Adv. Math. Fin. App., 2025, **10**(1), P.113-129

Advances in Mathematical Finance & Applications
www.amfa.iau-arak.ac.ir
Print ISSN: 2538-5569
Online ISSN: 2645-4610
Doi: 10.71716/amfa.2025.23011843

Original Research

Modeling the Co-Movement of Stocks Between Returns with Negative and Positive Shocks of Sentiment Arising from the Imbalance of Orders Using a Tree-Stage Clustering Approach

Kamal Ghanaei [a], Mehrdad Ghanbari [a, *], Babak Jamshidi Navid[a], Afshin Baghfalaki[b]

[a] *Department of Accounting, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran*
[b] *Department of Economics, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran*

ARTICLE INFO

Article history:
Received 2023-01-01
Accepted 2023-08-23

Keywords:
Emotional Shocks
Return Co-Movement
Three-Stage Clustering

Abstract

This study explores the impact of emotions on financial markets. Recent research highlights the role of psychological factors in financial crises. Investors, not always rational, base asset risk decisions on emotions and beliefs. Optimism, pessimism, and self-confidence influence decision-making processes over time. Irrational investors can cause market prices to deviate from fundamental values. This emotional price anomaly can be seen as a persistent sentiment risk factor significantly affecting stock returns. The research investigates the effect of emotional shocks (positive and negative) on order flow imbalances and their impact on price movements. A three-stage clustering approach is used. The sample includes 172 companies listed on the Tehran Stock Exchange from 2021, with daily data extracted from Rahvard Navin software. Information on independent and dependent variables is analysed using a three-step clustering method to identify different time scales. Finally, regression analysis in MATLAB software is used to examine the relationship between variables within specified time intervals. The results reveal a positive relationship between positive shocks caused by order imbalances and company returns. This relationship is reversed for negative shocks. Three-stage clustering separates companies based on co-movement, revealing distinct behaviour and relationships between variables within each cluster. These findings demonstrate the effectiveness of the three-step clustering method for analysing company data.

1 Introduction

Traditional financial theories are based on the two principles of the rationality of economic factors and the efficient market hypothesis, and they state that the competition between investors who are looking for unusual profits, causes the price of bonds to always approach its fundamental value. The emergence of exceptions and unusual phenomena in the market posed seri-

* Corresponding author. Tel.: +98 9181320812
E-mail address: *jamshidinavid@iauksh.ac.ir*

ous challenges to the two main foundations of traditional financial theories and raised the basic question of what factors other than the company's fundamental values affect stock price changes. In this context, behavioural finance literature states as a new approach in response to unusual phenomena in the market that traditional financial theories are unable to explain, some capital market players do not make rational decisions and have irrational behaviours, and their decisions result from it is their mentality or prejudices. Therefore, stock price changes do not only depend on the company's fundamental values, but also depend on the irrational behavior of investors, which is measured by investors' intentions [3]. Due to the high growth of the stock market index and the greater efficiency of the stock market compared to other parallel markets in the past two years and the attraction of small investors to this market, we have witnessed the sudden entry of a very large amount of capital into this market, which, along with the lack of sufficient knowledge of investors, has caused Investors' emotions and feelings have influenced the trading process and the direction of the market. The occurrence of inefficiency in monetary and financial markets due to the uninformed movements of people or in other words their momentary and short-term rush to any of these markets can cause economic crises. This issue has caused the investigation and recognition of investors' feelings and the resulting effects to become doubly important. Among the information that can affect investors' feelings and decisions is traders' knowledge about the movement of several stocks together (co-movement of stocks), which can play an important role in choosing people's stock. Meanwhile, the method of reviewing and analysing raw data is doubly important for decision makers because new and different methods of information processing are available to people, which sometimes produce different results. Due to the exponential increase in data and its communication tools, including the Internet and virtual networks and various information exchange channels, as well as the expansion of accounting and financial boundaries and entering interdisciplinary fields and creating new fields of research, including behavioural finance, working capital, and other capital market actors. They are faced with a huge amount of different data. This large amount of data alone does not help users of information; it also causes more confusion. Therefore, managing raw data and converting internal and external data of the organization into information and knowledge using various methods has a fundamental and pivotal role. One of the most basic methods for analysing different issues is to use patterns and examine them. The question that arises here is what the data have in common. Various methods have been used to check co-movement between data. Among them, data mining is one of the famous and widely used methods that can be performed on the database and obtain the required knowledge. Data mining is the art and science of intelligent data analysis with the goal of finding insights and knowledge from research data. Data mining can be used to perform tasks such as classification, prediction, estimation and clustering of data. To do this, methods have been developed, some of which are clustering algorithm, neural networks, genetic algorithm, nearest neighbour and decision tree [2]. Clustering is one of the most powerful methods for discovering natural groups and dependencies in a data set, as well as recognizing the structural patterns in it, without having any background about data characteristics. Considering the lack of sufficient research in the field of behavioural finance using new data mining methods, including the three-stage clustering method in behavioural finance research as described above, in this research, the co-movement of stock returns with negative and positive emotional shocks was investigated. From the imbalance of orders, it has been dealt with using a three-stage clustering approach [2].

2 Theoretical Foundations and Research Background

Over the past years, a large amount of different researches have been conducted in the field of financial behaviour and have analysed the impact of investors' tendencies on market prices and stock returns. For example, [3] showed through their theoretical model that the occurrence of irrational investors may cause the market price to deviate from its fundamental value. Due to the significant limitations of arbitrage and limited investment horizons, rational investors are discouraged from accepting the right position against other irrational investors, so that the stock price has not managed to return to its intrinsic values [3]. This price anomaly caused by irrational and emotional traders can be interpreted as a persistent sentiment risk factor that significantly affects stock returns. Proponents of behavioural finance argue that investors are not only exposed to fundamental risk, but also sentimental risk, given that irrational investors are exposed to this risk. Based on the return risk model, stocks that are highly exposed to emotions should have higher expected returns than stocks that are not highly exposed to emotions, which will result in their market prices being lower than their intrinsic value [5]. There is convincing evidence that the state or sentiment of the stock market can play an important role in influencing the extent to which individual stock returns move in tandem with other parts of the market [8]. Showed that the order flow imbalance of buying activity relative to selling activity is a useful measure of sentiment. This relation exists because an optimistic mood encourages more buying activity and less selling, while a pessimistic mood encourages more selling and buying. [8] Has used order flow imbalances as buyouts to correct sentiments and show that sentiments increase co-movement [5]. Hypothesis is that market sentiment is in the hands of real investors' sentiments, because these investors have access to information sources that institutions have, so they make more irrational investment decisions. However, there are evidences that show that legal investors also act recklessly and are influenced by emotions .Legal investors become emotional traders ,Because they are influenced by reputational trading, which encourages institutions to trade on sentiment to avoid mediocre performance, While the short-term predictability of sentiment strategies and the arbitrage barrier of low-cost evaluation make institutions able to use the benefits of short-term prediction caused by sentiment because it is profitable. Consequently, the order flow of institutional investors reflects their sentiments. In addition, a recent analysis of the order flow of real and legal investors by de [6]. Shows that legal investors are more influenced by emotions than real investors. Investors are trying to identify the most suitable stocks for investment and direct their capital resources to that direction. From the point of view of market traders, obtaining any information in market analysis is of great value; because having information can reduce investment risk and even cause big profits. In this regard, the traders' knowledge about the movement of several shares together can play an important role in the selection of people's stock portfolio, and the movement between the prices of companies' shares is a worthy phenomenon that has been of interest since the emergence of the first market theories. In addition, by examining the movement of prices in a group, it is possible to predict the stock prices of companies. Also, examining the co-movement and clustering of companies provides a new solution for preparing and presenting the necessary indicators to evaluate the market situation at the disposal of the Stock Exchange Organization. Clustering of companies based on their respective time series data will evaluate the co-movement of the companies' shares. As a result, companies that have the same movement trend, or in other words, the same

market shape, are placed in a cluster. These categories provide investors with useful information for market analysis by stating which companies have similar trends. In fact, the clustering of companies creates a suitable solution for predicting the stock price of companies according to the past movement trend or the movement trend of the price of other companies in its group [4]. Gudarzi Farahani, the Co-movement between Bit coin, Gold, USD and Oil: DCC-GARCH and Smooth Transition Regression (STR) Model. This study investigates the relationships between Bitcoin (BTC) prices and fluctuations in relation to gold, USD, and Iran's oil prices from 2019 to 2022. We employed the dynamic conditional correlation generalized autoregressive conditional heteroscedasticity (DCC-GARCH) method to model the fluctuations of financial variables. Additionally, the smooth transition regression (STR) method was applied to explore the relationships between the variables. The results reveal significant positive correlations between BTC prices and gold, as well as oil, and a negative correlation with USD prices. We observed volatility persistence, causality, and phase differences between BTC and other financial instruments and indicators. Notably, a negative relationship was identified between Bitcoin and the USD in both linear and non-linear aspects, with a larger coefficient in the second regime. Furthermore, a positive relationship was found between Bitcoin and the variables of gold and oil prices, with coefficients being larger in the second regime compared to the first. [9]. Khozein et al, Khozein, A Model of Investor Sentiment Based on Grounded Theory Approach, investor expectations regarding future economic processes are among the crucial factors that influence their decision-making. These expectations play a unique role as they are unmonitored variables capable of shaping observable economic phenomena. Psychological factors have a significant impact on both investor expectations and corporate market value. This study focuses on modeling investor sentiments with an emphasis on psychological factors, utilizing the Grounded Theory (GT) framework. The research is conducted through applied and mixed methods at its initial and subsequent stages. The statistical population for this study comprises 13 experts, senior managers of investment companies, and university professors. Participants were selected using purposive and snowball sampling techniques, continuing until theoretical saturation was achieved. Data col lection was carried out through semi-structured interviews, which were then coded using Atlas.ti 8 software. The research data were analyzed using an open coding method, leading to the identification of 46 categories and 6 key dimensions as the research results [10]. The results of some research on cost stickiness show that selling and general administration (SGA) costs, as well as costs of goods sold, have highly sticky behaviors [12]. Balounejad et al. Impact of Investors' Sentiments on Volatility of Stock Exchange Index in Tehran. The stock market is one of the main components of the economy, and various factors cause fluctuations in it, one of which is the effect of investors' behavior. Therefore, present study seeks to answer the question of whether the feelings and sentiments of investors might intensify the fluctuations in the Tehran Stock Exchange or not. To answer this question, at first, in order to quantify sentiments, as non-abstract variables, the Equity Market Sentiment Index (EMSI) was used that investors are classified in 5 categories of completely risk-averse, risk-averse, neutral-risk, risk-taking and completely risk-taking. Using GARCH-in-Mean, results indicate that the sentiments of investors will result in greater fluctuations in the Tehran Stock Exchange. Hence, if fluctuation is considered an indicator of market risk, the excitement associated with an abnormal rise in volumes will increase that risk [11]. In this research, the impact of negative and positive shocks on emotions is comprehensively investigated as imbalance shocks of the co-movement order flow using a three-stage clustering approach. Examining shocks has

advantages. First, the shocks do not have unexpected changes in the order flow imbalance, so the factor of the flow imbalance reflects the emotions that indicate an innovation or a change in behaviour, so it may have a different effect on returns and co-movement. In order to have levels of imbalance. The levels of order flow imbalances are nonstationary, but the change in order flow imbalances is stationary, which provides more incentive to focus on both negative and positive shocks. In this research, we consider two types of shocks. The first change is the level of order flow imbalance between successive periods. Using this concept of shock, in this research we estimate the average double correlation between excess market return and change in sentiment for each group of investors. In general, the issue that makes this research necessary is that what effect do positive and negative emotional shocks have on stock returns using three-stage clustering. The findings of some researchers showed that there is a significant relation between the stock market uncertainty changes in an economic boom and the investment risk in general, which is not significant in terms of the economic turndown. The Investment risk during both economic boom and recession is decreased by the unexpected increase in profit of each share and propagation of positive news. Although the risk is increased by the spread of negative forecasts in relation to shares [13]. Other findings suggest that cost stickiness has a positive impact on the relationship between institutional investors and passive institutional investors with conservatism [14]. Javaheri and zanjirdar showed that there was a significant relationship between the profit management and companies' performance. The profit management is also effective in forecasting future cash fund, in forcing solidarity between running and future yield [15].

3 Methodology

The present research is "applied" in terms of the purpose of implementation, "quantitative" in terms of data type, and "descriptive, survey and correlational" in terms of implementation method. The statistical population of the research is all the companies admitted to the Tehran Stock Exchange. In this research, the systematic elimination sampling method with the sieve method (or the judgment method) was used. The time period of the current research includes the beginning of 2021 to the end of this year .In the current research, the data required by the companies on a daily basis in the time interval of 6/1/1400 Until 12/28/1400, it will be received through Rahavard Navin software, it is worth mentioning that the data used are adjusted, that is, cash profit and increase and decrease of capital are included in the desired data. The number of working days during the desired period is 237 days .Since for conducting the present research according to the project methodology, the number of data should be the same for all the investigated companies, the number of data for each company was considered to be 200 data. On the other hand, due to the fact that a number of companies did not provide enough data during the review period, the companies whose available data was less than 200 data during the year were excluded from the comparison, and finally, 172 companies for Model implementation was used.

Hypothesis 1: Positive and negative shocks caused by investors' inclinations have an effect on companies' stock returns.

Hypothesis 2: The effect of positive and negative shocks on the efficiency of companies is greater in the same movement clusters according to the three-stage clustering method.

In order to achieve the objectives of the research, first the industries active in the stock exchange have been selected, then we collect the information related to the efficiency of the companies for the research period (2021) on a daily basis and we separate them from each other using the three-stage clustering method, then the relationship of the variables We analyse inde-

pendent and dependent according to the different clusters calculated in the previous step. Clustering is an unsupervised learning method. An unsupervised learning method is one in which the target dataset contains data without a target label or a group to which the data belongs. In general, it is used as a process to find a meaningful structure or pattern to group data. Clustering is the task of dividing the population or data points into a number of groups so that the data points in the group that is a member are most similar to It should have other data points in the same group and not be similar to data points in other groups. Basically, a set of objects or data are divided based on similarity and dissimilarity; It is a core task for data mining and a method for big data analysis that is used in many fields, including pattern recognition, image analysis, market research, information retrieval, bioinformatics, data compression, computer graphics, and machine learning. The features based on which these algorithms create these clusters include clusters with small distance between cluster members, clusters with high data density, distances and special statistical distributions. The data points in the graph below that are clustered together can be considered as a cluster. Clustering algorithms place data that have similar and close characteristics in separate categories called clusters.

First step: Reducing data dimensions and approximate clustering

[2] presented a new method based on PAA1 called SAX2. SAX is the first proposed symbolic method to reduce the dimensionality of time series defined by the approximate Euclidean distance function. SAX based on this assumption it is established that the data have a Gaussian normal distribution; therefore, the data must be normalized first. Based on the method PPA, by defining the breakpoints (β_2) to the data of each section according to the tables, it assigns a symbol and calculates the distance based on these defined symbols. In other words, the data reported from the SAX method after calculating the mean value for each PAA segment which is actually the average of each section, is discretized .This break is created by defining the number and location of breakpoints($\beta = \beta_1, \beta_2, \cdots, \beta_{a-1}$). $\hat{F}_1 = (\hat{f}_1, \hat{f}_2, \cdots, \hat{f}_w)$The data are symbolized by SAX method.[4] State that the minimum distance measurement method 3 (MINDIST) is a suitable method to calculate the distance for the SAX method. MINDIST based on the distance between the data Calculates the shortest distance between them. The problem of the MINDIST method is that it considers the distance between the index of adjacent locations (average data of each section) to be zero and also ignores the maximum and minimum points of the data. The APXDIST method is a modified MINDIST method that solves this problem. In this way, for each area, an index is defined based on the broken points obtained in other words, each indicator is the average sensitivity of the breaking points at the beginning and end of the said area.

$$0 < i < a \; Ind_i = \frac{\beta_{i-1} + \beta_i}{2} \tag{1}$$

Where $\beta_i$andβ_{i-1} are the maximum and minimum data points. β_oIs the overall minimum and $\beta_a$is the overall maximum for the time series data. Then the distance between the indicators is calculated based on equation (2) and time series clustering is done.

$$dis_{APXDIST} = \sqrt{\frac{n}{w}} \sqrt{\sum_i^w = 1 \left(dis\left(Ind_i, Ind_j\right)\right)^2} \tag{2}$$

Where w is the number of sections, n is the total number of data and $dis\left(Ind_i, Ind_j\right)$ the distance between two indexes$Ind_i \; and \; Ind_j$. After obtaining the distance matrix between the

companies, to cluster the data at the end of the first step, using the K-Medoids method which is from the group of K-Means algorithms is, will be used.

Second step: Determination of sub-clusters

The purpose of this step is to modify the clusters created in the previous step and improve its quality. In this step, the number of data sets is reduced by defining samples in each group. The distance scale at this stage is the Euclidean distance (ED).

$$dis_{ED}(F_i, F_j) = \sqrt{\sum_i^n = (F_i, F_j)^2} \tag{3}$$

Which F_i, F_j are the two data from the preliminary clustering of the previous step. In other words, at this stage, the clusters created in the first stage of review and sub-clusters are formed. The sub-clusters (SC_{ij}) are actually the data whose movement trend is similar in the same time period, which are created from the breakup of the clusters of the previous stage.

Consider the primary cluster (PC_i), the distance matrix between the data in this cluster is calculated and displayed as a M_{n*n} matrix, so that M_{ij}represents the distance between data i and j. The developed PCS1 algorithm separates the primary clusters using the defined distance matrix. In the second step, each cluster is built based on the threshold limit value of each time series, the concept of the cluster threshold limit is obtained from the CAST2 method [4] in such a way that for each data in the initial cluster PC_{ij}an approximate similarity threshold value is defined based on the following relationship:

$$a(F_x) = \frac{\sum_{y \in c} M_{xy}}{|SC|} \tag{4}$$

Where M_{xy} is the data distance matrix F_x ,F_y and $|SC|$ is the number of data in the new cluster. PCS creates clusters continuously based on dynamically defined thresholds. The output of the PCS method is the clusters created by separating the clusters of the previous step. In this process, each cluster is constructed by adding the time series that have the most similarity with the members within the cluster. The addition and subtraction of data from each cluster continues until no more changes are possible in the created cluster. In this step, to determine the number of sub-clusters created, a threshold limit $3(\alpha)$ is defined for each cluster to accept data entry in it:

$$\alpha = \frac{\sum_{xy\epsilon} PC_U, M_{xy} \geq \mu(M_{xy} - \mu)}{|PC_U|} + \mu \tag{5}$$

$$\mu = \frac{\sum M_{xy}}{|PC|}$$

$\boldsymbol{\mu}$ Is the initial threshold value in the algorithm and $|PC_u|$ is the remaining number of data in the initial clustering at each step. The threshold value defined in this step is calculated dynamically, which means that in each step, the number of remaining data in the initial cluster is checked and entered in the new cluster if the conditions are met. The parameter defined in the above equation controls the number and size of the created clusters, and this is one of the advantages of the method used in this research [10]. Then, for each created cluster, a sample (CS_j) is defined it can be done as follows:

$F_i = (F_{i1}, F_{i2}, \cdots, F_iT)$ the data in (CS_j) (6)

$R_i = (r_{i1}, r_{i2}, \cdots, r_{ix}, \cdots, r_{iT})$

R_i the saple is defind for the i-th cluster, where the r_{ix}values are calculated from the following equation:

$$r_{ix} = \frac{\sum_{i=1}^{n} f_{ix}}{n} \tag{7}$$

In other words, the sample definition for the data in each cluster reduces the complexity of calculations and increases the accuracy of clustering, by using this definition, remote points of the time series are also considered. Finally, all the data are checked and by considering the threshold limit for each cluster, more accurate and better quality clustering is the result. The third stage: Integration

Examining important or critical points (PIPs) is one of the important issues in time series data clustering. The concept of PIPs in financial data analysis means keeping the important points in the data and removing the outlying points. In another research, [7] proposes a model for using the concept of inflection points 2 in financial time series. Inflection points are important and critical points in financial data that should be considered. The output of the second stage provides indicators for each cluster ($\boldsymbol{R_i}$). In the third step, the distance between these indices is calculated to create the final clusters. In this step, local shifts are checked. As it was mentioned, examining the movement trend of companies' stock prices by considering time shifts will provide a more realistic result. According to the stated contents, since the sameness of the market shape of the companies is investigated, using the DTW3 method to measure the distance seems more appropriate. This distance measurement method achieves the best alignment between the data in the series by using the 4th swing axis. To better understand the calculation process of the third step, it is assumed that two clusters X and Y have been selected from the clusters created in the previous step. We assume that cluster X contains n companies and cluster Y contains m be a company to find the distance between the clusters $(SC_x), (SC_y)$ we consider the indices calculated for these two clusters:

$$\begin{aligned} R_x &= \{r_{x1}.r_{x2}.\cdots.r_{xi}.\cdots.r_{xn}\} \\ R_y &= \{r_{y1}.r_{y2}.\cdots.r_{yi}.\cdots.r_{yn}\} \end{aligned} \tag{8}$$

The distance matrix between the data in each index is calculated by considering the Euclidean distance; So that the result will be a matrix n*m, $M_{ij} = d(r_{xi}, r_{yj})$shows the distance obtained for company i-th from X cluster and company i-th from Y cluster. Now the optimal path between two clusters X and Y can be calculated using equation 9 which is the DTW distance measurement method.

$$\begin{aligned} &dis_{DTW}(R_x, R_y) = min\left(\sum_{k=1}^{K} W_k /K\right) \\ &\text{Swing point } W = (W_1.W_2.\cdots.W_u) \\ &W_u = \{(r_{x1}.r_{y2}).(r_{xi}.r_{yj}).\cdots.(r_{xn}.r_{ym})\} \end{aligned} \tag{9}$$

Finally, having the distance matrix between the cluster indices calculated using the above relationships, re-clustering is performed using the k-medoides method, the result of this clustering is the merging of clusters that had similar time series. Which were not placed in the same cluster in the previous steps [6]. Finally, using MATLAB software, we will investigate the relationship between independent and dependent variables in specified time intervals.In order to test the hypotheses, the variables of this research are divided into two groups of independent and de-

pendent variables. The independent variable of negative and positive emotional shocks is derived from the imbalance of orders, which is measured as follows:

In this research, investors' reactions (feelings of investors) are measured using the order imbalance variable, abbreviated as OIB. It is also the basis for identifying buying and selling transactions using the closing price of the previous day. Thus, if at a price higher than the closing price of the day Before (the positive scope of the transaction price fluctuation, we call it a buy order, and if the transaction is executed at a price lower than the previous day's closing price (the negative scope of the transaction price fluctuation, we call it a sell order), OIB is equal to the Rial value of the purchase orders Minus the Rial value of sales orders divided by the total orders placed on a share every day[5].

So we have:

$$OIB = \frac{B.V - S.V}{S.V + B.V} \tag{10}$$

Where

OIB= is the index of imbalance in sales (buying) orders,

B.V= the Riyal value of share purchase, which is calculated as follows:

$$B.V = \sum_{t=1}^{N} (Bvol_{it}, N_{it}, P_{it}) \tag{11}$$

Bvolit= Purchase order volume

Nit= Number of purchase order times

Pit= Share purchase price

S. V= Riyal value of share sale, which is calculated as follows:

$$S.V = \sum_{t=1}^{N} (Svol_{it}, N_{it}, P_{it}) \tag{12}$$

Svolit= Volume of sales order

Nit= Number of sales order times

Pit= Share sale price

Negative shocks are also equal to negative values OIB (NOIB), and positive shocks are also equal to values OIB (POIB). The dependent variable of this research is stock return, which is measured as follows:

The real yield of each ordinary share is calculated according to the fluctuation of stock price, cash profit, dividend and capital increase. The total efficiency can be calculated using the following equation:

$$\frac{Stock\ value\ at\ the\ end\ of\ yaer - Stock\ value\ at\ the\ start\ of\ yaer + Cash\ income\ of\ shareholders}{Stock\ value\ at\ the\ start\ of\ periode + Cash\ income\ of\ shareholders} \tag{13}$$

7 Findings

Financial time series have special characteristics compared to other time series. Among these features is that this category of time series is shown with few critical points, in other words, financial data are critical and important data at any time and require great accuracy in short-term and long-term reviews. The second special feature of the financial time series is the con-

tinuous, large and unlimited nature of these data. As stated, there are different algorithms for data clustering that create clusters with acceptable fundamental quality and accuracy and are effective for static data at a time point. Research has shown that classical data mining algorithms do not work well for time series clustering; Because the dimensions of the data and the correlation between the data are high, on the other hand, a large number of disturbances (noises) in the time series data make the clustering of these data difficult. One of the basic problems in time series data mining is the problem of searching for similarity or measuring distance. When the number of data is small, the similarity between the data can be checked by comparing them two by two; But when the number of data is very large, or the data has large dimensions, this method will not be effective. There are different methods to measure the distance between time series data, among which, ED and DTW methods are common methods in time series clustering. Although measuring the similarity distance between data using the Euclidean distance (ED) is a suitable measurement method in many existing clustering methods; But it does not have enough precision in examining the movement between the stock prices of the companies; Because this method only expresses the temporal similarity between the data. This is despite the fact that the final price of the companies' shares has time shifts and the similarity in the movement process over time should also be considered. The figure below shows the movement process of two companies C and B.

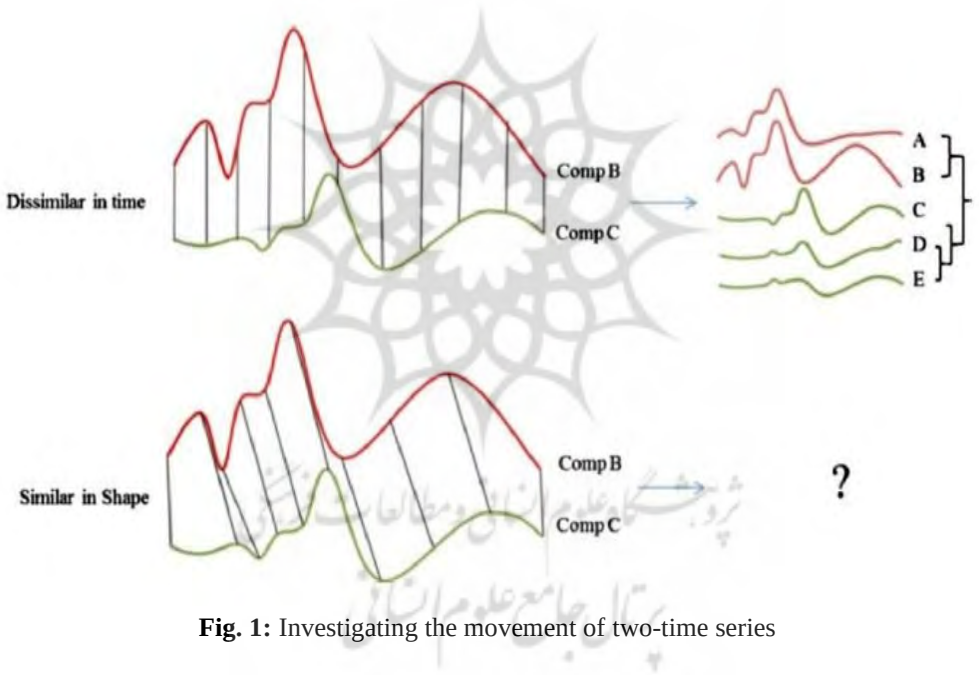

Fig. 1: Investigating the movement of two-time series

According to Shape1, two companies C and B may be placed in two different clusters in the usual clustering methods, considering the data at any moment of time, or there may be no relationship between their movement trends; While a closer look at the analysis of the movement trend shows that two companies C and B have a similar movement trend with some delay with a time shift, or in other words, there is a movement between the two time series; Therefore, clustering these data based on similarity at a point in time may not provide correct results.The advantage of the DTW method over the ED method is that DTW has the ability to consider time shifts and examine the movement of data over time. Figure 2 shows the difference between ED and DTW measurement methods. As it is known, the ED method only considers its time

equivalent point to check the co-movement between the data of two-time series at any moment of time; But in the DTW method, each time series data is compared with several other time series data, and finally the data that has the smallest distance with the examined data is used.
In other words, the DTW method also takes into account the problem of local shifts of time series and seeks to find series that do not show movement at any point in time; But their movement process over time is similar despite the shifts [7].

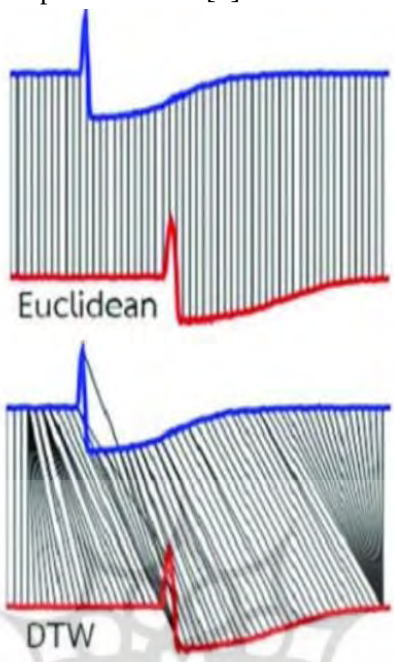

Fig. 2: Examining two methods of measuring distance ED and DTW

The three-stage clustering method also considers the existing time shifts and can create clusters based on the similarity in the movement process of time series. In other words, the three-stage clustering method, in addition to considering the data in a time unit, compares the data over time and provides a better result [6]. The number of working days during the period considered in this research is 237 days. Since the number of data for all investigated companies must be the same for conducting the present research according to the methodology, the number of data for each company is considered to be 200 data. On the other hand, due to the fact that a number of companies did not provide enough data during the review period, the companies whose available data was less than 200 data were excluded from the comparison, and finally 172 companies were used to run the model. The collected data were first arranged in the form of regular time series in Excel software in order to provide the basis for data analysis and model design and subsequent actions. In the first part, the clustering of the sample companies is done based on the yield variable and by the three-stage clustering method, and then within each cluster, the relationship between price shocks caused by investors' sentiments on stock returns is examined using regression. In this step, we first calculate the data related to the returns for all the sample companies in the mentioned time period, then we cluster the companies using the three-step clustering method. For this purpose, after entering the data in Excel, we enter the data in the Matlab software environment and using the facilities of this software, we provide the field of data review and analysis and model design to achieve the results.In the first step, the

defined value for the number of clusters (K) in the K-Medoieds method

It is equivalent to $\sqrt{\frac{n}{2}}$.

n is the total number of data in each time series, or in other words, the number of data of each company during the period under review (200). To divide the data into different parts, the parameter W is defined. In this research, the considered values for this parameter were 4, 6, and 8. The results of running the model are different considering different values of W parameter. To determine which value is more suitable for this parameter and shows better clustering, the sum of squared error (SSE) value for different values of W It is calculated using the following equation.

$$SSE = \sum_{j}^{K} = 1 \sum_{F_i \in SC_j} \left(dis(F_i, R_j)\right)^2 \quad (14)$$

The lower the value obtained for SSE, the better the model. Below are the results for the SSE values:

Table1: The Value Obtained for SSE

W=8	W=6	W=4
0.0385	0.1082	0.1317

According to the obtained values for SSE, choosing the value of W=8 will give better results.

The following graphs show the results of the execution of different stages of the model with the value of W=8.

Table 2: member companies of each cluster at the end of the first stage with W=8- first stage

1	2		3		4	5	6		7		8		9	10
foolad	vebsadea	zdasht	reksh	vasakt	vpol	hamra	petrol	fars	vasanat	froh	goharan	sedabir	defara	barkat
medko	vapost	dayro	salond	sestem	desja	akhaber	ap	valber	vmelat	fnft	faraor	vabahman	vapakhsh	dalqma
fasapa	thgarb	tutsa	vesapa	sfars	jabal	jam	sajen	shakbari	zagros	lavan	fkhas	frzin	jahrom	dezheravi
madaran	smaye	aryan	fanval	smaskan	fulad	hormuz	shahrdad	shafa	khakestar	shahrlo	shituka	sina	shiraz	foulah
shfn	thshahd	mehr	margham	vapoya	mofakhar	khuzestan	zanjan	flat	khadizl	shalamcheh	shireto	vatoka	chekaveh	va tejarat
vghdir	thnam	zokoother			payment	payment	afra	shimad	kharkar	kavir	hataco	fabahonar	hatoka	vabemelat
	novin	tehran			shaghdir	kermanshah	lasemerd	caspian	sina	falomineh	etela'at	sodor	hebandar	paksho
	thnosa	sepahan			toril	darab	toos	sejam	sekhash	faezeh	dayereh	ssafha	hfars	khodro
	vati	zagros			seshmal	shituka	khloran	soroud	dabid	sepahan	qasherin	kerazi	shabendar	talese
	vsepe	tooshe			sarom	jam	sharak	qonqash	sarsharg	shiraz	qanisha	ktoka		
	tekba	tanur			vohukmai	shenavar	qazvin	qahekamat	sekard	novin	kamrjan	fajr		
	shkhark	khavar			ravan	ghadir	semega	btrans	kerman	shireto	bemoto	zangan		
						ghanoosh	teknar	bekab			chekape	kesapa		
						ranfoor	epardaz	kahafez			saadi	kbahman		
							khattrak	kapars			ghapinoo	ploleh		
							isfahan	chekarn						
							alborz	parsian						

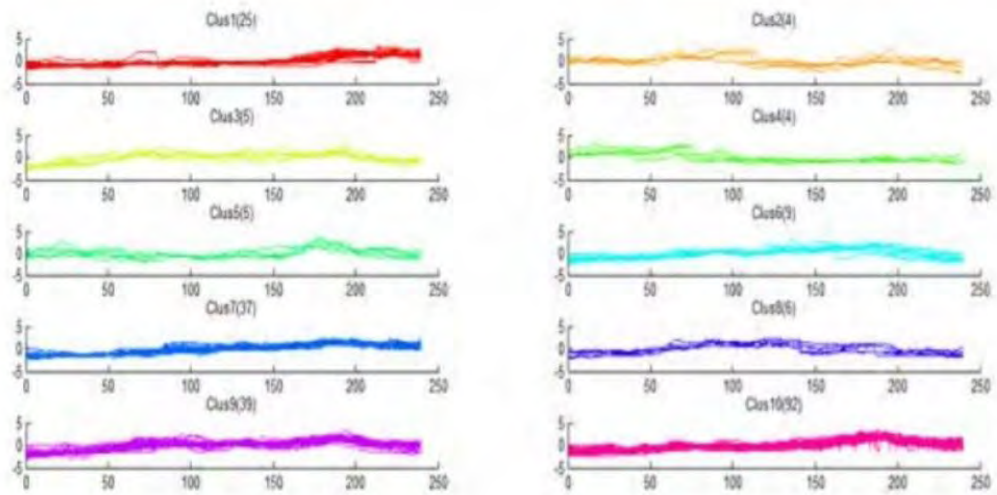

Fig.3: First Results for W=8

Table 3: The result of the second stage with W=8- second stage

1	2		3		4	5	6		7		8		9	10
folad	vsacht	zdasht	rakish	vsacht	voboli	hamrah	petrol	fars	v malat	farooqah	goharan	sedbir	dafara	barakat, barak, barakah
vebsad	voboli	vtoousa	thalond	sistem	dsbha	akhbar	up	valbert,	zagros	fannift	fakhas	vobahman	jahrom	dolqoma
zdasht	hamrah	vtoousa	fnaval	thfars	djaber pen_spark	vjmi	shasfa	shakbir	kgosrat	laparis	shtooka	saina	shiraz	votjarat
rakish	petrol	aryan	marqam	vpova	fulaz pen_spark	hormoz	zanjan	shafara	khkar	shaklr	hataid	vtoka	chekaveh	vbmelt
	fars	zkouther			mafakher pen_spark	fakhouz	afra	smavad	sena	kaveh	etela	fbahnar	hatoka	khodro
	vamlat	vskap			thnzam	kermanshah	vatus	hkazar	dhmid	falom	dsanco	sedor	habandar	telis
	foroui	zmgsa			shghadi	shtoli	claire	srod	sshargh	setareh	qasherin	krazii	hafars	
	goharan	thnor			touril	jam	sharak	qnosh	skrd	shiraz	kamrjan	ktoka	shabndar	
	sedbir	vkhavar			sshomal pen_spark	shnaft	qazvin	gkomak	kerman	novin	bamoto	fajr		
					vhekmat	ghasalm	samgha	bkab		shapna	chekapa	zangan		
						ghnoush	epardaz	khafez			ksaadi	khasapa		
						ranfor	dhub	kpars			gpeno	khbeman		
							alborz	parsyan				poloule		
11	12		13		14	15	16		17		18		19	20
fasapa pen_spark	vpost	sathran	vesapa	smaskan	sarum	pardak	thajen	flatland	sanat	shellard	faravar	farzin	vapaksh	foulah
madaran	samaya	vomehan			retko	darab	lesarma	sajjam	khadizel	fayra	shapetro	safa	dezheravi	paksho
	teshahod	vatooshe					teknar	btrans	sekhash		qanisha			
							khotrak	chakaran						

Table 4: The result of the second stage with W=8

Third stage			
1	2	3	4
Cluster7	Cluster 1	Cluster 5	Cluster 8
Cluster 18	Cluster 2	Cluster 6	Cluster 9
Cluster 19	Cluster 3		Cluster 10
	Cluster11		Cluster 13
	Cluster 12		Cluster 14
	Cluster 15		
	Cluster 16		
	Cluster 17		
	Cluster20		

According to the results presented in the above tables, using the three-stage clustering method, the companies were first divided into 10 primary clusters. Then, in the second stage, the sub-clusters were determined, and finally, in the third stage, the clusters were merged in the form of 4 main groups occurred. Finally, as stated, for example, clusters 7, 18, and 19 are placed in one group, and this means that the companies in this group have more movement with each other in terms of efficiency. In the continuation of the research, we will investigate the relationship between price shocks and the returns of companies in the form of regression method. First, we examine the entire research data without considering the clustering, then in order to examine the effect of the three-stage clustering on the accuracy of the results, within each general cluster (the merged group of clusters in the third stage of clustering) We relate dependent and independent variables. First, the appropriate model for the regression model has been selected. Then, using Limer's F test, the selection of the combined model against the combined data model has been done.

Table 5: Selection of pooled data against combined data (Limmer F test) and selection of fixed effects model against random model (Hausman test)

Model	$RE_{i,j} = \alpha + \beta_1 OIB_{i,j} + \varepsilon$		
Test type	The value of the test statistic	degrees of freedom	Probability of the test statistic
Limer's F	2/355	10	0/0149
Hausman	5/428	9	o/828

The results of Limmer's test show that the null hypothesis of equality of individual effects is rejected. Therefore, the appropriate model for the estimation of the investigated model is in the combined class (panel). In the next step, the Hausman test was performed in order to select the fixed effects model against the random effects model. The result of Limer's F test and Hausman test are presented in Table 5. The presented results indicate that the null hypothesis is not rejected based on the superiority of the random effects model over the fixed effects model. Therefore, the model is suitable for estimating the random effects model. The mixed regression model of random effects for all companies is shown in Table 6.

Table 6: The relationship between positive and negative shocks caused by the imbalance of orders and returns

Statistics / Variable		Model coefficients	The value of the t statistic	The significance level
Constant		0/2851	3/3950	0/0000
Positive shocks (POIB)		0/1296	7/106	0/0000
positive shocks (NOIB)		-0/2834	-4/066	0/0000
The coefficient of determination	Adjusted coefficient of determination	F statistics	The significance level	Watson camera
0/409	0/388	19/451	0/0000	1/782

The results in Table 5 show that the relationship between positive shocks caused by the imbalance of orders at the level of all companies is positive (0.1296) and is significant according to the t-statistic (106.7). This shows that the positive shocks caused by the imbalance of orders are directly related to the efficiency of companies. Also, the results show that the relationship between negative shocks caused by the imbalance of orders at the level of all companies is negative (-0.2834) and is significant according to the t-statistic (-4.066). It is necessary to explain, for the sake of brevity, other tests related to the regression assumptions are not given, but the results of the tests indicate the normality of the data and the results of the unit root test of the type of Levin, Lin and Cho test indicate the papacy of the model variables. In the following, we will examine the relationship between positive and negative shocks caused by the imbalance of orders on the stock returns of companies in each of the clusters.

We pay in full.

Table 7: The regression model for each cluster

Statistics / cluster	cluster	Model coefficients	The value of the t statistic	The significance level t	The value of the f statistic	The significance level f	Adjusted coefficient of determination
First	Positive shocks (POIB)	0/0521	6/544	0/0000	.5214 10	0./0000	0.5403
	positive shocks (NOIB)	-0/0577	-12/863	0/0000			
Second	Positive shocks (POIB)	0/0838	4/908	0/0000	.6325 11	0.0000	0/2259
	positive shocks (NOIB)	-0/1069	-1/732	0/0837			
Third	Positive shocks (POIB)	0/4218	12/159	0/0000	.4209 35	0/0000	0/7203
	positive shocks (NOIB)	-0/1167	-10/552	0/0000			
Fourth	Positive shocks (POIB)	0/3108	1/861	0/059	15/588	0/0000	0/5633
	positive shocks (NOIB)	-0/6806	-8/098	0/0000			

Table 6 The relationship between positive and negative shocks caused by the imbalance of orders and returns. The results in Table 6 show that, in general, positive shocks have a direct rela-

tionship with companies' returns. Negative shocks are also inversely related to returns. Also, the results show that the relationship between positive and negative shocks caused by the imbalance of orders and returns in different clusters According to the three-stage clustering, it is different, as it is clear in Table 7, the significance level of negative shocks variable in the second cluster (-0.837) and also positive shocks in the fourth cluster (0.059) indicate no significant relationship with returns. Also, the adjusted coefficient of determination in the case of regression is different for each cluster. The coefficient of determination in the third cluster has the highest value (0.7203) and the lowest value in the second cluster (0.2259). The results of this table show that the clustering Data can provide more accurate results.

8 Discussion and Conclusions

Clustering, as one of descriptive data mining methods, is a technique for grouping observations into k different clusters (groups). Stock market clustering provides useful information to individual investors and financial professionals to predict changes in stock prices of different companies. In recent years, the clustering of companies admitted to the stock exchange based on the similarity in the movement process or in other words the shape of their market has been considered. In this research, using the three-stage clustering method, the data related to the changes in the stock prices of the companies were analyzed in the form of stock price co-movement, and then the effect of price shocks caused by the imbalance of orders on stock returns. The companies within the cluster created by the above method were tested. As stated in the statistical tables of the previous section, the three-stage clustering provides useful information regarding the co-movement of positive and negative shocks caused by the imbalance of orders. The statistical tables show that there is a direct relationship between the positive shocks caused by the imbalance of orders and the returns of companies at the overall level. Also, this relationship is reversed regarding negative shocks. By using three-stage clustering and separating companies based on their co-movement, it was determined that the behavior and relationship between the mentioned variables within the clusters is different. As the results of the statistical tables showed, the intensity and extent of communication within each cluster is different from the other cluster. These results show the important function of the clustering method of companies' information, because by placing similar companies in one category, it is possible to examine the relationship between variables more precisely. Also, the results show that companies in the same industry have more synergy with each other. Considering the close relationship of companies within an industry in terms of information and the nature of activity, placing companies in a cluster in the three-stage clustering method does not necessarily mean that the said companies were selected from the same industry. As indicated in the above tables, the companies in the same cluster are sometimes from different industries, and the investigation of the reasons for the similar behavior of the mentioned companies in the investigated time period can provide useful information about the behavior of the companies. provide prices and provide useful information to users.

Resources

[1] Agha Babaei Mohammad, Madani Saeed, Investigating investors' sentiments and the simultaneity of stock returns in Tehran Stock Exchange, *Financial Management Perspective Journal*, 2019; 11(42): 100-129. 1911-3846.1997.tb00537.x

[2] Balounejad Nouri, R., Bagjavany, F., Amiri Hosseini, M., Impact of Investors' Sentiments on Volatility of Stock Exchange Index in Tehran Stock Exchange. *Advances in Mathematical Finance and Applications*, 2023; 9(1): 291-303. Doi:10.22034/amfa.2023.1963410.1773

[3] Chelley-Steeley, P, Lambertides, N, Savva, CS. Sentiment, order imbalance, and co-movement: An examination of shocks to retail and institutional trading activity. *Eur. Finance Manag*. 2019; 25 (2): 116-159. Doi:10.1111/eufm.12146

[4] Dagindan, Al-Hiyar; Ali Akbar Najafi and Alireza Taskhiri, Investigating the impact of investor sentiments on investment decisions in companies listed on the Tehran Stock Exchange, Third International Conference on Accounting and Management, Tehran, *Mehr Eshraq Conference Institute*, 2020; 12(45). Doi:10.22051/JFM.2016.2483

[5] Gudarzi Farahani, Y., Aghari Ghara, E., & Haghtalab, M. The Co-Movement Between Bitcoin, Gold, USD and Oil: DCC-GARCH and Smooth Transition Regression (STR) Model. *Advances in Mathematical Finance and Applications,* 2024; *7*(3): 900. Doi: 10.22034/amfa.2023.1963410.1773

[6] Khozein, A., Davoudi, A., Naderian, A., & Didehkhani, H. A Model of Investor Sentiment Based on Grounded Theory Approach. *Advances in Mathematical Finance and Applications*, 2023; 8(4). Doi:10.22034/amfa.2022.1948257.1668.

[7] Liang, W-I, Sensitivity to investor sentiment and stock performance of open market share. l. *Advances in Mathematical Finance and Applications.* 2015; **58**(3):117-158,190-199. Doi: 10.1016/j.qref.2015. 01.003.

[8] Ling, D.C., Naranjo, A. and Scheick, B., Investor sentiment and asset pricing in public and Miwa k, Investor sentiment, stock mispricing, and long-term growth expectations, *Research in International Business and Finance*, 2012; 4(12). Doi: 10.22103/jak.2013.523

[9] Niya Mohammad, Ahmad Poyanfar and Maliha Maliki, Modelling the co-movement of shares in Tehran Stock Exchange using the three-stage clustering approach, *Financial Management Perspective*, 2014; 10(7): 517-530. Doi:10.1016/S0927-538X (02)00040-9.

[10] Patricia C., S., Neophytos L.,Christos S. Savva, Sentiment, order imbalance, and co-movement: An examination of shocks to retail and institutional trading activity, *Europeans Financial Management private markets*,2016; **37**(2):1–22. Doi: 10.1016/j.pacfin.2016.02. 003.

[11] Shaari Anaghiz Saber, Yahya Hassah Yeganeh, Mehdi Sadidi and Benyamin Narhai, *Emotional decision making of investors, corporate governance and investment efficiency*, Financial Accounting Quarterly, 2018; **1**(1):139–153. Doi: 10.1016/S0165-1765(00)00403-1.

[12] Zanjirdar, M., Madahi, M., Khaleghi Kasbi, K. , Comparative analysis of sticky SGA costs and cost of goods sold: Evidence from Tehran Stock Exchange, *Management Science Letters*,2014;4(3): 521-526

[13] Zanjirdar, M., Moslehi Araghi, M., The impact of changes in uncertainty, unexpected earning of each share and positive or negative forecast of profit per share in different economic condition, *Quarterly Journal of Fiscal and Economic Policies*,2016;4(13): 55-76.

[14] Mohamadi, M., Zanjirdar, M., On the relationship between different types of institutional owners and accounting conservatism with cost stickiness, *Journal of Management Accounting and Auditing Knowledge*, 2018;7(28): 201-214

[15] Javaheri, M., Zanjirdar, M., The Relationship Between Profit Management and the Performance of Companies Studied in Bourse Securities of Tehran, *Productivity Management (Beyond Management)*,2017;11(42): 197-217