International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Statistical Modeling of S&P 500 Data Based on Time Lags of Apple Corporation

Ranju Karki[1], Doo Young Kim[2], Christ P. Tsokos[3]

[1, 3]University of South Florida, Department of Mathematics and Statistics,4202 East Fowler Ave, CMC365, Tampa, FL, 33620-5700, USA

[2]Sam Houston State University, Department of Mathematics and Statistics, Lee Drain Building 417B, Huntsville, TX, 77341, USA

*Corresponding Author.
Email addresses: ranjukarki[at]usf.edu (Ranju Karki), dkim[at]shsu.edu (Doo Young Kim), ctsokos[at]usf.edu (Christ P. Tsokos)

Abstract: ***Our objective is to select a company, AAPL, of the S&P 500 to be our leading company and proceed to predict the closing price of AAPL in conjunction with the other companies. We utilized the weighted five-day moving arc length as a measure of volatility and Self-Organized Maps to identify the appropriate cluster of companies that followed similar patterns with AAPL. We also developed predictive statistical models for the closing prices of the AAPL with Meta Platforms, Inc. (FB) and Microsoft Corporation (MSFT). One can select any company within the identified cluster to develop a predictive model using our procedures and methodologies.***

Keywords: Clustering, Statistical Modeling, SOMs, RVAR, Stock Price

1. Introduction

Since the beginning of the financial market such as stock markets, statisticians have been accumulating and analyzing all kinds of data from the markets so that we can have a better understanding on the dynamics of the market movement. Many studies have been conducted on the financial time-dependent information over time, and many statistical tools have been developed in order to predict the dynamic market movement. However, prediction of the market movement has been very challenging and difficult. There are many factors affecting the market movement, and there are non-linear relationships among those factors. Finding a better statistical model with a more accurate prediction of stock prices is desired, and it is our goal in the present study to accurately predict the stock prices. We are using a new statistical machine learning technique and a multivariate statistical modeling scheme to achieve our goals. An in-depth study of the stock market identified correlations between the closing prices of different stock companies. That implies the movement in the prices of one stock(leading stock) could impact on stock prices of one or more stock companies(lagging stocks). Therefore, we want to discover such price movement patterns of a leading stock so that the price movement pattern of other lagging stocks can predict more accurately. Our study mainly has driven by this key idea. We chose Apple corporation as the leading company, and we want to find out the other stock companies that are affected by the price changes of Apple corporation.

Apple is the world's largest technology company by revenue and the world's most valuable public company. Apple Inc. was founded by Steve Jobs and Stephen Wozniak in 1976. The headquarters of Apple Inc. is in Cupertino, California, and it belongs to the Information technology sector in *S&P* 500 index. On December 12, 1980, Apple went public at $22.00 per share. In August 2018, Apple became the first U.S. company worth a $1 trillion market cap. It increased to a $2 trillion valuation on August 19, 2020[8].

The *S&P* 500 index accounts for 80% of the market value of the U.S[1]. equities market. It is weighted by the market cap. Thus, the largest stock has a significant impact on the daily movement of the *S&P* 500 index and its long-term performance. Apple, Microsoft, Alphabet, Amazon, and Facebook Inc are the largest five companies of the *S&P* 500 index. They contribute to approximately 19% of the index market value. Only Apple covers about 7% of the *S&P* 500 total market cap. Hence, It has the highest influence on the index's movements. Therefore, we have chosen Apple Inc. as our leading company for this study.

2. Data Description

Over time, many indices have been created to measure the fluctuation in the stock market. The Dow Jones Industrial Average, Standard & Poor's 500, and Nasdaq Composite are the most popular and most followed indices in the United States' stock market. The Standard & Poor's 500, often abbreviated as the S&P 500, is the stock market index that consists of 500 stock companies with the largest market Capitalization (market cap) in the United States. These 500 companies are selected by a committee on the basis of certain criteria, such as market capitalization, the volume of shares traded on a stock exchange, and earning reports of the companies. These companies are also selected from various sectors so that they can reflect the diverse nature of the market. The S&P 500 index includes eleven business segments(sectors) such as communication services, consumer discretionary, consumer staples, energy, financials, healthcare, industries, information technology, materials, real estate, and utilities. In comparison to other indices, the S&P 500 index can represent the stock market of the U.S better as a whole, because it basically gives the idea of how the stock market is performing in general.

We obtained historical data of daily stock prices for 501 stock companies belonging to the S&P 500 index from

January 1, 2020, and December 31, 2020, from the Yahoo finance website. The data set includes opening price, closing price, high price, low price, and volume for all 501 companies. However, we are only focusing on closing price and volume for our study. Before the closing of a stock exchange, the last trading price for any stock on a given trading day is called the closing price. The sum of the number of shares for any stock traded on an exchange market in a given day is called the volume.

Figure 1, below, shows the structure of our data. Sector 1, Sector 2 ..., and sector 11 represent 11 business segments of the S& P 500 index. They are communication services(27 companies), consumer discretionary(63 companies), consumer staples(32 companies), energy(22 companies), financials(65 companies), healthcare(62 companies), industries(72 companies), information technology(74 companies), materials(28 companies), real estate(28 companies), and utilities(28 companies). Ticker 11, Ticker 12,..., Ticker $1n_1$are ticker symbols belonging to the communication sector, Ticker 21, Ticker 22,..., Ticker $2n_2$ are ticker symbols belonging to the consumer discretionary sector, ..., Ticker r1, Ticker r2,..., Ticker r_{n_r} are ticker symbols belonging to utilities sector respectively. Each stock company has given serial numbers based on their ticker position. For example, Ticker 11 is number 1, Ticker 12 is number 2,..., and Tickerr_{n_r} is number 501. There are 253 trading days in 2020. So, T equal to 253, the last day of trading. C_1,C_2,...,C_T represents the closing price of each ticker on day 1, day 2, ..., and day 253, respectively. V_1,V_2,...,V_T represents the volume of each ticker on day 1, day 2, ..., and day 253, respectively.

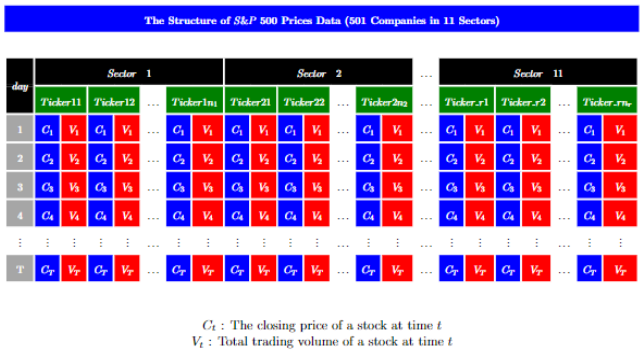

Figure 1: Structure of the Data.

1) Volatility of Stock Prices

The value of every security fluctuates in a different ratio. The price of some securities fluctuates dramatically over a short period of time, whereas other securities have steady value. The volatility is the measurement of such variation in the price of the security over time and measures the risk of a given security. The high volatility indicates that the security's value fluctuates dramatically, and hence it has a potentially high risk. The low volatility indicates that the value of the security changes at a steady pace, and hence it has low-risk [9]. Thus, by calculating the volatility of the stock prices, we can find information about the overall level of risk on the particular stock which helps us to construct our own portfolio and make constructive decisions. We have calculated five days' weighted volatility of all stock companies separately. We use the log-returns and arc-length of log-returns to calculate five-day-weighted volatility.

a) Log-Returns of Stock Prices

Any kind of statistical analysis of financial data is difficult because it is highly correlated and usually variance increases with time. So, it is always preferred to use the return of prices instead of raw prices in the financial analysis. If P_t is a discrete financial time series with equally spaced time points and P_t is the value of an asset at time t then the mathematical formula for log returns [11] is defined by

$$R_{Lt} = \ln\left(\frac{P_t}{P_{t-1}}\right) = \ln\left(1 + \frac{P_t - P_{t-1}}{P_{t-1}}\right) \quad (1)$$

We first calculated log-returns of the stock prices for all 501 companies by using above equation 1 separately.

Let C_{it} be the daily closing price of the stock i at time t, then the log-returns of closing price of the stock i at time t is given by below equation,

$$RC_{it} = C_{it} - C_{i(t-1)},\ where\ i = 1,2,\ldots,501,\ and\ t = 2,3,\ldots,253. \quad (2)$$

Let V_{it} be the daily volume of the stock i at time t, then the log-returns of volume of the stock i at time t is given by below equation,

$$RV_{it} = V_{it} - V_{i(t-1)} where\ i = 1,2,\ldots,501,\ and\ t = 2,3,\ldots,253. \quad (3)$$

For example, Figure 2 shows the time series plots of closing prices and log-returns of the closing price of the Apple Incorporation (AAPL) from January 2020 to December 2020. From the plot of Close price vs time, we can find that stock prices have a general increasing trend over time after day 50. From the plot of log return vs time, we can see that the mean of log return is almost zero and the volatility was larger around days 50 and 150.

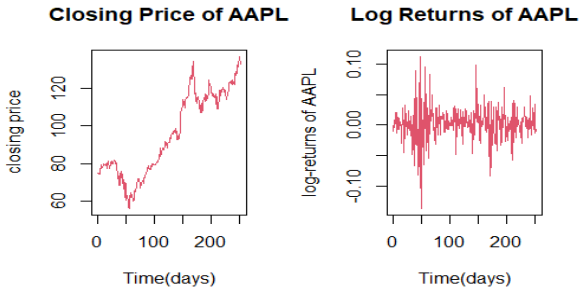

Figure 2: Closing Prices and Log-returns of Apple Incorporation (AAPL)

b) Arc-Length of Stock Prices

The arc-length is a statistical measurement for the volatility of datasets, that offers several advantages. The most important reason we use the arc-length is to measure the volatility of our time-dependent information and it can be measured in both univariate and multivariate time series for balanced or unbalanced data. Let P_t be the stock price corresponding to log price series lnP_t observed at times t = 1,2,...,n, then the sample arc length[10] of the log price series is defined below for sample points $(t, lnP_t)_{t=1}^{n}$:

$$\sum_{t=2}^{n} \sqrt{1 + {R_{Lt}}^2}, \quad (4)$$

whereR_{Lt} is the log-returns of stock price.

According to T.Wickramarachchi, F. Tunno[12], the difference between two sample arc lengths has an asymptotic normal distribution, and this result can be used to compare two series statistically in terms of volatility using arc-length as the measure.

In this study, we use a five-day sample moving arc-length in order to utilize the concept of the target lag, because the goal of this study is grouping stock companies having the same lag effect or volatility change over time. In another word, we want to cluster the stock companies based on the same amount of risk.

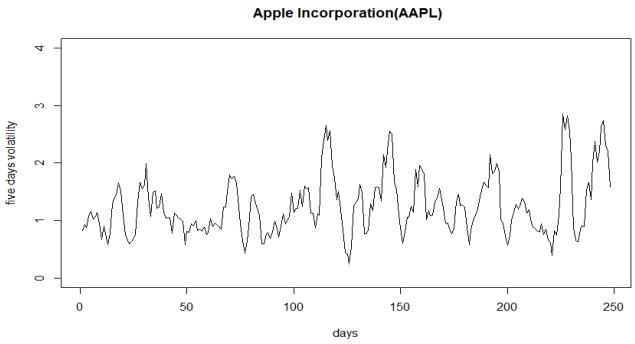

Figure 3: The Five-day Moving Volatility Plot of Apple Incorporation (AAPL).

Let RC_{it} the log-returns of closing price of the stock i at time t from equation 2 and RV_{it} log-returns of volume of the stock i at time t from equation 3. Then, we define the weighted five-day moving volatility of stock i at time t, λ_{it} as follows:

$$\lambda_{it} = \sum_{k=t-4}^{t} \sqrt{(RC_{it})^2 + (RV_{it})^2}.$$

In our example, t = 5,6,...,253, and $i = 1,2,\ldots,501$. After calculating the five days moving volatility of all companies, the dimension of our data is 501 by 248. Figure 3 above shows the time series plot for five-day moving volatility of Apple Corporation. From the figure, we can see the volatility of AAPL is higher on days between days 50 and 150. This shows when the price fluctuates more, the volatility of the stock price will be higher.

2) Clustering

Clustering is an unsupervised statistical machine learning technique that is used to group data sets into subsets on the basis of their specific features. The main purpose of clustering is to organize unlabeled data into the groups(sets) such that the distance among members in a group are minimized, and distances among the group members of different groups are maximized [13]. Clustering is statistically useful for partitioning any kind of unlabeled data types such as textual, binary, numerical, categorical, interval, relational, and so on. It is very useful to identify hidden patterns or internal structures of the data sets. In addition, clustering is also used for outlier detection or to find data points that are not part of any cluster.

We want to separate stock companies into different subgroups according to the same level of volatility they have in 2020. We have high dimensional data sets with 501 stock companies(features) and 248 observations(days) so that manual organization and annotation are not feasible. Clustering is going to help us to reach this goal. We have chosen Self Organizing Maps for clustering the data in this study because this method is useful to map multidimensional data onto lower-dimensional which reduces the complexity of problems for easy interpretation.

a) Self- Organizing Maps (SOMs)

Self-Organizing Maps were developed by Tuevo Kohonen in 1982. Self-Organizing Maps is a type of natural neural network technique which is inspired by a biological network. It is trained using an unsupervised learning algorithm which provides an easy visualization of high-dimensional data. Self-Organizing Maps is an effective clustering tool to convert complex, nonlinear statistical relationships among high-dimensional data sets into simple geometric statistical relationships on a low-dimensional grid (Tuevo Kohonen) [6]. The quality of SOMs depends on its initial conditions such as the number of iterations, sequence of training vectors, neighborhood function, learning rates, and weight of the map. The following are the stages that we must follow in implement of SOMs[5]:

Initial Stage:

Create k neurons (nodes) on a hexagonal grid and generate a p-dimensional initial random weight vector in each neuron. The value of p is determined by the dimension of the input vector. That is,

$$m(0) = (m_1(0), m_2(0), \ldots, m_k(0)).$$

In our case, the dimension of the input vector is 501.

Competitive Stage:

The first p-dimensional input vector x goes into the system and the winning neuron (node) is determined. The Euclidean distance is a statistical tool used to determine winning neuron (node) as shown in equation 6 below.

$$c^t(x) = arg\ \min_{k\epsilon\{1,2,\ldots,k\}} \|x - m_k(t)\|^2. \quad (6)$$

In equation 6 above, $c^t(x) \in \{1,2,..,k\}$is determined as the index of the best matching unit (BMU). The neuron with the minimum Euclidean distance wins and is called the winner (winning neuron). Suppose $c'(x) = i$, then i is the winner.

Cooperative Stage:

After finding the BMU, the weight vector of the winning neuron and its neighboring neuron are updated by using the neighborhood function. Usually, the Gaussian neighborhood function is preferred. That is,

$$h_{kc^t(x)}(t) = exp\left(\frac{-dist^2(k,c^t(x))}{2\sigma^2(t)}\right),$$

with

where $c^t(x)$is the winner and k is the number of nodes. The coefficients $\{\epsilon(t), t \geq 0\}$ are learning rate scaled valued and it is monotonically decreasing, and satisfies following conditions:

i) $0 < \epsilon(t) < 1$ ii) $\lim_{t\to\infty} \epsilon(t) \to \infty$ iii) $\lim_{t\to\infty} \epsilon^2(t) < \infty$

The weight vector of neuron(nodes) k is updated by using the following updating rule, iteratively,

$$m_k(t+1) = m_k(t) + \epsilon(t) h_{kc^t(x)}(t)(x - m_k(t)) \quad (9)$$

where *t* is the iteration step. As *t* increases, the number of updated neighborhood neurons is getting smaller.

In this study, we create SOMs with 10×10 hexagonal grid and circular neighborhood function. The training data include 501(companies) objects and the mean distance to the closest unit in the map is 6.945. This means companies have an average distance of 6.95 units to their closest cluster centroids. Another parameter of SOMs is iterations. We fixed the number of iterations as 1100 throughout the study. This means 1100 iterations the algorithm will execute. We can change the number of iterations if the convergence criteria are not fulfilled. We use 0.05 as the initial learning rate and it decreases until it gets to 0.01. It controls the size of how much weight is updated during each iteration. It controls the training process, and a higher learning rate is responsible for faster conversion. It decreases monotonically throughout the learning process.

There are two important plots we need to investigate [7]. These plots help to visualize the quality of generated SOM. One is a training progress plot, which shows the development of the average distance to the nearest cells on the map throughout the iterations process. As the SOM training iterations continue, the distance from each node's weights to the samples represented by that node will be declined. This distance reaches a minimum plateau after a certain number of iterations. If the curve is continually decreasing(not converge), more iterations are required. Another important plot is the 'counts plot' which gives us the information about the number of members that are mapped into each node on the map [7]. Large values in some nodes suggest that a larger map(grid) would be beneficial. Empty nodes indicate that the map size is too big for the number of samples. It also shows how many members are in each cluster by the color gradient.
We have applied SOMs with the parameter explained above. The data sets(501 stock companies) are divided into 100 clusters (10×10 hexagonal grid) based on the similarity between their five-day-moving volatility. However, we are only interested in the cluster to which Apple Incorporation (AAPL) belongs because Apple Incorporation (AAPL) is our leading company. The reason behind choosing a leading company is that we want to see a relationship between the leading company with all other companies where all other companies are lagging companies. The following subsections present the results and graphical representations of SOMs at lag zero through lag five for Apple Incorporation (AAPL) as the leading company. The lag-n version data set are created from the original data set by moving the series values forward *n* period. we use the below series for lag *l* clustering: $(t, \lambda_{Lt})_{t=5}^{253-l}$ for leading stock *L* and $(t, \lambda_{Ot})_{t=5+l}^{253}$ for all other lagging securities *O*. The detailed series are explained below.

3) Stock Companies with Apple Incorporation at Lag 0:
The original data set has dimension 501 by 248, and we have to create data set for Apple Incorporation as the leading company at zero lag. This means there is no time gap between leading and lagging stock companies. Hence, if the price of apple change, then the price of all other lagging companies which share the same cluster with AAPL will change on the same day. We applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit on the map is 7.03.

The following figure 4 are the graphical representations of self-organizing maps at lag zero for Apple Incorporation (AAPL) as the leading company. The training progress plot, figure 4(left) shows the training progress is gradually decreasing until 900 iterations (nearly) and then rapidly decreasing after that. There is not a significant improvement from (nearly) 950 iterations. This means that the clustering system is converging. Hence, we do not need to increase the number of iterations. The count plot figure 4(right) shows our 100 clusters. Each cluster has at least one member. The color intensity increases towards blue when the number of companies increases in the cluster. The darkest blue color cluster has a maximum number of members, 11 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag 0. In this study, all other clusters are not considered.

In Figure 4, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the O'Reilly Automotive, Inc. (ORLY), the Williams Companies, Inc. (WMB), Exxon Mobil Corporation (XOM), American International Group, Inc. (AIG), Principal Financial Group, Inc. (PFG) are together with Apple Inc. (AAPL) at lag 0. Thus, if the price of AAPL changes, then we expect there will be some changes in the price of these companies on the same day.

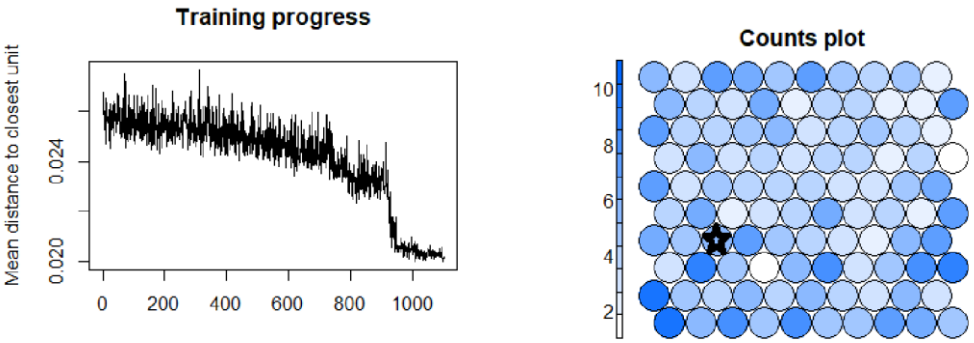

Figure 4: Training Progress and Counts Plot at Lag 0 of Stock Securities When Apple Incorporation is the Leading Company.

4) Stock Companies with Apple Incorporation at Lag 1:
The original data set has dimension 501 by 248, and we have to create data set for Apple Incorporation as the leading company at one lag. This means there is one day time gap between leading and lagging stock companies. Hence, if the price of apple change, then the price of all other lagging companies which share the same cluster with AAPL will change on the next day. The new data for lag 1 has dimensions 501 by 247. The data were created by using the following code in Rstudio: *rbind*(*data.zero*[1 : 343,2 : 248],*data.zero*[344,1 : 247],*data.zero*[345 : 501,2 : 248]) , where data. zero is the name of the original data.

We applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit in the map is 6.996.

The following Figure 5 are the graphical representations of self-organizing maps at lag one for Apple Incorporation (AAPL) as the leading company. The training progress plot, Figure 5(left) shows the training progress was started converging from (nearly) 1000 iterations. The count plot Figure 5(right) shows our 100 clusters. The darkest blue color cluster has a maximum number of members, 12 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag 1.

In Figure 5, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the Meta Platforms, Inc. (FB), Alphabet Inc. (GOOG), Alphabet Inc. (GOOGL), Amazon.com, Inc. (AMZN), Microsoft Corporation (MSFT), and Apple Inc.(AAPL) are in the same node (cluster). Thus, if the price of AAPL changes, then we are expecting that there will be some changes in the price of these companies on the next day.

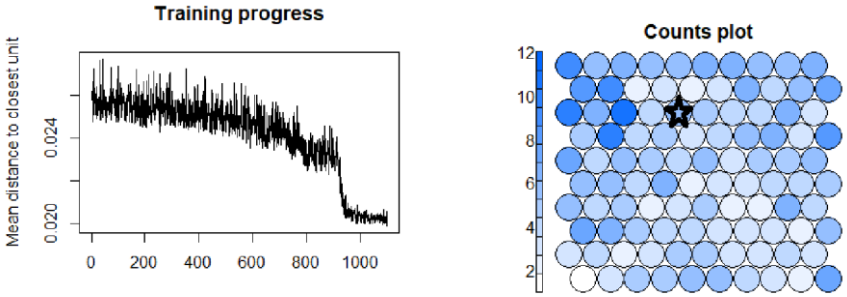

Figure 5: Training Progress and Counts Plot at Lag 1 of Stock Securities When Apple Incorporation is the Leading Company

5) Stock Companies with Apple Incorporation at Lag 2:
First, we have to create data set for Apple Incorporation as leading companies at two lag. This means there is two day time gap between leading and lagging stock companies. Hence, if the price of apple change then the price of all other lagging companies which share the same cluster with AAPL will change on the second day. The new data for lag 2 has dimensions 501 by 246. The data were created by using the following code in Rstudio: *rbind*(*data.zero*[1 : 343,3 : 248],*data.zero*[344,1 : 246],*data.zero*[345 : 501,3 : 248]) , where data.zero is the name of the original data.

As above, we applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit on the map is 6.945.

The following Figure 6 are the graphical representations of self-organizing maps at lag two for Apple Incorporation (AAPL) as the leading companies. The training progress plot, Figure 6(left) shows the training progress was started converging from (nearly) 1000 iterations. In the count plot Figure 6(right), the darkest blue color cluster has a maximum number of members, 12 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag two.

In Figure 6, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the Meta Platforms, Inc. (FB), Ford Motor Company (F), American International Group, Inc. (AIG), Advanced Micro Devices, Inc. (AMD), Microsoft Corporation (MSFT), NVIDIA Corporation (NVDA), and Apple Inc.(AAPL) are in the same node (cluster). If the price of AAPL changes today, then we are expecting that there will be some changes in the price of these companies on the second day.

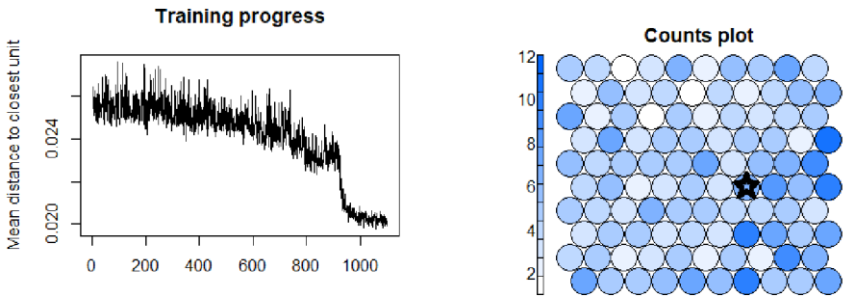

Figure 6: Training Progress and Counts Plot at Lag 2 of Stock Securities When Apple Incorporation is the Leading Company.

6) Stock Companies with Apple Incorporation at Lag 3: First, we have to create data set for Apple Incorporation as leading companies at three lag. This means there is three day time gap between leading and lagging stock companies. Hence, if the price of apple change then the price of all other lagging companies which share the same cluster with AAPL will change on the third day. The new data for lag 3 has dimensions 501 by 245. The data were created by using the following code in Rstudio: *rbind*(*data.zero*[1 : 343,4 : 248],*data.zero*[344,1 : 245],*data.zero*[345 : 501,4 : 248]) , where data.zero is the name of the original data.

As above, we applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit in the map is 6.99.

The following Figure 7 are the graphical representations of self-organizing maps at lag three for Apple Incorporation (AAPL) as the leading companies. The training progress plot, Figure 7(left) shows the training progress was started converging from (nearly) 1000 iterations. In the count plot Figure 7(right), the darkest blue color cluster has a maximum number of members, 15 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag three.

In Figure 7, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the Meta Platforms, Inc. (FB), Alphabet Inc. (GOOG), Alphabet Inc. (GOOGL), Amazon.com, Inc. (AMZN), Starbucks Corporation (SBUX), Lam Research Corporation (LRCX), Microsoft Corporation (MSFT), and Apple Inc. (AAPL) are in same node (cluster). If the price of AAPL changes today, then we are expecting that there will be some changes in the price of these companies on the third day.

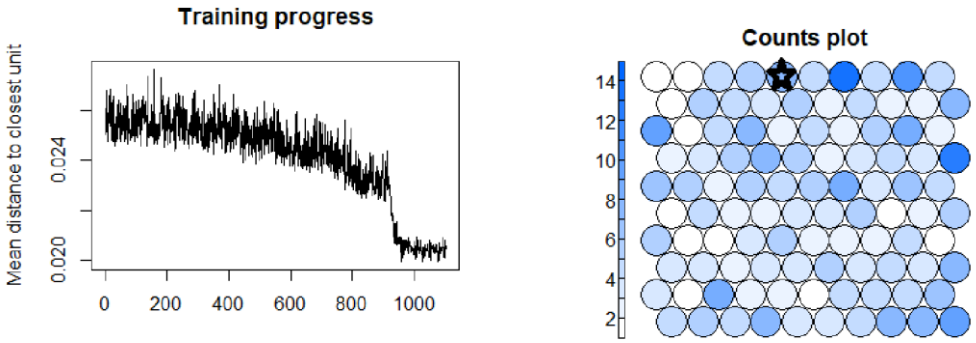

Figure 7: Training Progress and Counts Plot at Lag 3 of Stock Securities When Apple Incorporation is the Leading Company

7) Stock Companies with Apple Incorporation at Lag 4: First, we have to create data set for Apple Incorporation as the leading company at four lags. This means there is four day time gap between leading and lagging stock companies. Hence, if the price of apple change, then the price of all other lagging companies which share the same cluster with AAPL will change on the fourth day. The new data for lag 4 has dimensions 501 by 244. The data were created by using the following code in Rstudio: *rbind*(*data.zero*[1 : 343,5 : 248],*data.zero*[344,1 : 244],*data.zero*[345 : 501,5 : 248]) , where data.zero is the name of the original data.

As above, we applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit on the map is 6.952.

The following Figure 8 are the graphical representations of self-organizing maps at lag four for Apple Incorporation (AAPL) as the leading companies. The training progress plot, Figure 8(left) shows the training progress was started converging from 1000 iterations. In the count plot Figure 8(right), the darkest blue color cluster has a maximum number of members, 11 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag four.

In Figure 8, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the AutoZone, Inc. (AZO), Tapestry, Inc. (TPR), Advanced Micro Devices, Inc. (AMD), Microsoft Corporation (MSFT), NVIDIA Corporation (NVDA), and Apple Inc. (AAPL) are in same node (cluster). If the price of AAPL changes today, then we are expecting that there will be some changes in the price of these companies on the fourth day.

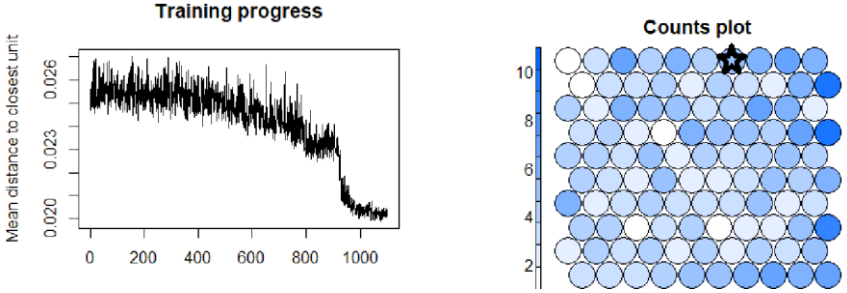

Figure 8: Training Progress and Counts Plot at Lag 4 of Stock Securities When Apple Incorporation is the Leading Company.

8) Stock Companies with Apple Incorporation at Lag 5: First, we have to create data set for Apple Incorporation as the leading company at five lags. This means there is five day time gap between leading and lagging stock companies. Hence, if the price of apple change, then the price of all other lagging companies which share the same cluster with AAPL will change on the fifth day. The new data for lag 5 has dimensions 501 by 243. The data were created by using the following code in Rstudio: *rbind*(*data.zero*[1 : 343,6 : 248],*data.zero*[344,1 : 243], *data.zero*[345 : 501,6 : 248]) , where data.zero is the name of the original data.

As above, we applied SOMs with parameters 10×10 hexagonal grid, 1100 iterations, and 0.05 as the initial learning rate. The training data includes 501 objects and the mean distance to the closest unit on the map is 6.949.

The following Figure 9 are the graphical representations of self-organizing maps at lag five for Apple Incorporation (AAPL) as the leading company. The training progress plot, Figure 9(left) shows the training progress was started converging from (nearly) 1000 iterations. In the count plot Figure 9(right), the darkest blue color cluster has a maximum number of members, 12 stock companies. The lightest blue color cluster has a minimum number of members, one stock company. We are only interested in those stock companies which has the same level of volatility as Apple Incorporation in 2020 at lag five.

In Figure 9, the Apple Incorporation belongs to the node(cluster) with the black star. In the cluster, the Ford Motor Company (F), Advanced Micro Devices, Inc. (AMD), Microsoft Corporation (MSFT), Service Now, Inc. (NOW), NVIDIA Corporation (NVDA), and Apple Inc. (AAPL) are in the same node (cluster). If the price of AAPL changes today, then we are expecting that there will be some changes in the price of these companies on the fifth day.

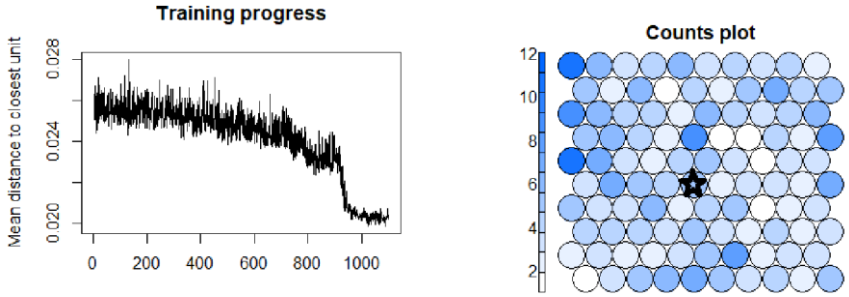

Figure 9: Training Progress and Counts Plot at Lag 5 of Stock Securities When Apple Incorporation is the Leading Company.

Table 1 gives us information about the list of securities in the same cluster with AAPL. At the beginning of data processing, we have given the serial number to each stock companies. These serial numbers are the identification of each ticker. After the SOM process, we identify the name of companies belongs to each lag by using their serial number (Number).

Table 1: List of Securities Related to AAPL

No.	Lag zero		Lag one		Lag two		Lag three		Lag four		Lag five	
	Number	Ticker	Number	Ticker	Number	Ticker	Number	Ticker	Number	Ticker	Number	Ticker
1	70	ORLY	9	FB	9	FB	9	FB	31	AZO	47	F
2	143	WMB	12	GOOG	47	F	12	GOOG	81	TPR	352	AMD
3	144	XOM	13	GOOGL	146	AIG	13	GOOGL	352	AMD	389	MSFT
4	146	AIG	29	AMZN	352	AMD	29	AMZN	389	MSFT	393	NOW
5	190	PFG	389	MSFT	389	MSFT	78	SBUX	395	NVDA	395	NVDA
6					395	NVDA	385	LRCX				
7							389	MSFT				

3. Introduction to Vector Autoregressive Models

The vector autoregression (VARs) model was first proposed by Christopher Sims in 1980. It is a natural extension of the univariate autoregressive model to dynamic multivariate time series. A univariate autoregression model is a single variable linear model with a single equation in which the current value of a variable is described by its own lagged values. The VAR model is a *n*-variable linear model with *n*-equations in which the current value of a variable is described by its own previous values added with current and past values of the remaining $(n-1)$ variables. Because of this simple framework, the VAR model has become one of the most successful, reliable, and convenient models for the analysis of multivariate time series. The VAR model is useful for forecasting and exploring the dynamic behavior of financial and economic time series. It is also used for structural interference and policy analysis [6].

Let $Y_t = (y_{1t}, y_{2t}, \dots, y_{nt})'$, denote a time series vector with length n. The vector autoregressive model with lag p is defined by

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + \varepsilon_t, \quad p > 0, t = 1, \dots, T, \tag{10}$$

where c is a n-dimensional vector, the Π_i's are $(n \times n)$ coefficient matrices, and ε_t is a sequence of $(n \times 1)$ independent white noise vector process with zero mean and time invariant covariance matrix Σ.

The general form of the VAR(p) is defined as

$$Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \cdots + \Pi_p Y_{t-p} + \Phi D_t + G X_t + \varepsilon_t, \quad (11)$$

where Φ and G are parameter matrices, Y_t is a $(m \times$ *l*matrix of exogenous variables and *Dt* is a *l*×*l*matrix of deterministic components.

In this paper, we shall investigate the best model for the financial time series which are in the same cluster with Apple Incorporation (AAPL) to forecast future price changes of those companies based on AAPL. We use a restricted VAR model for modeling the data. The VAR (p) model includes all the predictors which may not be significantly important for the model whereas the restricted VAR model re-estimated each equation according to the t-value of the regressors [2]. The regressors whose absolute value of t is greater than the threshold set value only stay in the model, and we have to assign the threshold value. We use one as a threshold value. The forecasting with a VAR model gives us a nearly approximate result if we choose the correct number of true lags for VAR. So, the first step in VAR is the selection of true lag values.

3.1 Lag Length Selection

The lag selection criteria are the most commonly used strategy to determine the maximum number of lags $(p = 0, \ldots, p_{max})$ that should be included in VAR(p) model. The Akaike (AIC), Schwarz-Bayesian (BIC) and Hannan-Quinn (HQ) are three of the most commonly used lag value selection criteria [6]:

$$AIC(p) = \ln\left|\hat{\sum}(p)\right| + \frac{2}{T}(pn^2)$$

$$BIC(p) = \ln\left|\hat{\sum}(p)\right| + \frac{\ln T}{T}(pn^2)$$

$$HQ(p) = \ln\left|\hat{\sum}(p)\right| + \frac{2 ln ln T}{T}(pn^2)$$

where, T is the sample size and $\hat{\sum}(p) = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_t'$ is the residual covariance matrix. The BIC and HQ criteria estimate the order consistently under general conditions if the order of lag p is less than or equal to p_{max}, whereas the AIC criterion asymptotically overestimates the order with positive probability.

In this study, we investigate the best model for the financial time series which are in the same cluster with Apple Incorporation (AAPL). We use a restricted VAR model for modeling the data. The VAR (p) model includes all the predictors which may not be significantly important for the model whereas the restricted VAR model re-estimated each equation according to the t-value of the regressors [3].

The regressors whose absolute value of t is greater than the threshold set value only stay in the model, and we have to assign the threshold value. We use one as a threshold value. The forecasting with a VAR model gives us a nearly approximate result if we choose the correct number of true lags for VAR. So, the first step in VAR is the selection of true lag values.

According to [4], if we choose a higher order lag than the true lag, then the value of the mean square forecast errors from the VAR model will increase, and if we choose a lower order lag than the true lag, then the value of the autocorrelation will increase. In our case, we choose the maximum order of lag for the model according to their lag relationship with Apple.

We also calculate the eigenvalues (unit-root structure) of the RVAR(p) model for checking the stability of RVAR(p) model. If the moduli of the eigenvalues from the companion matrix of the fitted model are less than one, then the RVAR model is stationary.

Now, we present the modeling process of the Meta Platforms, Inc. (FB) and the Microsoft Corporation (MSFT) securities when the Apple Incorporation (AAPL) is considered as a leading stock by using a restricted vector Auto regressive Model. The reason behind choosing these companies are they are both U.S biggest companies of S&P 500 index. In addition, FB was presented in lag 1,lag 2 and lag 3 continuously with Apple whereas MSFT was presented in all five lags except lag 0. From the clustering section, table 1, we have shown, in terms of volatility on the prices they are connected to each other.

3.2. Modeling of Meta Platforms, Inc. (FB) with Apple Incorporation (AAPL) as Leading Stock

We use the closing prices of the Meta Platforms, Inc. (FB) securities for our model. From the clustering section, we know that the stock price of the Meta Platforms, Inc. (FB) company's securities fluctuates the day after the fluctuation in the price of Apple Incorporation occurred. There are also effects on the second and third day. Therefore, we choose three as the maximum lag value in the RVAR.

Each variable is a linear function of the past values of itself and the past values of all the other variables and an error term in VAR (p) model. Table 2 and table 3 summarize the ordinary least squares (OLS) results of the fitted RVAR (3) for AAPL and FB, respectively. In 2 and table 3, the T value is the estimate divided by its standard error. The standard error is an estimate of the standard deviation of the estimates. The regressors whose absolute value of *t* is greater than the threshold set value only stay in the model, and the assign threshold value is one for RVAR(3) model.

In Table 2, the value of the estimate for each independent variable gives us how much the dependent variable (closing price of the securities of AAPL) is expected to increase when that independent variable increases by one, holding all the other independent variables constant. The significant independent variables, in this case, are AAPL at lag 1 and AAPL at lag 2.

Table 2: Regression Results for AAPL

	Estimate	Std.Error	T value	Pr(>|t|)
AAPL.l1	0.82171	0.06249	13.150	<2e−16
AAPL.l2	0.18097	0.06265	2.889	0.00421

In Table 3, the value of the estimate for each independent variable gives us how much the dependent variable (closing price of the securities of FB) is expected to increase when that independent variable increases by one, holding all the other independent variables constant. The significant independent variables, in this case, are FB at lag 1, AAPL at lag 2 and AAPL at lag 3.

Table 3: Regression Results for FB.

	Estimate	Std.Error	T value	Pr(>|t|)
FB.l1	0.92707	0.03092	29.982	<2e−16
AAPL.l2	0.30613	0.15804	1.937	0.0539
AAPL.l3	-0.19274	0.15328	-1.257	0.2098
const	6.51826	2.91227	2.238	0.0261

Table 4 shows the restriction matrix of the fitted RVAR (3) for AAPL and FB respectively. The VAR (3) model includes all the predictors which are not be significantly important for the model so we use the restricted VAR (3) model which re-estimated each equation according to the t-value of the regressors and improved the goodness of the fits of our models. In Table 4, *0* indicates the restricted variables in the RVAR (3) and *1* indicates included variables in the RVAR (3). In other words, the independent variables FB at lag 1, FB at lag 2, AAPL at lag 3, FB at lag 3, and constants are removed from the fitted RVAR (3) model for AAPL. The fitted RVAR (3) model for AAPL is given in equation 12. In addition, the independent variables AAPL at lag 1, FB at lag 2, and FB at lag 3 are removed from the fitted RVAR (3) model for FB. The fitted RVAR (3) model for FB is given in equation 13.

Table 4: Restriction Matrix of the Fitted RVAR (3).

	AAPL.l1	FB.l1	AAPL.l2	FB.l2	AAPL.l3	FB.l3	const
AAPL	1	0	1	0	0	0	0
FB	0	1	1	0	1	0	1

The fitted RVAR (3) model for FB when AAPL as the leading company is

$$\hat{C}_{A,t} = 0.82C_{A,t-1} + 0.18C_{A,t-2} \quad (12)$$

and

$$\hat{C}_{F,t} = 6.52 + 0.93C_{F,t-1} + 0.31C_{A,t-2} - 0.19C_{A,t-3} \quad (13)$$

where, $\hat{C}_{A,t*}$ and $\hat{C}_{F,t*}$ represent the closing price of the securities of AAPL and FB at time t^*, respectively. The t^* means different time lags presented on above equations 12 and 13.

Equation 12, is the fitted RVAR (3) model that estimates the prices of AAPL at time *t*. Equation 13, estimates the price of FB at time *t* in which we are interested. According to equation 13, when there are no changes in the closing price of all other factors in this model, then the following results can be derived:

- If the price of FB increased by 1 dollar yesterday, then we are expecting that the change in the price of FB today will be increased by 93 cents.
- If the price of AAPL increased by 1 dollar the day before yesterday, then we are expecting that the change in the price of FB today will be increased by 31 cents.
- Three days ago, if the price of AAPL is increased by 1 dollar, then we are expecting that the change in the price of FB today will be decreased by 19 cents.

The accuracy of our fitted models are tested by using the coefficient of determination statistic, R^2 and $R^2 - adjusted$. The R^2 is used to measure the goodness-of-fit of a statistical model. It estimates the proportion of variation in the response variable explained by the model attributable risk factors. The higher the R^2 statistic the better the goodness-of-fit of a statistical model. In general, the R^2 is defined by

$$R^2 = 1 - \frac{SSR}{SST},$$

where SSR is the regression sum of squares representing the variation explained by the proposed models, and SST is called the total sum of squares is the proportional to the sample variance. The SST is equals to the sum of SSR and SSE where SSE is the residual sum of squares representing the variation in the proposed model left unexplained. Generally, the R^2 has the problem of increasing by increasing the number of parameters or predictors in the model. Therefore, it is recommended that we estimate the R^2 along with the $R^2 - adjusted$ to adjust for the degrees of freedom of the model and is given by

$$R^2 - adjusted = 1 - \frac{SSR/(n-p)}{SST/(n-1)},$$

where, n is the total number of observations and p is the number of predictor variables. Our proposed statistical model given in equation 13 resulted in equal amount of R^2 and $R^2 - adjusted$ of 99.93%. This means the proposed model explains 99.93% variation in the response variable (i.e., the closing price of the securities of Facebook). It is a very good quality model. The R-squared does not always indicate that the model has a good fit. Therefore, we have further investigated the visual diagram of fitted models.

In Figures, 10, the diagram of the fitted line plot displays the relationship between fit and residuals for the FB model present in equation 13. The standardized residuals plot in Figure 10(bottom half) shows the residuals are distributed randomly around zero which shows the residuals from the model have a constant variance. In the diagrams of the autocorrelation plot(bottom left) and partial autocorrelation (bottom right), we can see that all the spikes of residuals of the model are in between 95% confidence bounds. There is not significant autocorrelation in the residuals of the model which indicates no loss of information by the fitted RVAR (3) model that estimates the prices of FB at time *t* given in the Equation 13. In addition, all moduli of eigenvalues from the companion matrix are less than one. Therefore, the fitted RVAR (3) model given in Equation 13 is stationary. Its statistical properties such as mean, variance, and autocovariance do not change over time. Hence, the conclusions from the analysis of stationary series are reliable.

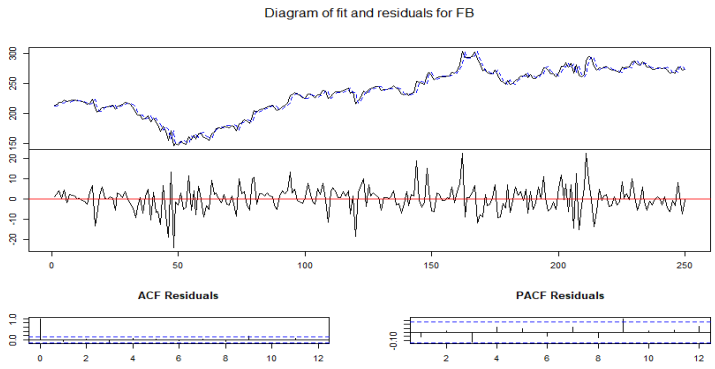

Figure 10: FB Model - Diagram of Fit, Residuals, ACF and PACF

In Figures 11, the diagram of the fitted line plot displays the relationship between fit and residuals for the AAPL model present in equation 12. In this diagram also the standardized residuals plot in Figure 11(bottom half) shows the residuals are distributed randomly around zero. In the diagrams of the autocorrelation plot(bottom left) and partial autocorrelation (bottom right), all the spikes of residuals of the model are in between 95% confidence bounds. There is not significant autocorrelation in the residuals of the model which indicates no loss of information by the fitted RVAR (3) model that estimates the prices of AAPL at time *t* given in the Equation 12 when FB is lagging company and it is also stationary.

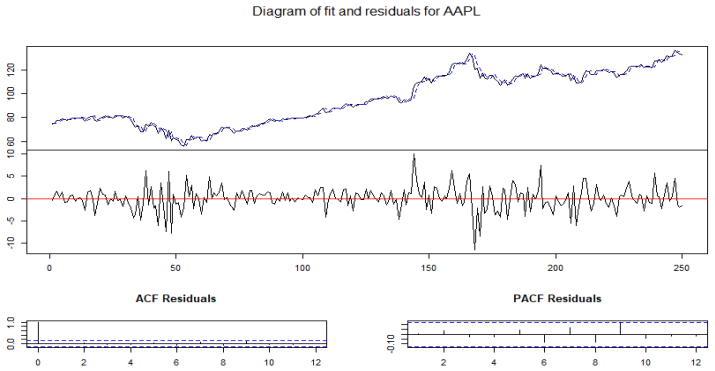

Figure 11: AAPL Model - Diagram of Fit, Residuals, ACF and PACF

3.3 Modeling of Microsoft Corporation (MSFT) with the Apple Incorporation (AAPL) as a Leading Stock

The Microsoft Corporation (MSFT) and the Apple Incorporation (AAPL) are both U.S giant technology companies. In the S&P 500, the Apple stock and Microsoft stock are the top two stocks by market weight. We would like to investigate the connection between these two companies in terms of volatility on the prices. We use a restricted VAR model for modeling the closing prices of the Microsoft Corporation (MSFT) securities when Apple Incorporation (AAPL) is a leading company. From the clustering section, Table 1, we know that the stock prices of the MSFT fluctuate the days after the fluctuation of the price of Apple Incorporation occurred. It also has effects on the second day, third day, fourth day, and fifth day. Therefore, we choose five as the maximum lag value in the RVAR.

Tables 5 and 6 above summarize the ordinary least squares (OLS) results of the fitted RVAR (5) for AAPL and MSFT, respectively. In 5 and 6, the regressors whose absolute value of *t* (T value) is greater than the threshold value which is one only stay in the model.

In Table 5, the value of the estimate for each independent variable gives us how much the dependent variable (closing price of the securities of AAPL when MSFT is a lagging company) is expected to increase when that independent variable increases by one, holding all the other independent variables constant. The significant independent variables are AAPL at lag 1, MSFT at lag 1, MSFT at lag 2, AAPL at lag 3, MSFT at lag 3, MSFT at lag 4, AAPL at lag 5, and MSFT at lag 5.

Table 5: Regression Results for AAPL

	Estimate	Std.Error	T value	Pr(>|*t*|)
AAPL.l1	0.96830	0.07290	13.282	<2*e*−16
MSFT.l1	-0.11501	0.04869	-2.362	0.0190
MSFT.l2	0.17440	0.04340	4.019	7.84*e*−05
AAPL.l3	-0.12232	0.09700	-1.261	0.2085
MSFT.l3	0.07078	0.06152	1.151	0.2511
MSFT.l4	-0.05245	0.04319	-1.214	0.2258
AAPL.l5	0.15056	0.07454	2.020	0.0445
MSFT.l5	-0.07455	0.04899	-1.522	0.1294

In Table 6, the value of the estimate for each independent variable gives us how much the dependent variable (closing price of the securities of MSFT when AAPL as a leading

company) is expected to increase when that independent variable increases by one, holding all the other independent variables constant. The significant independent variables are AAPL at lag 1, MSFT at lag 1, MSFT at lag 2, AAPL at lag 3, MSFT at lag 3, AAPL at lag 4, MSFT at lag 4 and constant.

Table 6: Regression Results for MSFT.

	Estimate	Std.Error	T value	Pr(>|t|)
AAPL.l1	0.21385	0.13278	1.611	0.1086
MSFT.l1	0.56487	0.08902	6.345	1.10e−09
MSFT.l2	0.39473	0.07640	5.167	5.01e−07
AAPL.l3	-0.44368	0.20284	-2.187	0.0297
MSFT.l3	0.19709	0.11201	1.759	0.0798
AAPL.l4	0.27180	0.17478	1.555	0.1212
MSFT.l4	-0.21358	0.09867	-2.164	0.0314
const	7.26867	3.86540	1.880	0.0613

Table 7 shows the restriction matrix of the fitted RVAR (5) for AAPL and MSFT respectively. The VAR (5) model includes all the predictors which are not be significantly important for the model so we use the restricted VAR (5) model which re-estimated each equation according to the t-value of the regressors and improved the goodness of the fits of our models. In Table 7, 0 indicates the restricted variables in the RVAR (5) and 1 indicates included variables in the RVAR (5). The independent variables AAPL at lag 2, AAPL at lag 4, and constants are removed from the fitted RVAR (5) model for AAPL. The fitted RVAR (5) model for AAPL is given in equation 14. In addition, the independent variables AAPL at lag 2, AAPL at lag 5, and MSFT at lag 5 are removed from the fitted RVAR (5) model for MSFT. The fitted RVAR (5) model for MSFT is given by equation 15.

Table 7: Restriction Matrix of the Fitted RVAR (5).

	AAPL.l1	MSFT.l1	AAPL.l2	MSFT.l2	AAPL.l3	MSFT.l3	AAPL.l4	MSFT.l4	AAPL.l5	MSFT.l5	const
AAPL	1	1	0	1	1	1	0	1	1	1	0
MSFT	1	1	0	1	1	1	1	1	0	0	1

The fitted RVAR (5) model for MSFT when AAPL as leading company is

$$\hat{C}_{A,t} = 0.97C_{A,t-1} - 0.12C_{M,t-1} + 0.17C_{M,t-2} - 0.12C_{A,t-3} + 0.07C_{M,t-3} - 0.05C_{M,t-4} + 0.15C_{A,t-5} - 0.07C_{M,t-5} \quad (14)$$

and

$$\hat{C}_{M,t} = 7.27 + 0.21C_{A,t-1} + 0.56C_{M,t-1} + 0.39C_{M,t-2} - 0.44C_{A,t-3} + 0.20C_{M,t-3} + 0.27C_{A,t-4} - 0.21C_{M,t-4}, \quad (15)$$

where, $\hat{C}_{A,t*}$ and $\hat{C}_{M,t*}$ stand for the closing price of securities of AAPL and MSFT at time t^*, respectively. The t^* means different time lags presented on above equations 14 and 15.

Equation 14, is the fitted RVAR (5) model that estimates the price of AAPL at time t. Equation 15, estimates the price of MSFT at time *t* in which we are interested. According to 15, when there are no changes in the closing price of all other factors in this model, then the following result can be derived:

- If the price of MSFT increased by 1 dollar yesterday, then we are expecting that the change in the price of MSFT today will be increased by 56 cents.
- If the price of MSFT increased by 1 dollar two days ago, then we are expecting that the change in the price of MSFT today will be increased by 39 cents.
- If the price of MSFT increased by 1 dollar three days ago, then we are expecting that the change in the price of MSFT today will be increased by 20 cents.
- If the price of MSFT increased by 1 dollar four days ago, then we are expecting that the change in the price of MSFT today will be decreased by 20 cents.
- If the price of AAPL increased by 1 dollar yesterday, then we are expecting that the change in the price of MSFT today will be increased by 21 cents.
- Three days ago, if the price of AAPL increased by 1 dollar, then we are expecting that the change in the price of MSFT today will be decreased by 44 cents.
- Four days ago, if the price of AAPL increased by 1 dollar, then we are expecting that the change in the price of MSFT today will be increased by 27 cents.

The accuracy of our fitted models are tested by using the coefficient of determination statistic, R^2 and $R^2 - adjusted$. Our proposed statistical model given in equation 15 resulted in equal amount of R^2 and $R^2 - adjusted$ of 99.95%. This means the proposed model explains 99.95% variation in the response variable (i.e. The closing price of the securities of MSFT). It is a very good quality model. In addition, we have further investigated the visual diagram of fitted models.

In Figures 12, the diagram of fitted line plot displays the relationship between fit and residuals for the MSFT model present in equation 15. The standardized residuals plot in Figure 12(bottom half) shows the residuals are distributed randomly around zero which shows the residuals from the model have a constant variance. In the diagrams of the autocorrelation plot (bottom left) and partial autocorrelation (bottom right), we can see that there is not significant autocorrelation in the residuals of the model which indicates no loss of information by the fitted RVAR (5) model that estimates the prices of MSFT at time *t* given in the Equation 15 when AAPL is leading company.

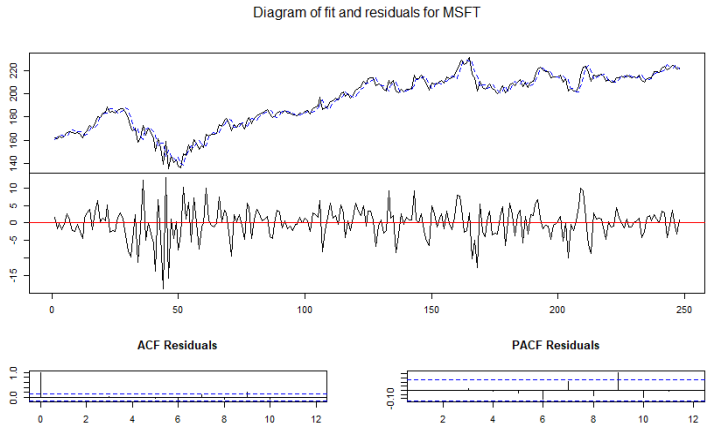

Figure 12: MSFT Model - Diagram of Fit, Residuals, ACF and PACF

In addition, all moduli of eigenvalues from the companion matrix are less than one. Therefore, the fitted RVAR (5) model given in Equation 15 is stationary. Its statistical properties such as mean, variance, and autocovariance does not change over time. Hence, the inference from this analysis of will be reliable.

In Figures 13, the diagram of fitted line plot displays the relationship between fit and residuals for the AAPL model present in equation 14. In this diagram also the standardized residuals plot in the Figure 13(bottom half) shows the residuals are distributed randomly around zero. In the diagrams of the autocorrelation plot (bottom left) and partial autocorrelation(bottom right), there is not significant autocorrelation in the residuals of the model since all spikes of residuals of the model are in between 95% confidence bounds. which indicates no loss of information by the fitted RVAR (5) model that estimates the prices of AAPL at time *t* given in the Equation 14 when MSFT is a lagging company, and it is also stationary.

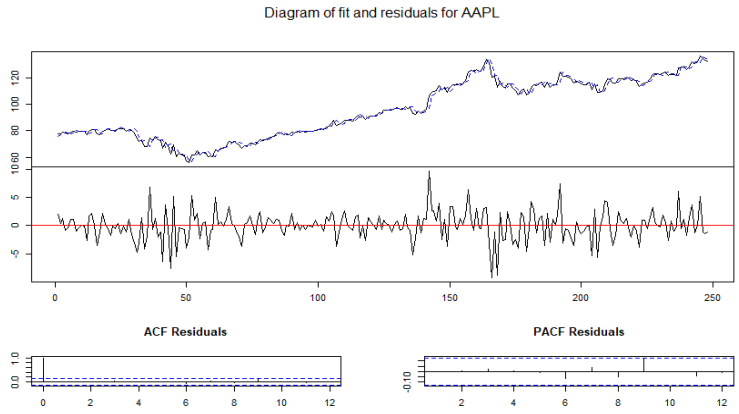

Figure 13: AAPL Model - Diagram of Fit, Residuals, ACF and PACF

4. Conclusions

This study focused on two main topics. First, the clustering of the closing price of securities of the S&P 500 when Apple Incorporation (AAPL) is the leading company. We calculated log returns, and the weighted five-day moving volatility of the closing price of all stock companies presented in the S&P 500 on 2020. We used this manipulated data set for clustering using self-organizing maps. The following are the results we established from clustering when AAPL is the leading company.

- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of ORLY, WMB, XOM, AIG, and PFG at the same day because they share same cluster 33 with AAPL at lag 0.
- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of FB, GOOG, GOOGL, AMZN, and MSFT on the next day because they share same cluster 75 with AAPL at lag 1.
- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of FB, F, AIG, AMD, MSFT, NVDA two days later because they share same cluster 47 with AAPL at lag 2.
- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of FB, GOOG, GOOGL, AMZN, SBUX, LRCX, and MSFT three days later because they share same cluster 95 with AAPL at lag 3.
- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of AZO, TPR, AMD, MSFT, and NVDA four days later because they share same cluster 97 with AAPL at lag 4.

- If the AAPL return changes today, then we can expect that there will be some price changes on the returns of F, AMD, MSFT, NOW, NVDA five days later because they share same cluster 45 with AAPL at lag 5.

Second, we focused on the modeling the closing price of the securities FB and MSFT when Apple Incorporation (AAPL) was the leading company, by using restricted VAR (p). The following are our developed analytical models:

- The fitted RVAR (3) model for FB with AAPL is

$$\hat{C}_{F,t} = 6.52 + 0.93C_{F,t-1} + 0.31C_{A,t-2} - 0.19C_{A,t-3} \; . \quad .$$

- The fitted RVAR (5) model for MSFT with AAPL is

$$\hat{C}_{M,t} = 7.27 + 0.21C_{A,t-1} + 0.56C_{M,t-1} + 0.39C_{M,t-2} - 0.44C_{A,t-3} + 0.20C_{M,t-3} + 0.27C_{A,t-4} - 0.21C_{M,t-4}.$$

We have also fitted model of the closing price of the AAPL securities when itself as the leading company by using restricted VAR (p). These models are not our focus of interest so excluded from conclusion.

We can predict the current stock prices of the FB and MSFT companies if we know the previous day's prices of AAPL, FB and MSFT securities by using above models. The proposed model helps us to build a profitable portfolio by predicting the future price of the security with a high level of confidence. This provides the huge advantage over traditional strategies in investment that are relying on rumors. Our study can be highly beneficial for investors because they can secure their investment before a market crash or correction. On the basis of our model, they also can make the investment on a stock before the value of it scales up. Thus, our research has a high potential to be the best tool for investors in their strategies for building a profitable portfolio.

References

[1] S&PDowJonesIndices.S&P500. [Online]. Available: https://www.spglobal.com/spdji/en/indices/equity/sp500/overview. [AccessedAug. 28,2021].
[2] A.Sarda-Espinosa,"Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package,"cran.r.pp. 1–41, 2017.[Online].Available:https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf.[Accessed Jan10,2018].
[3] B. Pfaff and K. Taunus, "Using the vars package VAR: Vector autoregressive models Definition," University of Bayreuth. pp. 1–34, 2007. [Online]. Available: http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/vignettes/vars/vars.pdf. [Accessed Jan. 2, 2018].
[4] D. Miljkovic, "Brief Review of Self-Organizing Maps," Mipro 2017/Cts, pp. 1252–1257, 2017.´
[5] H. Yin, "The self-organizing maps: Background, theories, extensions and applications," Stud. Comput. Intell., vol. 115, pp. 715–762, 2008.
[6] M. T. Series, "Vector Autoregressive Models for Multivariate Time Series," Model. Finance. Time Ser. with S-PLUS®, pp. 385–429.
[7] R. Rakotomalala, "Introduction Representation of data using a Kohonen map, followed by a cluster analysis," Tanagra Data Mining, pp. 1–21, 2017.
[8] A. Ramkumar,"Alphabet Becomes Fourth U.S.Company to Reach $1 Trillion Market Value",Jan. 16, 2020.[Online]. Available:https://www.wsj.com/articles/alphabet-becomes-fourth-u-s-company-to-ever-reach-1-trillion-market-value11579208802.[Accessed Sep 14,2021].
[9] S.S.Chung,"A Class of Nonparametric Volatility Models: Applications to Financial Time Series," Semanticscholar, pp. 1–40, 2012. [Online]. Available: https://pdfs.semanticscholar.org/8d07/427adec97ef173f16b42a8c5ced9a00ba8f3.pdf [Accessed Jan10, 2018].
[10] Tan,Steinbach, Kumar, and Ghosh, "K-Means Algorithm," Ucs, 2002. [Online].Available: http://www.ucs.louisiana.edu/ xxw8007/kdd/-PPT/Kmeans-ICDM06.pdf [Accessed Jan15, 2018].
[11] T. Wickramarachchi, "Asymptotics for the Arc Length of a Multivariate Time Series and Its Applications as a Measure of Risk," All Dissertations. Paper 1040., 2012
[12] T. Wickramarachchi and F. Tunno, "Using Arc Length to Cluster Financial Time Series According to Risk," Communication in Statistics, 1(4) (2015).
[13] T. Warren Liao, "Clustering of time series data - A survey," Pattern recognition., vol. 38, no. 11, pp. 1857–1874, 2005. [dataset]Yahoo Finance Historical Data. [Online].Available: https://finance.yahoo.com/[Accessed Jan 20,2021].