of each investor. Regarding the relations with the selected variables, the AMBIMA market index (IMAB) and the SMLL index (small caps) are the variables that most relate to the clusters and the IPCA and Ibovespa variables are the ones that showed less significance in the application econometric approach proposed in this article. This information can assist in financial decisions and mitigate risks, since investors can choose assets and / or clusters that are related to different variables. The objective of future work is to continue the cluster analysis using other methods, starting with the confirmatory factor analysis method. The objective of future work is to continue the cluster analysis using other methods, starting with the confirmatory factor analysis method.

References

[1]. Anderson, T. An introduction to multivariate statistical analysis. New York: John Wiley & Sons, p. 675.1984.

[2]. Bussab, W., Miazaki, E., Andrade, D. Introdução à análise de agrupamentos. São Paulo: Associação Brasileira de Estatística, p. 105. 1990.

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

[4]. Halkini, M., Batistakis, Y., Vazirgiannis, M. On Clustering Validation Techniques, 2001.

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

[6]. 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.

[7]. Linden, R. Técnicas de Agrupamento. Revista de Sistemas de Informação da FSMA, p. 18-36, n. 4. 2009.

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

[9]. Nievola, J. Análise de Agrupamento. PPGIa, PUCPR. 2006.

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

[11]. 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.

[12]. Quintal, G. Análise de clusters aplicada ao Sucesso/Insucesso em Matemática. Universidade da Madeira Departamento de Matemática e Engenharias.2006.

[13]. [13] Reis, E. Estatística multivariada aplicada. Lisboa: Edições Silabo, p. 342. 1997.

[14]. [14] 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, v.9, n.1. 2012.

[15]. Sarajane M., Clodoaldo A. Técnicas de Agrupamento (Clustering). 2015.

[16]. 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.

[17]. Totti, R., Vencovsky, R., Batista, L. Utilização de métodos de agrupamentos hierárquicos em acessos de Paspalum (Graminea (Poaceae)). São Paulo, 2001.

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