Portfolio composition based on hedge funds using clustering and partial correlation coefficients

Francisco Santos
Instituto Superior Técnico
Universidade de Lisboa
Lisboa, Portugal

***Abstract*— In this work, it was used publicly available information about the investments made by two big different groups of hedge funds to solve the problem of portfolio composition with the main goal to beat the S&P 500. It was used three different methods to create portfolios of stocks the first method chooses the stocks which have the highest investment by the funds to create the portfolio, the second method used the division of stocks in sectors and in the third method, it was implemented an alternative system to divide the stocks. This system first calculates the partial correlation coefficients of the companies between each other instead of the normal correlation because with partial correlation it was possible to remove the influence of the index which is a factor that drives correlation between the returns of the stocks up, after this, it was applied an agglomerative hierarchical clustering algorithm to divide the stocks. Only companies belonging to S&P 500 index were used and the simulation period was between fifteen of August 2013 and fifteen of May 2017. The obtained results were promising because not only it was possible to obtain better returns than the S&P 500, it was obtained average returns of 76.14 % whereas S&P 500 obtained 44.6 %, but also it was possible to understand the different kinds of strategy that works with different kinds of funds.**

***Keywords-* Clustering; Portfolio Composition; Partial Correlation; Hedge Funds**

I. INTRODUCTION

Financial markets are without a doubt important for the health of the economy because in a very simple way they allow moving funds where they are in excess to where they are needed the most making the economy more efficient. These markets are operated by humans which bring a random factor to it making a lot of aspects not predictable and therefore an interesting field of study. With a completely different approach to the existent funds in that time, it's in 1949 that the first hedge fund appears by the hand of Alfred Winslow Jones. The protection to the investors in these funds are less than in the other funds such as mutual funds because hedge funds are destined to experienced investors that know what they are doing because if it's possible to have higher returns investing in hedge funds it's also possible to have higher loses. Besides this fact, these funds started to get a lot of attention because of the fees their managers receive, these are the only funds with such high fees where the managers receive a management and a performance fee. Most of the times the managers are people highly qualified in the financial world considered the best funds managers and investors are willing to put huge sums of money in the funds, there are even certain situations where the position which a fund has in a stock can affect the direction where the market goes regarding that stock. Financial analysts, investors and the government have been classifying the companies based on their industry since 1930 [1]. In the last few years, the main used system has been the Global Industry Classification Standard (GICS) one of the goals of this work is to create an investment strategy based on a new industry classification system proposed by Jung and Chang in 2016 [2] applying it to the S&P 500, this alternative system uses an agglomerative hierarchical clustering algorithm to divide the companies based on their partial correlation coefficients, the use of this correlation allows to remove the influence of the market represented by the index S&P 500. Using this strategy, it's possible to create a portfolio of stocks based on publicly available information about the position of hedge funds with the goal to generate not only better returns than the index S&P 500 but also better than Whale Index, which is an index that is specialized in the use of positions taken by hedge funds, such a strategy can benefit from the knowledge of world-class managers without having to pay the huge fees. The main contributions of this work are not only the testing of a new industry dividing system in the S&P 500 but also apply this to the portfolio composition problem combined with the use of the publicly available information about the positions of two groups of hedge funds.

In this paper Section II presents the state-of-the-art for the partial correlation, clustering and portfolio composition. Section III describes the proposed system. In Section IV the case studies and results are discussed. The conclusions and future work and appear in Section V.

II. RELATED WORK

In the United States hedge funds managers that have under managements equity assets of at least hundred million dollars in value they must fill a 13F form to report their holdings to the Securities and Exchange Commission (SEC). These forms must be filled within the forty-five days of the end of each quarter and they are available to the public through Electronic Data Gathering, Analysis and Retrieval (EDGAR) in the link on [3]. In these forms, it's possible to know the long positions of the funds at the end of the quarter, the put and call options, number of stocks and their value. The short positions, the total cash of the fund and other asset classes are not revealed in these forms. Naturally, with the increase of the curiosity about hedge funds Whale Index was created. This is a long only index and uses the information from the 13F forms of the hedge funds, it's composed of one hundred stocks with the same weight and it's updated forty-six days after the end of each quarter when the new forms are available. This index uses a mechanism to classify the funds according to several indicators based on risk and return and then a score is attributed, in the end, there are chosen the best sixty funds to be used by the index.

A. *Correlation*

Correlation is a statistical measure that measures the linear relationship between two variables, it's possible to measure the degree of association between the variables using the correlation coefficient. These coefficients can take positive values when the variables move in the same direction, a negative value when the variables move in opposite directions and a value of zero when there is no relationship between the variables. The Pearson correlation coefficient is represented in (1) and it takes values in the interval of [-1,1].

$$\rho(X,Y)=\frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{(N-1)\sigma_x\sigma_Y} \qquad (1)$$

Where σ is the standard deviation, $\overline{x}$ and $\overline{y}$ are the average values of the variables. The partial correlation coefficient is represented in (2) and it measures how two variables X and Y are correlated when a common factor that affects both is removed.

$$\rho(X,Y:Z)=\frac{\rho(X,Y)-\rho(X,Z)\rho(Z,Y)}{\sqrt{[1-\rho^2(X,Z)][1-\rho^2(Y,Z)]}} \qquad (2)$$

The partial correlation coefficient is based on the Pearson coefficient and it also takes values from [-1,1]. To evaluate the statistical significance of the obtained coefficients, it's necessary to do a null hypothesis test in which is tested the hypothesis of the obtained coefficient to be zero compared to the hypothesis of not being zero.

$$H_0:\rho_{xy,z}=0\, vs\, H_1:\rho_{xy,z}\neq 0 \qquad (3)$$

In case of the null hypothesis being rejected then the other is accepted. To test the null hypothesis it's used a statistical test [4] that has the distribution of a t student and is represented in (4).

$$t=r_{xy,z}\sqrt{\frac{n-2-k}{1-r^2_{xy,z}}}\sim t_{n-2-k} \qquad (4)$$

The sample size is represented by **n**, **k** is the number of conditioning variables and $r_{xy,z}$ is the value of the partial correlation coefficient. To reject the null hypothesis condition on (5) has to be respected where α is the significance level.

$$|t|>t_{n-2-k,\alpha/2} \qquad (5)$$

Partial correlation has been used to study the markets in several ways. Xing Li et al. [5] used an impact coefficient based on partial correlation which removes the influence of the index but also the influence of a certain stock, to study the influence of intern and extern indexes to the Chinese market. Regarding the intern indexes, they concluded that although the stocks suffer influence from stocks in other sectors the biggest influence comes from the stocks in the same sector. They also concluded that the influence of extern indexes was low in the Chinese market. Focused now on S&P 500 index Kenett et al. [6] studied, between 2000 and 2010, how the companies are influenced by the several sectors and what is the influence between them. They calculated a coefficient that indicates how a company is influenced by the other sectors. The conclusions revealed that some companies are influenced the same way by all sectors and others receive more influence by one or two sectors. The energy and financial sectors are not really influenced by others and sectors such as industry and materials are severely influenced by others. Studying the 300 hundred biggest companies in the New York Stock Exchange (NYSE) between 2001 and 2003 Kenett et al. [7] tried to identify the dominant stocks. Two different methods were used one based on a partial correlation network and the other using a ranking of partial correlations according to its intensity. The authors concluded that the financial sector influences constantly the other sectors. Between the period of 2000 and 2007 in the S&P 500 index was studied what was the origin of the partial correlation [8] if it was due to the correlation between the stocks and the index or if it was due to other factors such as the use of the same kind of materials, business partnerships or sector organisation. This revealed a complex problem because after the influence of the index was removed the correlation between the companies reduced drastically however for some companies this was not true for all the companies which indicate that there are other factors involved.

B. *Clustering*

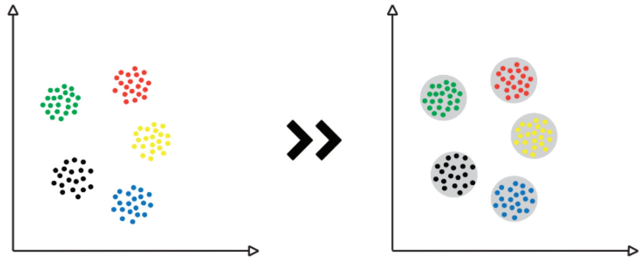

Fig. 1 - Clustering example

Clustering is the process of dividing objects that can be physical or abstracts into groups (clusters) according to a similarity measure. The goal is that the objects that belong to the same cluster are the most similar possible to each other and the most dissimilar possible to the objects in other clusters.

In this work, an agglomerative hierarchical clustering algorithm was used to divide the companies. In the agglomerative hierarchical methods, the initial number of cluster is the same as the number of objects and then the objects/clusters will be grouped until the desired number of clusters is achieved which, in this case, was ten clusters because it was the same number of sectors used in the GICS system. To group the clusters, it's necessary to calculate the distance between them, the distance between clusters is calculated by the average distance between all objects of each cluster. The Euclidean distance represented on (6) it's used as a measure of the distance between the objects.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2} \quad (6)$$

Jung and Chang [2] used an agglomerative hierarchical clustering algorithm with partial correlation to study the market structure dividing the companies of the Korean Stock Market (KOSPI) into clusters and then compare with the sectors of the GICS system. The results showed clusters composed by companies of several sectors not having a single cluster dominated by one sector.

C. Portfolio Composition

The problem of portfolio composition can be solved using several methods such as neural networks, decision trees and logistic regression models [9]. An approach using genetic algorithms is also commonly used, it can be single objective trying to increase the return [10] or it can be multi-objective trying to increase the return and reduce the risk [11] and [12]. Silva, Neves and Horta [11] based their strategy on the use of fundamental indicators to choose the stocks whereas the technical indicators were used in the process of buying and selling stocks and then the genetic algorithm was used to optimize the return and reduce the risk associated. With this strategy, it was possible to obtain better results than investing in the S&P 500 index. Besides the use of the genetic algorithm Pinto, Neves and Horta [12] also used technical indicators but now using the volatility index (VIX) to decide when to buy or sell the stocks. An approach using Clustering can also be used Nanda, Mahanty and Tiwari [13] tested three different algorithms: K-means, Fuzzy C means and self-organizing maps to divide the stocks then chose the best initial number of clusters and the best **N** stocks of each one. To create an efficient portfolio the Markowitz method was used to maximize the returns to a certain level of risk, the results were compared against the Sensex index and the best results were obtained using the K-Means algorithm. Using clustering as well Long et al. [14] applying their strategy to the Stock Exchange of Thailand, firstly the stocks are divided into **K** clusters using the algorithm fuzzy c means, having chosen the **K** number of clusters based on cluster validation methods to obtain the optimal number of clusters. After the stocks are divided its chosen **N** stocks representing each cluster and then the genetic algorithm was used to choose the importance of the stocks in the portfolio. Lorio, Frasso, D'Ambrosio and Siciliano [15] applied clustering to the P-spline coefficients and then the portfolio is built using only the best cluster. According to the authors in this case the clustering algorithm is not very important, what is important is to choose the optimal number of clusters. With a completely different approach, Yua, Chenb and Zhang [16] used Support vector machine (SVM) with principal component analysis (PCA) with the goal of finding the stocks that can originate a higher return and build the portfolio with only that stocks. It's applied PCA to the annuals reports of the companies to reduce the number of variables and obtaining new ones after this SVM is applied to select the stocks that are going to have the best returns.

III. SYSTEM ARCHITECTURE

The system was developed using Python programming language. The system is organized in layers as showed in *Fig. 2*

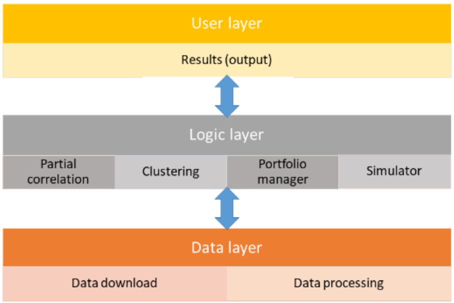

Fig. 2 - System Architecture

The user layer is responsible for every interaction between the system and the outside world, the logic layer is responsible for the implementation of all the concepts applied in this work. The data layer is responsible to manage all the data used mainly the download of information and then their processing.

It was implemented a long-term investment strategy, so the system is designed to work every time that the 13F forms are available to the public which corresponds to forty-six days after the end of a quarter and work again when new forms are out again. An example of this can be seen for the year of 2015 in the Table 1.

Quarters	Beginning quarte	End quarter	Beginning of the simulation	End of the simulation
1.º	1 of January of 2015	31 of March of 2015	18 of May of 2015	14 of August of 2015
2.º	1 of April of 2015	30 of June of 2015	17 of August of 2015	16 of November of 2015
3.º	1 of July of 2015	30 of September of 2015	17 of November of 2015	16 of February of 2016
4.º	1 of October of 2015	31 of December of 2015	17 of February of 2016	16 of May of 2016

Table 1 - Example of simulation dates

The system can work in two different ways: the first one it only uses the stocks used by the hedge funds to invest, the second, it also uses the stocks, however, first divide them into sectors or clusters and then chooses the stocks from them to invest. A sequence of steps for the second method is the following:

1- Read the config file which has the necessary parameters defined by the user.
2- Read the info on the intended hedge funds and process their data.
3- In case of clustering calculate the partial correlation matrix for that quarter and apply the agglomerative hierarchical clustering or divide the stocks into sectors if that's the case.
4- Build the portfolio according to the criterion defined by the user and simulate in the correspondent period.

If the used method is the first one, then the third step it's not done.

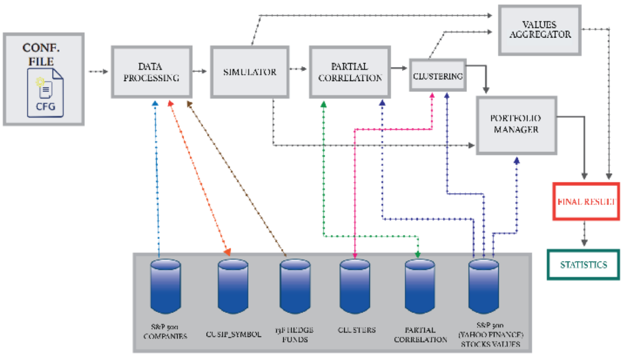

Fig. 3 - System overview

A. *Data Layer*

This layer is divided into two modules. One that is responsible to download the data and the other one that is responsible to process the data.

1) *Download data module*

All the information that needs to be download from the internet is done by this module. It was created a web crawler that received the names of the funds and then downloaded all the 13F forms from the EDGAR 2.0 database. Furthermore, it's also downloaded the historical values of S&P 500 companies using pandas data reader [17] that uses the API of Yahoo finance, this process is done for the period between 3 of January of 2011 and 1 of June of 2017.

2) *Data processing module*

This module is responsible to do all the data processing related to the 13F forms of the funds. On each quarter all the 13F forms are read and then normalized. After this, the stocks from S&P 500 are filtered and its constructed a super portfolio where each stock has the information of how much in total was invested by the funds, how many shares of that company the funds had, the number of funds that had that company on their portfolio and the average weight that this company had in the funds' portfolios. Besides this, it's also added the ticker of the company that was obtained looking in the database for that CUSIP.

B. *Logic Layer*

This layer is subdivided into four modules. The partial correlation module, the clustering module, the portfolio management module and the simulator module.

1) *Partial Correlation module*

The main functionality of this module is to calculate the partial correlation coefficient matrix, so the clustering module can use it. The adjusted close values of the companies are read and then the logarithmic returns are calculated for all the companies the next step is to apply the partial correlation formula in (2) in which the period used was thirty months. The coefficients are calculated for all the companies and then the matrix is filled and saved.

2) *Clustering*

The clustering is done only in the companies that the hedge funds had in that quarter instead of all the S&P 500 leading to a variable number of companies to cluster each quarter. This module implements the agglomerative hierarchical clustering.

3) *Portfolio Manager*

This module is responsible to create and simulate the portfolio. The portfolio can be created in two different ways: the first one is using solely the information of the funds which in this case it can be copied the exact portfolio of a certain fund, the funds can be ranked according to their performance and then choose only a few of them or it can simply use the number of funds defined by the user. After this, the maximum number of stocks defined by the user are chosen. The second way to create the portfolio is after all the information about the stocks of the funds are read they are divided into sectors according to GICS or in clusters using the proposed method. After this, the defined number of clusters/sectors are chosen according to a criterion that can be the return in the last quarter of the cluster/sector or by the amount of money the funds have invested in that cluster/sector. To maintain the diversity the same number of stocks are chosen from each cluster/sector based on the number of funds that have them and the amount of money invested in that stock. After the portfolio is done it's simulated to see how it behave in the market.

4) *Simulator*

This module is responsible for making the interpretation of the config file filled by the user and then implement the right order of steps executing the desired simulation, for this, it must establish communication with all the other modules.

C. *User Layer*

In the user layer is built all the graphics and statistics necessary to be used in the results section and to analyse the results.

IV. EXPERIMENTS AND RESULTS

A. *Partial correlation and clustering results*

1) *Partial correlation*

The calculus of the partial correlation coefficients took place every quarter after the 13F forms are available so, it's calculation has a sliding window of three months. Both the Pearson correlation and partial correlation coefficients were calculated and compared and although [2] studied the Korean market the difference between the average value of both coefficients also had a magnitude order difference as obtained in this work for the S&P 500. This can lead to the conclusion that the market index drives the correlation between the stocks up, however, this is not the only factor that influences the correlation between the stocks because the partial correlation coefficients were not zero after the index removal. To verify that the partial correlation coefficients obtained were statistically significant the null hypothesis test explained in Section II.A was made. Unfortunately, it was not possible to reject the null hypothesis for all the coefficients, so for that ones, it was assumed that the partial correlation was zero. This result can be explained as for some companies they don't have any common factors which is according to what was studied in the past by [8].

2) *Clusters analyses*

After calculation of the partial correlation coefficients, the agglomerative hierarchical clustering algorithm was applied to only the stocks held by hedge funds. This allowed a reduction from working with 431 stocks to work with between 200 and 300 stocks on each quarter. It was observed that for all the quarters simulated the funds chose stocks from all the GICS sectors available. Analysing the obtained clusters, it's possible to see that there is always a cluster constituted by several sectors. Whenever there are a lot of stocks belonging to the financial sector this one is divided almost all the time into two clusters with only financial stocks, the healthcare sector had almost all the time a cluster with its own stocks. The energy sector was another sector that had almost all the time a cluster with only its stocks with sometimes a few stocks of the materials sector. Other sectors like materials, information technology, consumer staples and telecommunications services had never a cluster with only their companies. The industrial, utilities and consumer discretionary were also grouped with other sectors most of the quarters with a few quarters that they had their own cluster.

B. *Case study I*

In the first case study to build the portfolio, it was used only the most invested stocks by all the hedge funds. Two groups of funds were used the first group is very known to the public because their managers are constantly in the Forbes lists of the richest billionaires, the second group is a less know group of funds used in one quarter by Whale Index. There are a few differences between these two groups of funds, the first one is that the investment made by the funds belonging to the billionaires is ten times bigger in the S&P 500 than the other group, the second is regarding the number of stocks. The number of stocks held by the billionaires is seven times bigger than in the other group, of all the stocks less than a quarter of the portfolio is from S&P 500 whereas in the group of less known funds one-third of their portfolio are stocks from the S&P 500. Which can lead us to the conclusion that the funds from the billionaires have a wider range of stocks held.

In Fig. 4 it's represented the lines of the returns of the portfolios. In blue, there are the lines correspondent to the funds of Whale Index and in green, there are the funds correspondent to the billionaire funds. Focusing now on the blue lines it's possible to see that no matter the size of the portfolio the returns will always be greater than the S&P 500 index. Three different portfolio sizes were tested: one with all the stocks the funds had in that quarter (line 6), the second using the size used by Whale Index which corresponds to 100 stocks (line 4) and the last is the most used portfolio size in the financial world which is twenty stocks (line 3). It was possible to see that the returns get better with fewer stocks than with more, this can lead us to the conclusion that the stocks chosen by these funds when they have more importance in the portfolio lead to better results, so they do smart choices.

Focusing now on the green lines it is possible to see that the results were not as good as the ones in the blue lines. For this type of funds, the best result was not the one with fewer stocks but the one 100 stocks although it was not even better than the S&P 500 index.

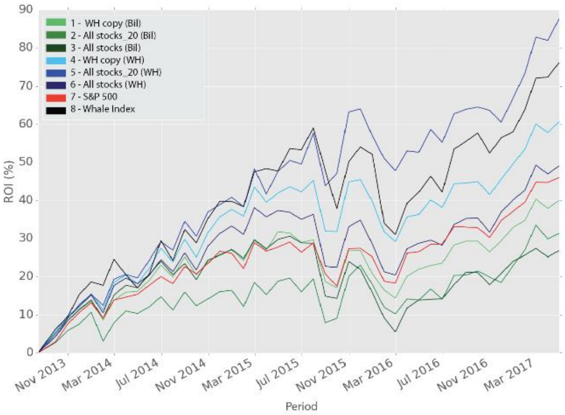

Fig. 4 - ROI obtained in case study I

	ROI (%)
Best result (Top 3 quarter (WH))	115.29
Worst result (All stocks (Bil))	26.46
S&P 500	44.6
Whale Index	76.29

Table 2 - ROI's summary case study I

C. *Case study II*

Some common investments strategies use the GICS system to base their investment. In this case study, it was used the stocks held by hedge funds and then divided them into their sectors to make the portfolio. To choose stocks from sectors two criteria were used: the first one is to choose the sector based on their ROI in the last quarter plus the 45 days it takes for the 13F forms to get available. The second criterion is to choose the stocks based on the amount of money invested by the funds in the sectors. It was constructed portfolios between 5 and 20 stocks. To make the graphics more readable, it was drawn only the lines of the returns of five sectors instead of the all the ten, the criterion to choose the lines was the one that obtained the biggest and smallest average ROI and the other ones were chosen based on the major differences between each other. It was done the simulations with the Whale Index funds and the billionaire ones but as it happened in the first case study the results of the billionaire funds were not as good as the Whale Index ones, so they are not presented.

1) Choose sectors based on the average return

It was studied the effect of choosing several number of sectors to portfolio composition. In this case study, it was possible to see that with this criterion to choose the sectors it's not linear that choosing more sectors will originate better returns. It was obtained better results when choosing only one sector compared to choosing two, three, four and five sectors but then when you start choosing six sectors and more the returns tend to increase again. To invest in a strategy like this it would be needed analysis to choose the best number of sectors to invest. The results obtained by the billionaire funds were worse than the ones obtained by the Whale Index funds, the billionaire ones didn't obtain a single result better than the S&P 500.

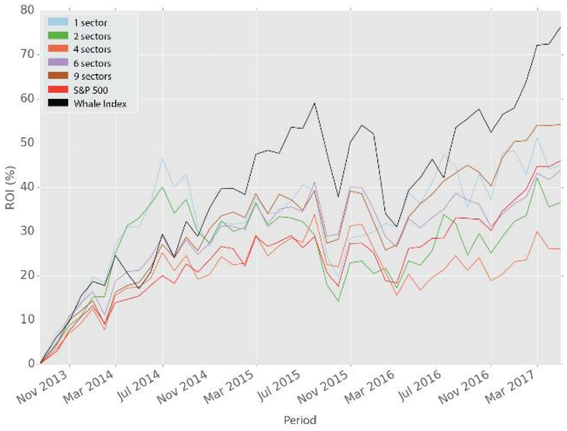

Fig. 5 - ROI for the Whale Index funds for the first experience case study II

2) Choose sectors based on the fund's investment

To use all the knowledge that hedge funds have in picking stocks this criterion was used. Comparing with the other criterion in this case study it's possible to see that the average results are clearly better with a criterion based on the knowledge of hedge funds. Choosing eight sectors presented the worst returns and choosing one sector was the best option. Analysing in detail the case where it was only chosen one sector it's possible to see that in all the simulated period it was only chosen the sector of information technology which it was clearly a smart choice, because this sector was clearly in expansion in the last few years. It was investigated if there is also a second sector that the hedge funds showed a clear preference and it was possible to identify the financial sector which when there is an option of choosing two sectors it was chosen ten times out of fifteen possible. Once again, the returns originated from the billionaire's hedge funds got worse returns than the ones of Whale Index even when the same sectors were chosen. This indicates that the choices made by them, even in similar conditions, are worse than the Whale Index funds.

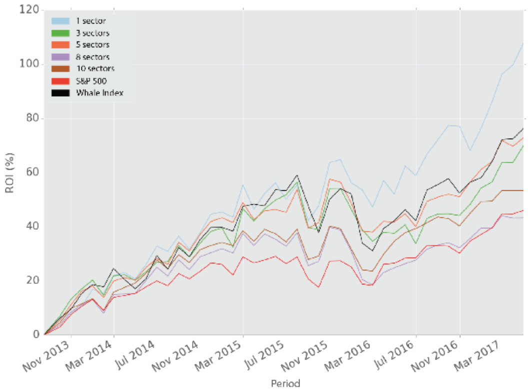

Fig. 6 - ROI for the Whale Index funds for the second experience case study II

		ROI (%)
Best result Whale Index funds (case 1 (1 sector))		118.54
Worst result Whale Index funds (case 2 (4 sectors))		27.6
Average case 1	Billionaires	32.57
	Whale Index	47.43
Average case 2	Billionaires	41.37
	Whale Index	71.89
S&P 500		44.6
Whale Index		76.29

Table 3 - ROI's summary case study II

D. Case study III

This case study is like case study II, however, the method used to divide the stocks is completely different. In this case study, it was used the method implemented using the agglomerative hierarchical clustering method based on the partial correlation coefficients, it was chosen partial correlation instead of Pearson correlation because with partial correlation it's possible to remove the influence of the index that drives up the correlation between stocks. This is an alternative method to divide the companies to increase the diversity of the portfolio because sometimes using the GICS system could be not enough. Once again in the returns graphics, it's only showed the results for five cases to the turn the graphic more legible.

1) Choose clusters based on their average return

The case with the best returns it was when eight clusters were chosen and the case with the worst return it was when one cluster was chosen. All the obtained returns no matter what number of clusters were chosen were better than S&P 500 index which is an excellent indicator and the results get better with the increase in the number of clusters chosen. Comparing this experience with the correspondent one in case study II it's possible to see that for all the cases using this method of stocks division the results obtained are better, not only the returns are bigger but also values such as minimum loss is lower. Once again, the results obtained by the Whale Index funds were better than the billionaire ones, however, there is something important that needs to be said. The results obtained by the billionaire funds in this experience obtained the best returns from all the cases simulated with this type of funds which can indicate that this method can be a good choice when there are a high number of stocks to be chosen because as it was seen in case study I some important differences between these types of funds are the number of stocks held and the quality in choosing them. So, when there are more stocks that are not so well chosen this method turned out a good alternative to invest.

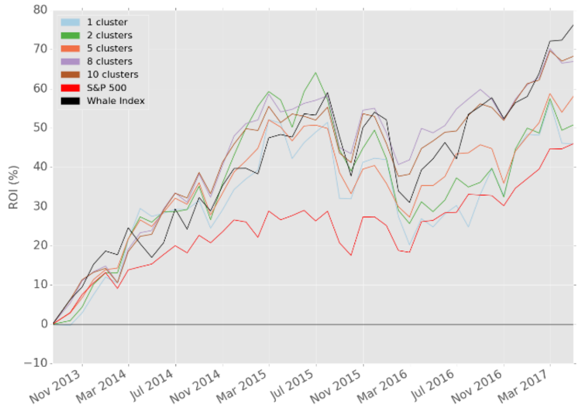

Fig. 7 - ROI for the Whale Index funds for the first experience case study III

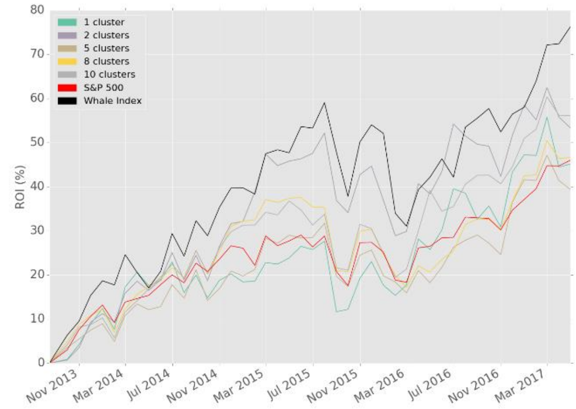

Fig. 8 - ROI for the billionaire's funds for the first experience case study III

2) *Choose clusters based on funds investment*

In this experience, as it had been used in the one in case study II the clusters were chosen based on the total investment made by the hedge funds. All the obtained results were higher than the S&P 500 ones and the best return was obtained when eight clusters were chosen, this one was even higher than the Whale Index itself, and the worst result was obtained when six clusters were chosen. As it can be seen in the results it's not linear that increasing the number of clusters chosen the returns will increase. Comparing this experience with the one that the criteria to choose the clusters is based on the return it's possible to see that what happened in case study II choosing clusters using the knowledge of hedge funds obtain better returns. Comparing this case with the similar case in the case study II we can conclude that the biggest advantage of the clustering method is obtained when more than seven clusters are chosen. When fewer clusters are chosen like one or two is worth to choose investing using the GICS method this can be explained because some funds base their investment in the GICS system so like this they choose certain sectors to invest which when the diving stock method is using clustering this fact doesn't happen. Once again, the results obtained by the billionaire's hedge funds were lower than the ones from whale index.

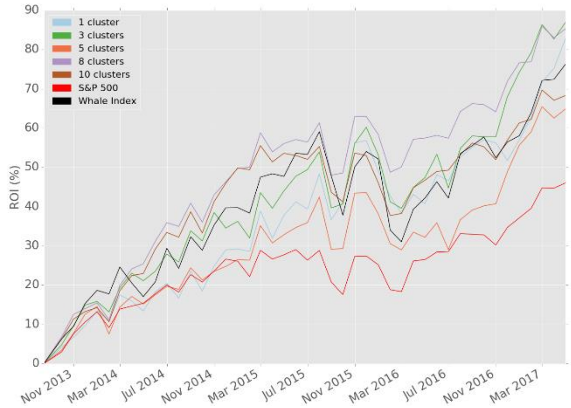

Fig. 9 - ROI for the Whale Index funds for the second experience case study III

		ROI (%)
Best result Whale Index funds (case 1 (8 clusters))		89.71
Worst result Whale Index funds (case 2 (1 cluster))		44.8
Average case 1	Billionaires	50.93
	Whale Index	59.87
Average case 2	Billionaires	39.86
	Whale Index	76.14
S&P 500		44.6
Whale Index		76.29

Table 4 - ROI's summary case study III

V. CONCLUSIONS

The analysis of the results obtained in this work allowed to understand a few things. Firstly, it's possible to build a profitable strategy using solely the publicly available information of hedge funds. However, the funds used need to be quality ones and their main strategy needs to be focused on S&P 500 stocks. If the funds used make smart choices when the investment strategy is based on the GICS sectors it's possible, using a criterion based on the funds, to identify the best sector to invest and have great results. The tested strategy using clustering and partial correlation originated better results than the S&P 500 and in some cases even better than the Whale Index, this strategy revealed more importance in the case of billionaire funds where the criterion to choose the clusters doesn't depend on the funds allowing to obtain higher returns than the S&P 500 which didn't happen without it, this can indicate that this is an interesting strategy to use when there is a lot a stocks to divide.

REFERENCES

[1] C. A. Ambler e J. E. Kristoff, “Introducing the North American Industry Classification System,” *Government Information Quarterly,* vol. 15, pp. 263-273, 1998.

[2] S. S. Jung e W. Chang, “Clustering stocks using partial correlation coefficients,” *Physica A,* vol. 462, pp. 410-420, 2016.

[3] SEC, "Edgar 2.0 database," [Online]. Available: https://www.sec.gov/edgar/searchedgar/companysearch.html. [Accessed March 2018].

[4] M. g. kendall, The advanced theory of statistics, Griffin, 1975.

[5] T. Q. G. C. L.-X. Z. ,. X.-R. W. Xing Li, “Market impact and structure dynamics of the Chinese stock market based on partial correlation analysis,” *Physica A,* vol. 471, pp. 106-113, 2017.

[6] D. Y. Kenett, X. Huang, I. Vodenska, S. Havlin e H. E. Stanley, “Partial correlation analysis: Applications for financial markets,” *Quantitative Finance,* vol. 15, pp. 569-578, 2015.

[7] M. T. A. M. G. G.-G. R. N. M. E. B.-J. Dror Y. Kenett, “Dominating Clasp of the Financial Sector Revealed by Partial Correlation Analysis of the Stock Market,” *PLoSone,* vol. 5, nº e15032, 2010.

[8] D. K. E. B.-J. Y. Shapira, “The Index cohesive effect on stock market correlations,” *The European Physical Journal B,* vol. 72, pp. 657-669, 2009.

[9] Y. H. Carol Hargreaves, “Prediction of Stock Performance Using Analytical Techniques,” *Journal of emerging technologies in web intelligence,* vol. 5, pp. 136-142, 2013.

[10] A. Gorgulho, R. Neves e N. Horta, “Applying a GA kernel on optimizing technical analysis rules for stock picking and portfolio composition,” *Expert Systems with Applications,* vol. 38, pp. 14072-14085, 2011.

[11] A. Silva, R. Neves e N. Horta, “A hybrid approach to portfolio composition based on fundamental and technical indicators,” *Expert Systems with Applications,* vol. 42, pp. 2036-2048, 2014.

[12] J. M. Pinto, R. F. Neves e N. Horta, “Boosting Trading Strategies performance using VIX indicator together with a dual-objective Evolutionary Computation optimizer,” *Expert Systems with Applications,* vol. 42, pp. 6699-6716, 2015.

[13] S. Nanda, B. Mahanty e M. Tiwari, “Clustering Indian stock market data for portfolio management,” *Expert Systems with Applications,* vol. 37, pp. 8793-8798, 2010.

[14] N. C. Long, N. Wisitpongphan, P. Meesad e H. Unger, “Clustering stock data for multiobjective portfolio optimization,” *International Journal of Computational Intelligence and Applications,* 2014.

[15] G. F. A. D. R. S. Carmela Iorio, “A P-spline based clustering approach for portfolio selection,” *Expert Systems With Applications,* vol. 95, pp. 88-103, 2018.

[16] R. C. G. Z. Huanhuan Yua, “A SVM Stock Selection Model within PCA,” *Procedia Computer Science,* vol. 31, pp. 406-412, 2014.

[17] PyData, "Pandas datareader," [Online]. Available: http://pandas-datareader.readthedocs.io/en/latest/. [Accessed January 2018].