IJAAF; Vol. 7, No. 3, Summer 2023, Serial 24

Iranian Journal of
Accounting, Auditing & Finance

Ferdowsi University of Mashhad

Quarterly

RESEARCH ARTICLE

Portfolio Diversification Based on Clustering Analysis

Marziyeh Nourahmadi
Department of Accounting and Finance, Faculty of Economics, Management and Accounting, Yazd University, Yazd, Iran
Hojjatollah Sadeqi*

Department of Accounting and Finance, Faculty of Humanities and Social Sciences, Yazd University, Yazd, Iran

How to cite this article:
Nourahmadi, M., & Sadeqi, H. (2023). Portfolio Diversification Based on Clustering Analysis. Iranian Journal of Accounting, Auditing and Finance, 7(3), 1-16. doi: 10.22067/ijaaf.2023.74812.1092
https://ijaaf.um.ac.ir/article_43078.html

ARTICLE INFO

Article History
Received: 2023-03-18
Accepted: 2023-06-01
Published online: 2023-07-14

Keywords:
Hierarchical Clustering, Portfolio Optimization, Portfolio Diversification, K-means.

Abstract

Forming an investment portfolio is one of the main concerns of managers and investors who strive in order to create the best investment portfolio to get the best return from the market. So far, many methods have been presented to construct a portfolio, of which the most famous is the Markowitz approach. Our research aims to offer a classical portfolio selection using cluster analysis. We trained four models using k-means clustering with daily log returns as features and agglomerative clustering methods with complete, single and average linkages based on correlation-based distances. Four equally weighted portfolios of 30 stocks each were formed by selecting the stock with the highest Sharpe ratio from each cluster. Based on the silhouette scores and Sharpe ratio, we selected agglomerative clustering with average linkage trained on last year's data as our final model. The performance of our selected portfolios over the test period was better than random selection in terms of Sharpe ratio but worse than the overall index. The results in terms of volatility showed better performance; our selected portfolio had an annualized volatility lower than the random selection and the average volatility of all clusters and relatively close to that of the equally weighted portfolio consisting of all 334 stocks in the data.

https://doi.org/10.22067/ijaaf.2023.43078.1092

NUMBER OF REFERENCES
36

NUMBER OF FIGURES
9

NUMBER OF TABLES
6

Homepage: https://ijaaf.um.ac.ir
E-Issn: 2717-4131
P-Issn: 2588-6142

*Corresponding Author: Hojjatollah Sadeqi
Email: sadeqi@yazd.ac.ir
Tel: 09124946469
ORCID: 0000-0001-5852-4198

1. Introduction

Data mining is introduced as the science of data analysis to gain insights and knowledge about the data under study. Researchers in most scientific fields, such as management, business, medicine, engineering and biology, face the rapid growth of information and high-dimensional data, so this method is used to try to understand the relationships between existing phenomena (Williams, 2011).

Clustering is one of the most critical data mining methods to extract useful information from high-dimensional data sets (Kumar and Wasan, 2010). In other words, clustering is a process in which a group of objects is clustered, so objects in one cluster are similar and different from objects in other clusters (Chaudhuri and Ghosh, 2016; Jain & Dubes, 1988). In recent years, different clustering methods have been proposed and developed that can be defined and designed as a mathematical technique to uncover the classification structures in data collection of real-world phenomena (Majewski et al., 2014).

One of the most critical investment issues facing different investors is choosing an optimal investment portfolio and balancing risk and return to maximise investment returns and minimise investment risk (Kolm et al., 2014). Markowitz first introduced the theory of portfolio analysis in “Portfolio Selection” (1952), which was used by investors and financial institutions for a long time (Pardalos et al., 1994). In the following years, some mathematical approaches have been used in financial decisions (Detemple, 2014).

Securities optimization is a significant financial problem, and the issue of choosing the optimal portfolio of stocks has long concerned investment professionals. One of the basic assumptions in finance is that due to a shortage of resources, all economic options are subject to some exchange. In deciding to invest, a rational investor faces the fundamental problem of choosing between the level of return they want to earn and the level of risk they are willing to accept for that return. A key step in the investment process is allocating one’s financial resources optimally (Bechis, 2020).

According to Gallup (Jones, 2017), the average percentage of Americans owning stocks from 2009 to 2017 is 54%, which has decreased by 8% since the financial crisis. Many people have withdrawn from the stock market because they are not getting ideal returns or because there is no high-yield investment strategy to generate reasonable profits while withstanding the market's volatility. This problem affects not only retail investors but also many institutional investors. Modern portfolio theory suggests that investors can achieve this goal through portfolio diversification by reducing risk by spreading a portfolio across many different investments (Bodie et al., 2014). Portfolio diversification allows investors to avoid over-exposure to a single source of risk; an investor with a well-diversified portfolio can be immune to many of the company's risks (Hull, 2018). This project aims to use clustering methods to construct well-diversified portfolios that reduce volatility and losses and increase capital preservation.

In the first part of the paper, portfolios are constructed using different clustering methods for 334 stocks and the stock with the highest Sharpe ratio is selected from each cluster. For k-means clustering, we use daily log returns as features. In contrast, we use correlations between stocks for agglomerative clustering as a user-defined distance metric, as suggested by several previous studies (León et al., 2017). Using the Sharpe ratio and silhouette scores, we select our final model to test with the test data. In the second part of the portfolio, we evaluate and compare the test results based on the Sharpe ratio with the overall index and portfolios of randomly selected stocks as benchmarks. We also calculate the annualized volatility of the portfolio and compare it to that of each cluster and the overall index to evaluate our model. We conclude our study by interpreting the results of the first and second parts and suggesting several improvements.

2. Literature Review

Clustering is one of the most critical tasks in data mining and one of the unsupervised learning

models. This method aims to naturally group a set of objects and data into different sections, and then the quantitative comparison of the features of each section allows the discovery and investigation of hidden structures in the data (Jain, 2010). The clustering of time series data is commonly used to discover patterns in time series datasets (Wang et al., 2002).

This task itself is divided into two separate sections. The first part is to find patterns that occur frequently in time series (Chung et al., 2001; Chiu et al., 2003), and the second part is to methods that examine patterns that occur infrequently in time series.

It also explores events that have surprising effects on the time series process (Keogh et al., 2002; Leng et al., 2009). Clustering categorizes data by reducing the volume of data and finding patterns. The general approaches of clustering algorithms are shown in Figure 1.

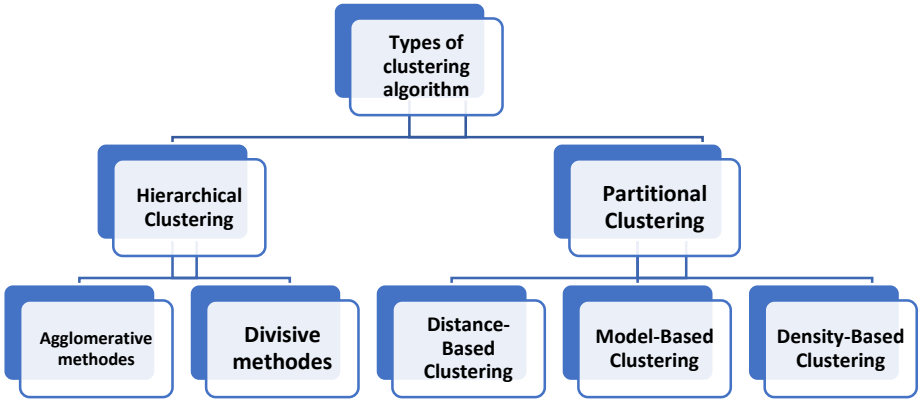

Figure 1. Types of clustering algorithms
Source: (Saxena et al., 2017)

2.1 Types of clustering algorithms (Saxena et al., 2017)

Clustering is divided into two categories: partial and hierarchical, which are defined and categorized below:

2.1.1 Partial clustering (segmentation):

They divide datasets into non-overlapping subsets so that each piece of information is contained in exactly one subset.

Hierarchical clustering is divided into two categories:

2.1.2.1 Agglomerative methods (bottom to top method)

Starts with each dataset in a cluster. Repeatedly, it combines clusters close to each other at each stage to remain a cluster finally.

2.1.2.2 Divisible methods (top to the bottomed method)

Starts the entire data as one cluster. Repeatedly splits the data into one of the clusters until there is only one dataset per cluster.

In this study, k-means clustering and hierarchical clustering methods are used, both of which are

explained in detail.

2.1.2 Clustering techniques

There are many clustering techniques, depending on the strategy and identification categories. The choice of which technique to use depends on the type and structure of the data. In this section, two clustering methods are discussed:

2.1.2.1 K-means clustering

K-Means is JB's best known clustering method. MacQueen proposed this method in 1967 as a classical clustering algorithm for scientific research and industrial applications. The k-means algorithm aims to find and group data points in similar classes, where this similarity is perceived as the opposite of the distance between the data. The closer the data points are to each other, the more likely they are to belong to a cluster. The basic idea of this algorithm is to divide n^{th} data objects into n^{th} clusters such that the sum of the squares of the data points in each cluster is the smallest distance from the center of the cluster (Thuraisingham and Ceruti, 2000).

The algorithm finds the center of "k" and assigns each data point to exactly one cluster to minimize the variance within the cluster (called inertia). This method usually uses Euclidean distance (the typical distance between two points), but other distance criteria can be used. The k-means algorithm provides a local optimum for a given K and proceeds as follows:

1. This algorithm determines the number of clusters.
2. The data points are randomly selected as cluster centers.
3. Each data point is assigned to the cluster center closest to it.
4. The cluster centers are updated to the average assignment.
5. Steps 3 and 4 are repeated until all cluster centers remain unchanged.

2.1.2.2 Hierarchical clustering

Hierarchical clustering creates clusters that have a dominant order from top to bottom. The main advantage of hierarchical clustering is that the number of clusters does not need to be determined. The model itself determines them and solves this problem. This clustering method is divided into two types: agglomerative hierarchical clustering and divisive hierarchical clustering.

Agglomerative hierarchical clustering is the most common type used to group objects based on similarity. This is a bottom-up approach where each observation starts in its own cluster, and cluster pairs are merged as they move up the hierarchy. The agglomerative hierarchical clustering algorithm provides a *local optimum* that works as follows:

1. Think of each data point as a one-point cluster, starting at N.
2. Consider two data points closer together and group them into N-1 clusters.
3. Consider two clusters close to each other and combine them into N-2 clusters.
4. Repeat step 3 to stay with only one cluster.

Divisive hierarchical clustering works "top-down" and separates the remaining clusters to form distinct subgroups of each. Both methods create the N-1 hierarchical level and facilitate clustering at the level that best divides the data into homogeneous groups.

Hierarchical clustering allows the drawing of dendrograms, an image of a binary hierarchical clustering. A dendrogram is a tree diagram that shows hierarchical relationships between different data sets. Dendrograms provide an exciting and informative representation of the results of hierarchical clustering that includes the memory of the hierarchical clustering algorithm, making it possible to express the formation of clusters simply by looking at the diagram.

One of the advantages of hierarchical clustering is that it is easy to implement, the number of

clusters does not need to be fixed, and the dendrograms generated are very useful for understanding the data. However, the time complexity of hierarchical clustering can lead to longer computation times than other algorithms, such as K-Means. For a large dataset, it is difficult to determine the correct number of clusters by observing the dendrogram. Hierarchical clustering is very sensitive to outliers in the data, which significantly affects the model's performance (Tatsat et al., 2020).

The agglomerative method is one of the hierarchical clustering algorithms. It classifies objects by collecting small clusters from the bottom up in a tree structure. The clustering process starts by declaring each point as its cluster, and then the two most similar clusters are merged into a single cluster according to their linkage. The termination criterion in scikit-learn is the number of clusters entered, so the above process is repeated until the specified number of clusters is left. Unlike k-means, Agglomerative Clustering in scikit-learn allows a user-defined distance metric. Therefore, a correlation-based distance metric can be used. The custom distance metric we use is as follows:
$distance$=(1−$correlation$)

There are four different linkage criteria to determine how similarities between two clusters are measured: single, complete, average, and station. The two with the least minimum distance between their points are merged in a single linkage. The two clusters with the least maximum distance between their points are merged in the complete linkage. The two clusters with the least average distance between all their points are merged in average linkage. Ward linkage, the default setting in scikit-learn, merges two clusters so that all clusters' variance increases the least. Since Ward only allows the Euclidean distance metric, it is omitted for our purpose of using correlation distances.

2.2 Sharpe ratio

The Sharpe ratio, widely used to evaluate portfolio performance, is used in this project to examine portfolio performance. It measures how much a portfolio outperforms the risk-free return on a risk-adjusted basis. The higher the ratio, the better the performance. The Sharpe ratio is calculated using the following formula:

$$Sharp\ Ratio(SR) = \frac{R_p - R_f}{\sigma_p} \qquad (1)$$

Clustering is one of the data mining techniques that group data based on a similarity criterion without knowing the number and characteristics of the groups. Clustering based on the similarity of trends can be very useful in evaluating the common movement of prices. So far, various research works have been conducted in the field of clustering and studying the correlation or convergence between stocks on a stock exchange, the overall index or the index for a particular industry on the stock exchanges of different countries or the index of different industries on a stock exchange, which are discussed and presented below.

Raffinot (2017) proposes an asset allocation method based on hierarchical clustering using network theory and machine learning techniques. His experimental results show that the hierarchical clustering based portfolio is stable, truly diversified, and performs better risk adjustment than traditional optimization techniques (Raffinot, 2017).

In a paper by Ding et al. (2019), the CSI800 index was clustered using the stock K-Means benchmark. In this study, hierarchical clustering diagrams and similarity structure diagrams were drawn and analyzed, and it was found that clustering approaches in stock analysis have visual characteristics and ease of analysis (Ding et al., 2019).

In their research, Nakagawa et al. (2019) used the pattern of stock price fluctuation, which is not yet fully used in the financial market, as an input feature for prediction. They extracted the representative stock price fluctuation patterns with k- Medoids Clustering with the Indexing DTW

method (Nakagawa et al., 2019).

In their study, Huarng et al. (2008) investigated the structural changes using the K-Means clustering method to analyze a time series in the capitalization-weighted stock index of the Taiwan Stock Exchange. This study also illustrates the advantages of using the cluster method to determine structural changes (Huarng et al., 2008).

Liao et al. (2008) investigated a two-stage data mining method to summarize and visualize the data of the Taiwan Stock Market. In the first stage, a series of methods were used to illustrate the patterns and rules to suggest stock categories. In the second stage, K-Means clustering was implemented to identify stock category clusters and provide helpful information to investors (Liao et al., 2008).

Prior studies (Barziy and Chlebus, 2020; Snow, 2020; Molyboga, 2020; Jaeger et al., 2021; Nourahmadi and Sadeqi, 2021) used the HRP approach. (Lohre et al., 2020) In their paper, they examine diversification strategies based on hierarchical clustering. (Raffinot, 2018) Their results show that HERC portfolios based on descending risk criteria perform statistically better than CDaR criteria for risk adjustment.

3. Research Methodology

This paper aims to use clustering algorithms to create a diversified portfolio to reduce volatility and the overall risk of an investment portfolio. The first step in data preparation is data mining. For this purpose, we extracted the adjusted daily data of all listed companies from 01/01/2017 to 07/30/2020 (about 660 stocks) using the Noavarn Amin software. All calculations are performed by Python version 3.8. The second step in data preparation is data preprocessing. In this stage, we first need to clean the data from noise, outdated data and missing data that compromise the quality of the data. First, the number of trading days was calculated for all stocks, and based on the number of trading days, the remaining 334 stocks and the rest were removed from the statistical population because they did not have enough data.

This study used the adjusted final price as the primary variable for clustering. It starts with the P_{it} raw price series, which shows the stock price of company *i* on day *t*, and Pit-1 also shows the stock price of the company *i* on day *t-1*. Then the log returns of the companies' stocks are calculated according to Equation 2.

$$R_{it} = ln\frac{P_{it}}{P_{it-1}} \quad (2)$$

In order to determine the degree of similarity between two time series, it is necessary to detect the extent to which time series A can explain time series B. This value is determined using the following equation, known as the correlation coefficient, where the original diameter is one, and the other elements indicate their correlation coefficient:

$$\rho_{AB} = \frac{\mathrm{Cov(A,B)}}{[\mathrm{Var(A)Var(B)}]^{\frac{1}{2}}} \quad (3)$$

Equation 4 is used to convert the correlation coefficient into a metric criterion:

$$\mathrm{Dist}_{\rho}(\mathrm{A,B}) = \sqrt{2(1-\rho_{\mathrm{AB}})} \quad (4)$$

After processing the data, we try to achieve order in the data in the model learning phase. As mentioned earlier in this study, we want to use the clustering method as an unsupervised learning method and use it to optimize the portfolio. We divide the period into two parts: the testing and training periods. The training period, 02/07/2019 to 01/01/2017, covers 768 days, and the testing

period, 02/08/2019 to 07/30/2020, covers 539 days. We use the overall index as the criterion for comparing the results.

Figure 2 shows the returns of all stocks.

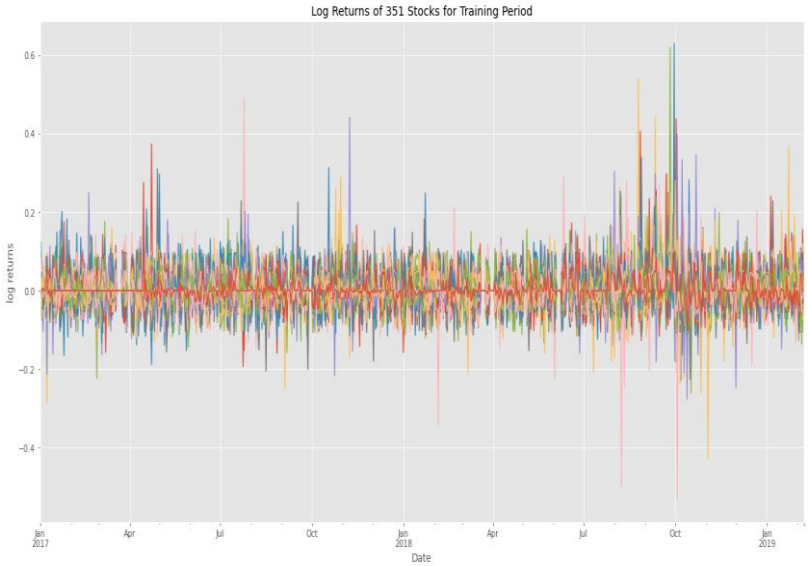

Figure 2. Return of stocks
Source: Research findings.

The literature review states that the portfolio selection problem can be solved more efficiently by grouping stocks into clusters and then selecting stocks in clusters to form efficient portfolios (Gubu et al., 2019). The general framework in this paper is shown in Figure 3.

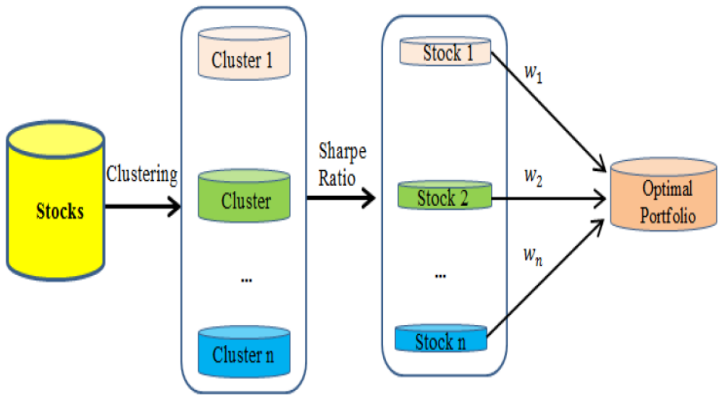

Figure 3. General steps of portfolio selection
Source:(Gubu et al., 2019)

In order to assess how well stocks are clustered, we use the Sharpe ratio, silhouette scores, and annualized volatility. A higher Sharpe ratio indicates a better portfolio risk-adjusted return, suggesting the portfolio is well diversified. We compare the Sharpe ratio in the created portfolios and the benchmarks. Since we are simply comparing different models over the same period, we used the daily Sharpe ratio: the mean of the log returns over the period divided by the standard deviation of

the log returns over the period. The result is successful if our portfolio has a higher Sharpe ratio than the benchmarks. The silhouette score measures how similar an asset is to its own cluster compared to other clusters. It ranges from -1 to +1, with a higher value indicating that the object is a good match to its cluster and a poor match to neighboring clusters. If most objects have a high value, the cluster configuration is appropriate. If any items have a low or negative value, there may be too many clusters in the model.

Finally, annualized portfolio volatility is calculated for each cluster to assess diversification:

$$vol = \sqrt{w^T \sum w} \quad (5)$$

Where $\sum$ is the covariance matrix of returns, w and w^T are the portfolio weights and their transpose. If our clusters are well constructed and we group similar stocks in each cluster, the volatility in each cluster should be higher than the portfolio's volatility. We also compare the volatility of the portfolio to the volatility of benchmarks: a portfolio of 30 randomly selected stocks and an equally weighted portfolio consisting of all 334 stocks in the data.

Two benchmarks indicate the market performance used in this project. The first is the overall index.

To construct this benchmark portfolio, 30 stocks are randomly selected based on a uniform distribution. However, instead of simply drawing 30 stocks at random once, we constructed 30 such portfolios and calculated the average Sharpe ratio of the 30 portfolios to control for variability. The same benchmark is also used to evaluate volatility.

We calculate the Sharpe ratio for all stocks and the overall index in the first step. The Sharpe ratio for the overall index is 0.399126. Then we randomly select 30 stocks from the data and calculate the Sharpe portfolio ratio, which consists of 30 randomly selected stocks and is 0.33981.

4. Result

We will create a portfolio using the k-means method in the next step. First, we cluster 334 stocks using the k-means method. We set the number of clusters to 30 and then performed the clustering. We use daily stock prices as a feature. In the next step, based on the Sharpe criterion, we select the stock from the cluster with the highest Sharpe criterion and select it as the selected stock to form the portfolio. The Sharpe criterion for a portfolio consisting of the k-means method is 0.32347.

4.1. The correlation matrix

We will use the correlation distance between the price histories of each stock for hierarchical clustering, so let us create our correlation matrix and visualize how our clustering results are formed. We assume that correlation distance works better than daily prices because it can account for price and movement similarities. The heat map does not show the 30 clusters we construct but visualizes the results of our practice in a more general sense.

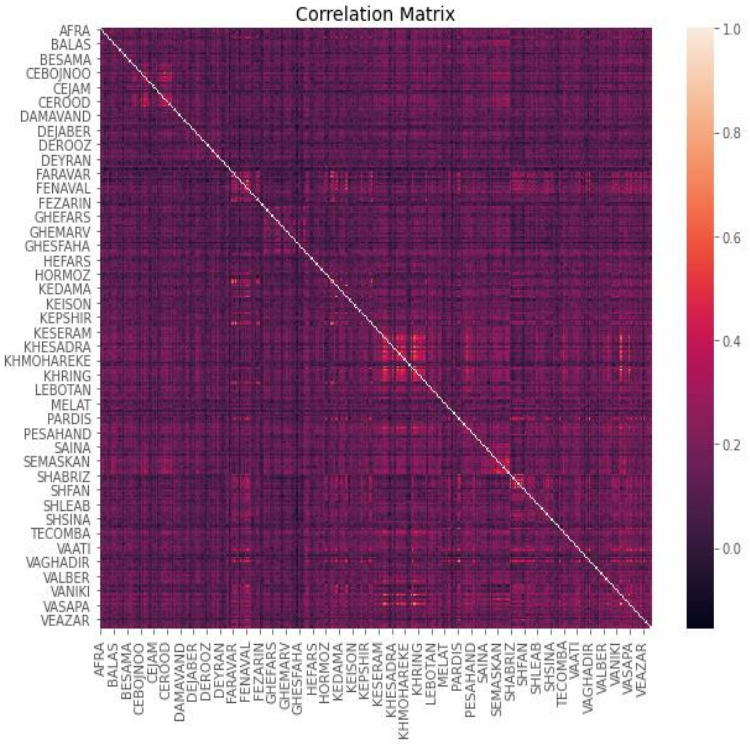

Figure 4. Heat Map (Correlation matrix)
Source: Research findings.

Figure 5 shows the Heat Map of Stocks Clustered by Correlation Distances, Average Linkage in Training Period 1.

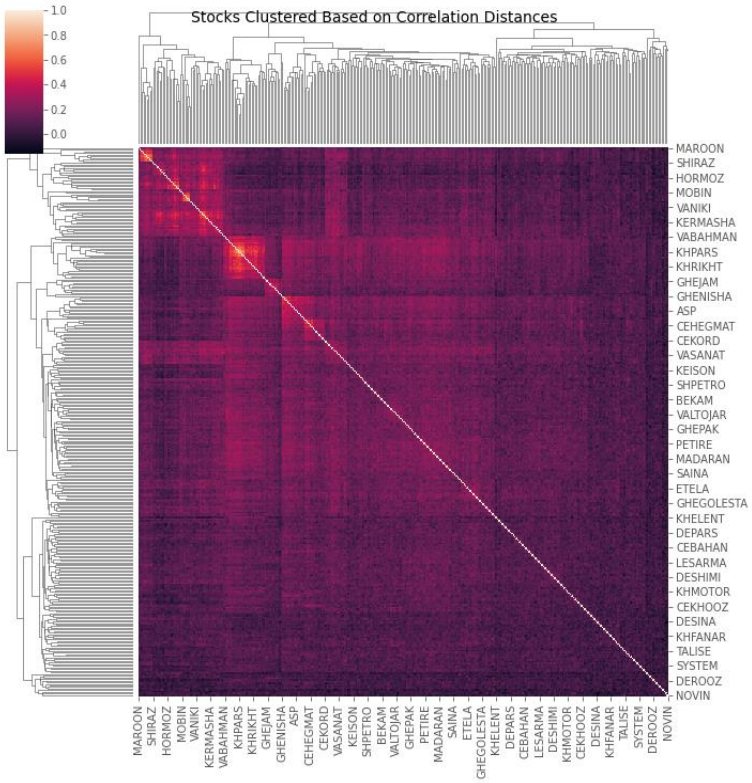

Figure 5. Heat Map (Stocks clustered by correlation distances, Average linkage, training period 1)
Source: Research findings.

Table 1 Shows 30 stocks selected from 334 stocks using various clustering methods.

Table 1. Hierarchical (Agglomerative) Clustering and k-means

Average		Complete		Linkage		K-means	
Sharpe	**Group**	**Sharpe**	**Group**	**Sharpe**	**Group**	**Sharpe**	**Group**
FEOLAD	0	FEROS	0	FEOLAD	0	SHKARBON	0
GHESALEM	1	CEKHAF	1	SEDABIR	1	FEPANTA	1
REANFOR	2	MOBIN	2	VEAZAR	2	SHSINA	2
PAKSHOO	3	GHEPIRA	3	PAKSHOO	3	GHEJAM	3
KEKHAK	4	DAMAVAND	4	NOVIN	4	KHRING	4
KHODKAFA	5	KEMANGANEZ	5	VAETEBAR	5	KEHRAM	5
ENERGY	6	VABOALI	6	KEMARJAN	6	KAZERO	6
VAETEBAR	7	GHEHEKMAT	7	FEKHAS	7	SHPARS	7
CEBOJNOO	8	SHPAKSA	8	GHEMAHRA	8	SHZANG	8
VADEY	9	KEKHAK	9	KEDAMA	9	CEKHAF	9
KEMARJAN	10	CEFANO	10	GHESALEM	10	FENAVAL	10
DESINA	11	GHEFARS	11	KOSAR	11	CEBAGHER	11
HEFARS	12	CHEKAREN	12	KEHRAM	12	SENOSA	12
CHEKAREN	13	VAETEBAR	13	ETEKAM	13	LEKEMA	13
CEFANO	14	PAKSHOO	14	HEFARS	14	CEKHASH	14
GHESHESFA	15	FLAMI	15	CEBOJNOO	15	FEOLAD	15
NOVIN	16	GHEMAHRA	16	BEMAPNA	16	BESAMA	16
KOSAR	17	HAMRAH	17	FEPANTA	17	SEPARDIS	17
VAKAR	18	VADEY	18	CHEFIBER	18	NIROO	18
SYSTEM	19	ENERGY	19	DEFRA	19	VAETEBAR	19
GHEMAHRA	20	SHKARBON	20	DETMAD	20	PELOLEH	20
KHNASIR	21	FEOLAD	21	DERAZAK	21	FARAVAR	21
SEDABIR	22	BESAMA	22	FLAMI	22	KEGHAZVI	22
FEKHAS	23	KESERAM	23	FAJR	23	KHELENT	23
DAROO	24	KEMARJAN	24	KAZERO	24	KHMOTOR	24
FAJR	25	DESINA	25	CEKHASH	25	SHETRAN	25
FLAMI	26	KEHAMEDA	26	FESEPA	26	BEKAB	26
VEAZAR	27	FEKHAS	27	DEROOZ	27	KHAZIN	27
KHELENT	28	GHESALEM	28	CEHORMOZ	28	VALTOJAR	28
CEKERMA	29	NOVIN	29	GHEFARS	29	PAKHSH	29

In Table 2, we compare Sharpe ratios of different clustering methods.

Table 2. Sharpe Ratio (Training Period 1)

	K-Means	**Single**	**Complete**	**Average**	**Total index**	**Random**
Sharpe ratio	0.323	0.347	0.305	0.328	0.399	0.339

Source: Research findings

The silhouette value measures how similar a point is to its cluster (*cohesion*) compared to other clusters (*separation*). The range of the silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. We may have created too many clusters if many dots have a negative silhouette value (Tatsat et al., 2020).

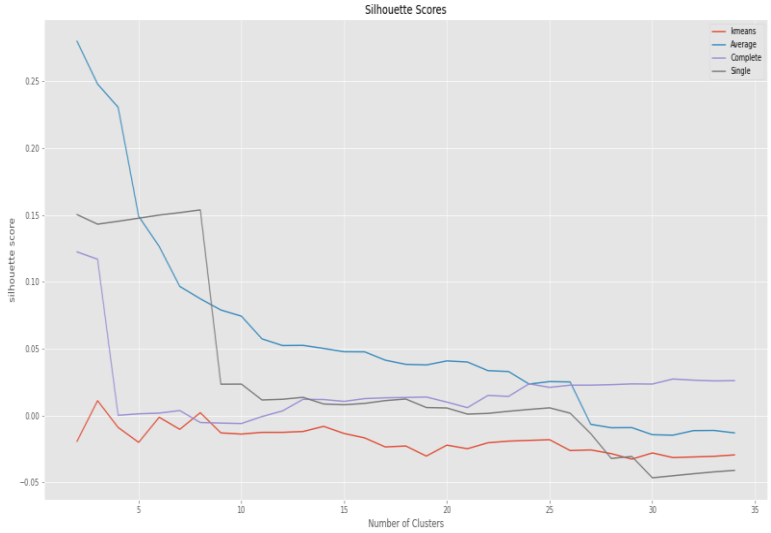

Figure 6. Silhouette Scores (Training Period 1)
Source: Research findings.

The above results do not seem consistent with our expectations or the silhouette values. Contrary to our expectations, k-means clustering has the highest Sharpe ratio, followed by single linkage. The silhouette scores indicate the worst performance for average linkage. This may be because we add noise using training data that goes back too long. Stock prices from 4 years ago do not necessarily match today's prices. Let us try a shorter training period and see if it works better.

Below we present the results from Training Period 2 (07/07/2019 to 07/07/2020).

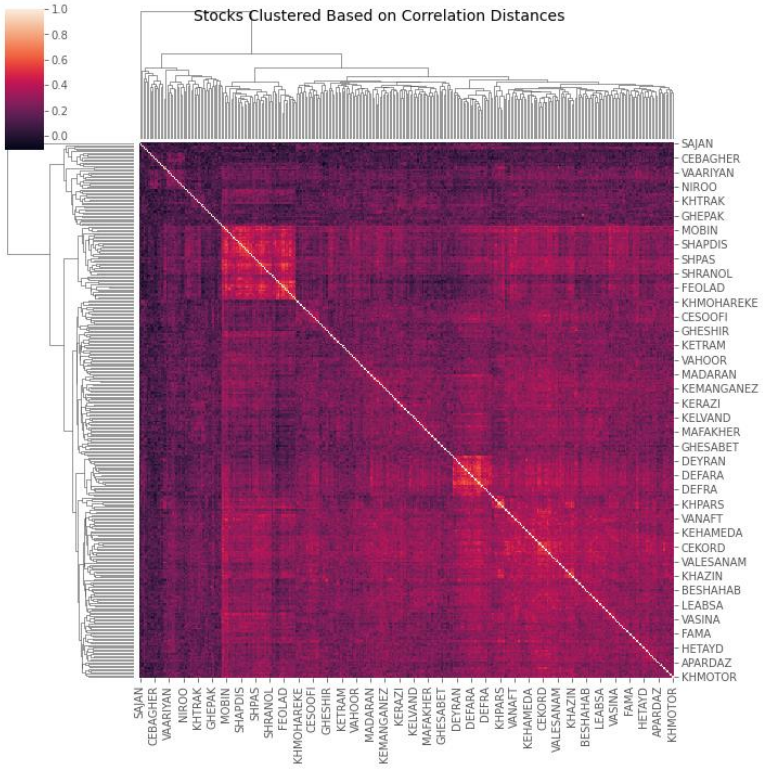

Figure 7. Heat Map (Stocks Clustered by Correlation Distances, Average Linkage, Training Period 2)
Source: Research findings.

Table 3. Hierarchical (Agglomerative) Clustering and Kmeans

Average		Complete		Single		K-means	
Sharpe	**Group**	**Sharpe**	**Group**	**Sharpe**	**Group**	**Sharpe**	**Group**
FEOLAD	0	REANFOR	0	FEOLAD	0	SHFARS	0
KHETOGHA	1	PEDERAKHSH	1	VEAZAR	1	HAMRAH	1
SHNAFT	2	FEROS	2	VAKADO	2	DAMAVAND	2
GHESALEM	3	CEKHAF	3	SEGHAZVI	3	SHKARBON	3
KEMARJAN	4	KEKHAK	4	TEPCO	4	KESERAM	4
GHESABET	5	FARAVAR	5	SHSINA	5	GHEHEKMAT	5
CEBAGHER	6	SHNAFT	6	TEPUMPI	6	DEABOR	6
VAKADO	7	GHESALEM	7	KHODRO	7	ZAGROS	7
FLAMI	8	FLAMI	8	KHTRAK	8	SHPAKSA	8
VEAZAR	9	SHTOOKA	9	NIROO	9	VATOOSHE	9
TEPUMPI	10	KHFANAR	10	SAJAN	10	PARDIS	10
TEPCO	11	KEFRA	11	KESAVEH	11	KHETOGHA	11
KHFANAR	12	KESERAM	12	KHODKAFA	12	FEROS	12
LEKEMA	13	HAMRAH	13	PELOLEH	13	TIPIKO	13
CEKHAF	14	TEPCO	14	GHEFARS	14	CEOROUM	14
DAMAVAND	15	VEAZAR	15	ENERGY	15	VATOSAM	15
SAJAN	16	LEKEMA	16	GHESHESFA	16	CABZEVA	16
GHEGOL	17	GHEHEKMAT	17	GHEGOL	17	DEAMIN	17
VABAHMAN	18	KHELENT	18	GHEMAHRA	18	FEOLAD	18
GHEFARS	19	CEBAGHER	19	CHEKAREN	19	KELVAND	19
KHODRO	20	ENERGY	20	KAZERO	20	KEAMA	20
ENERGY	21	SHAMLA	21	KEFPARS	21	GHSHEKAR	21
KEDAMA	22	DAMAVAND	22	GHEJAM	22	RETAP	22
KEFPARS	23	SAJAN	23	FEPANTA	23	ENERGY	23
KHODKAFA	24	SHKARBON	24	GHESALEM	24	GHEPINO	24
FEKHAS	25	LEKHAZAR	25	CEBAGHER	25	KHODRO	25
KESAVEH	26	SEGHAZVI	26	KHFANAVAR	26	CEBAGHER	26
PELOLEH	27	KHODRO	27	KEHRAM	27	GHESALEM	27
CHEKAREN	28	FEOLAD	28	KESERAM	28	KEMASEH	28
KELVAND	29	VABAHMAN	29	CEJAM	29	BOURSE	29

Table 5 compares the results of the Sharpe ratios of different clustering methods.

Table 5. Sharpe Ratio (Training Period 2)

	K-Means	**Single**	**Complete**	**Average**	**Total index**	**Random**
Sharpe ratio	0.319	0.334	0.328	0.341	0.399	0.339

Source: Research findings.

Table 5 shows the results for the second training period. The portfolios constructed by k-means clustering and agglomerative methods with single, complete and average linkages have Sharpe ratios of about 0.319, 0.334, 0.328 and 0.341, respectively.

Figure 8 illustrates the silhouette scores in the second training period. The results show that the single-linkage model performs worst at 30 clusters, consistent with the Sharpe ratio results. The results in the second training period are much more consistent with expectations and silhouette scores, suggesting that better clusters are formed.

Considering that our data consists of 334 firms, we can claim that the volatility of our portfolio is satisfactory. Our model also performs better than random selection in terms of Sharpe ratio and volatility.

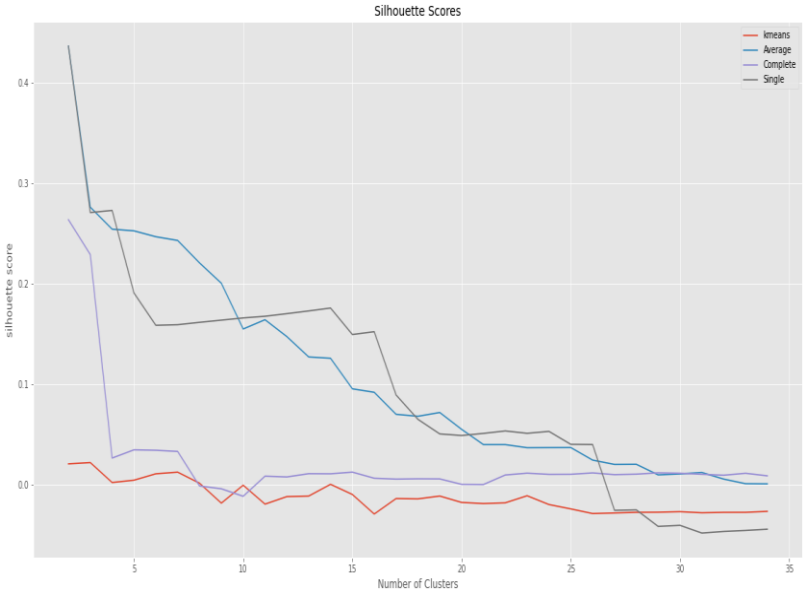

Figure 8. Silhouette Scores (Training Period 2)
Source: Research findings.

Table 6. The Volatility

Volatility	
Total index	0.200
Portfolio	0.168
Random	0.210
Cluster	0.383

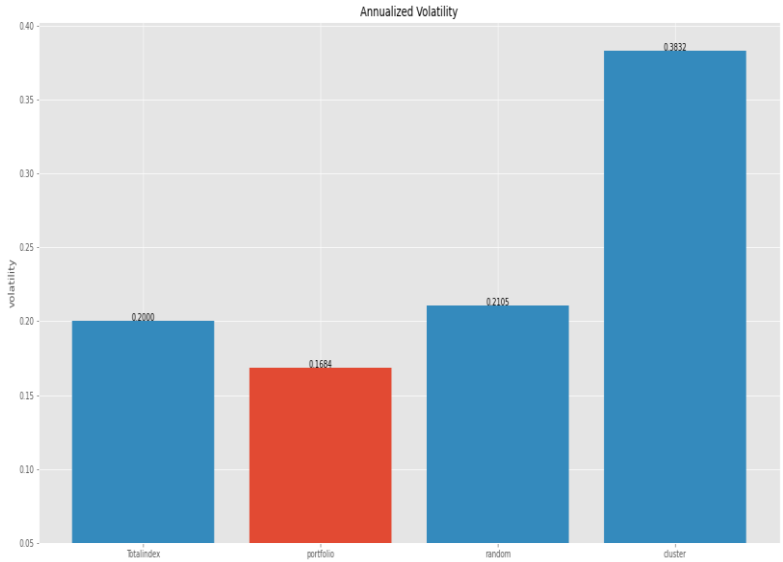

Figure 9. Volatility
Source: Research findings.

5. Conclusion and Recommendations

Data mining is one of the most powerful tools for extracting information and knowledge from raw data, and clustering, as one of the standard methods in data mining, is a suitable method for grouping

data into different clusters, which helps in understanding and analyzing relationships. In general, clustering is one of the data mining methods in which similar data is classified into related or homogeneous groups (Rai and Singh, 2010). Investors who intend to buy and add new stocks to their portfolio or investors who want to construct an optimal portfolio must first pay attention to the degree of mobility or, in other words, the correlation between different stocks because this reduction of investor risk is risk averse and increase investor return is risky so that they can use clustering methods. One of the most critical investment issues facing different investors is choosing an optimal investment portfolio and balancing risk and return to maximise investment returns and minimize the investment risk (Kolm et al., 2014). Thus, this project aims to create a well-diversified portfolio using clustering. We trained four models using k-means clustering with daily log returns as features and agglomerative clustering with average, full, and single linkages based on correlation-based distances. Four equally weighted portfolios of 30 stocks each were formed by selecting the stock with the highest Sharpe ratio from each cluster. Based on the silhouette scores and Sharpe ratio, we selected agglomerative clustering with average linkage trained on last year's data as our final model. The performance of our selected portfolio over the test period was better than random selection in terms of Sharpe ratio but worse than the overall index. The results in terms of volatility showed better performance; our selected portfolio had an annualized volatility lower than the random selection and the average volatility of all clusters and relatively close to that of an equally weighted portfolio consisting of all 334 stocks in the data.

There are a few ways to improve the performance of our model potentially. First, we could further adjust the length of the training period. We could try to adjust the number of clusters since we obtained higher silhouette values with smaller clusters.

Another way to ensure well-constructed clusters is to use a distance threshold. Distance thresholds define the maximum distance within a cluster such that the components of a cluster are "similar" enough. Finally, we could improve the portfolio's risk-adjusted return by weighting the individual stocks in the portfolio based on an optimization problem with maximizing the Sharpe ratio as a constraint.

6. The implications

In terms of performance evaluation, we could add additional means of portfolio evaluation such as Sortino ratio, Calmer, Max Drow Down, Omega ratio, VaR and CVaR to draw more concrete conclusions. We can also construct another benchmark by selecting three stocks with the highest Sharpe ratio in several industry sectors.

In addition, we can perform another qualitative analysis using the industry data to understand the commonalities within each cluster by examining whether the stocks in the typical clusters are similar in terms of sectors. We can also use time series analysis to build a model that describes the changes in stock prices over the entire period.

References

1. Majewski, S., Majewska, A., and Nermend, K. A. (2014). Comparison of k-means and Fuzzy c-means Clustering Methods for a Sample of Gulf Cooperation Council Stock Markets. *Folia Oeconomica Stetinensia* 14(2), pp. 19.39. http://dx.doi.org/10.1515/foli-2015-0001
2. Barziy, I., and Chlebus, M. (2020). HRP performance comparison in portfolio optimization under various codependence and distance metrics. Working papers, Warszawa, Poland.
3. Bechis, L. (2020). Machine learning portfolio optimization: hierarchical risk parity and modern portfolio theory. LUISS Guido Carli, Roma, Italy.
4. Bodie, Z., Kane, A., and Marcus, A. (2014). Investments, Edition 10, London, McGraw-Hill

Education-Europe. Hentet Seember. https://www.amazon.com/Investments-10th-Zvi-Bodie/dp/0077861671

5. Chaudhuri, T. D., and Ghosh, I. (2016). Using clustering method to understand Indian stock market volatility, *Communications on Applied Electronics*, 2(6), pp. 35-44, https://doi.org/10.5120/cae2015651793
6. Chiu, B., Keogh, E., and Lonardi, S. (2003). Probabilistic discovery of time series motifs. *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 493–498. https://doi.org/10.1145/956750.956808
7. Chung, F. L. K., Fu, T. C., Luk, W. P. R., and Ng, V. T. Y. (2001). Flexible time series pattern matching based on perceptually important points. *In Workshop on Learning from Temporal and Spatial Data in International Joint Conference on Artificial Intelligence.* pp. 1-7.
8. Detemple, J. (2014). Portfolio selection: a review. *Journal of Optimization Theory and Applications*, 161(1), pp. 1–21. https://doi.org/10.1007/s10957-012-0208-1
9. Ding, B., Li, L., Zhu, Y., Liu, H., Bao, J., and Yang, Z. (2019). Research on Comprehensive Analysis Method of Stock KDJ Index based on K-means Clustering. *3rd International Conference on Mechatronics Engineering and Information Technology*, pp. 484–491. https://doi.org/10.2991/icmeit-19.2019.78
10. Gubu, L., Rosadi, D., and Abdurakhman. (2019). Classical portfolio selection with cluster analysis: Comparison between hierarchical complete linkage and ward algorithm. *AIP Conference Proceedings*, 2192(1), https://doi.org/10.1063/1.5139174
11. Huarng, K.-H., Yu, T. H.-K., and Kao, T.-T. (2008). Analyzing structural changes using clustering techniques. *International Journal of Innovative Computing, Information and Control*, 4(5), pp. 1195–1202.
12. Hull, J. C. (2018). Options, Futures and Other Derivatives, 10e. Aufl., New York.
13. Jaeger, M., Krügel, S., Papenbrock, J., and Schwendner, P. (2021). 'Adaptive Seriational Risk Parity'and other Extensions for Heuristic Portfolio Construction using Machine Learning and Graph Theory. Working paper, https://doi.org/10.3905/jfds.2021.1.078
14. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), pp. 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
15. Jain, A. K., and Dubes, R. C. (1988). Algorithms for clustering data. Prentice-Hall, Inc, New Jersey, U.S.
16. Jones, J. M. (2017). US stock ownership down among all but older, higher-income. *Gallup Economy*, Washington, D.C., United States.
17. Keogh, E., Lonardi, S., and Chiu, B. (2002). Finding surprising patterns in a time series database in linear time and space. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, PP. 550-556. https://doi.org/10.1145/775047.775128
18. Kolm, P. N., Tütüncü, R., and Fabozzi, F. J. (2014). 60 Years of portfolio optimization: Practical challenges and current trends. *European Journal of Operational Research*, 234(2), pp. 356–371. https://doi.org/10.1016/j.ejor.2013.10.060
19. Kumar, P., and Wasan, S. K. (2010). Comparative analysis of k-mean based algorithms. *International Journal of Computer Science and Network Security*, 10(4), pp. 314–318.
20. Leng, M., Lai, X., Tan, G., and Xu, X. (2009). Time series representation for anomaly detection. *2nd IEEE International Conference on Computer Science and Information Technology*, 10868358, pp. 628–632. https://doi.org/10.1109/ICCSIT.2009.5234775
21. León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., and Niño, J. (2017). Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science*, 108, pp. 1334-

1343. https://doi.org/10.1016/j.procs.2017.05.185
22. Liao, S.-H., Ho, H., and Lin, H. (2008). Mining stock category association and cluster on Taiwan stock market. *Expert Systems with Applications*, 35(1–2), pp. 19–29. https://doi.org/10.1016/j.eswa.2007.06.001
23. Lohre, H., Rother, C., and Schäfer, K. A. (2020). Hierarchical Risk Parity: Accounting for Tail Dependencies in Multi-asset Multi-factor Allocations. *Machine Learning for Asset Management: New Developments and Financial Applications*, pp. 329-368. https://doi.org/10.1002/9781119751182.ch9
24. Molyboga, M. (2020). A Modified Hierarchical Risk Parity Framework for Portfolio Management. *The Journal of Financial Data Science*, 2(3), pp. 128–139.
25. Nakagawa, K., Imamura, M., and Yoshida, K. (2019). Stock price prediction using k-medoids clustering with indexing dynamic time warping. *Electronics and Communications in Japan*, 102(2), pp. 3–8. https://doi.org/10.1002/ecj.12140
26. Nourahmadi, M., and Sadeqi, H. (2021). Hierarchical Risk Parity as an Alternative to Conventional Methods of Portfolio Optimization: (A Study of Tehran Stock Exchange). *Iranian Journal of Finance*, 5(4), pp. 1–24. https://doi.org/10.30699/ijf.2021.289848.1242. In Persian.
27. Pardalos, P. M., Sandström, M., and Zopounidis, C. (1994). On the use of optimization models for portfolio selection: A review and some computational results. *Computational Economics*, 7(4), pp. 227–244.
28. Raffinot, T. (2017). Hierarchical clustering-based asset allocation. *The Journal of Portfolio Management*, 44(2), pp. 89–99. https://doi. org/10.3905/jpm.2018.44.2.089
29. Raffinot, T. (2018). The hierarchical equal risk contribution portfolio. *Risk Management Journal*, http://dx.doi.org/10.2139/ssrn.3237540
30. Rai, P., and Singh, S. (2010). A survey of clustering techniques. *International Journal of Computer Applications*, 7(12), pp. 1–5.
31. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., and Lin, C.-T. (2017). A review of clustering techniques and developments. *Neurocomputing*, 267, pp. 664–681. https://doi.org/10.1016/j.neucom.2017.06.053
32. Snow, D. (2020). Machine Learning in Asset Management—Part 2: Portfolio Construction—Weight Optimization. *The Journal of Financial Data Science*, 2(2), pp. 17–24. http://dx.doi.org/10.3905/jfds.2020.1.029
33. Tatsat, H., Puri, S., and Lookabaugh, B. (2020). Machine Learning and Data Science Blueprints for Finance From Building Trading Strategies to Robo-Advisors Using Python. O'Reilly Media, Inc, California, United States.
34. Thuraisingham, B. M., and Ceruti, M. G. (2000). Understanding data mining and applying it to command, control, communications and intelligence environments. *Proceedings 24th Annual International Computer Software and Applications Conference*, Taipei, Taiwan, pp. 171–175. http://dx.doi.org/10.1109/CMPSAC.2000.884710
35. Wang, H., Wang, W., Yang, J., and Yu, P. S. (2002). Clustering by pattern similarity in large data sets. *Proceedings of International Conference on Management of Data*, pp. 394–405. https://doi.org/10.1145/564691.564737
36. Williams, G. (2011). Data mining with Rattle and R: The art of excavating data for knowledge discovery. Springer Science & Business Media. New York, London. DOI:10.1007/978-1-4419-9890-3