Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

RESEARCH METHODOLOGY – PART B

EconoPhysics - Financial Forecasts

Second semester 2022-23

By Maroulis Adamadios

Professor: Avraam Xarokopoulos

**Linear analysis, complex networks, hierarchical clustering of time series analysis**

Table of Contents

[Abstract 3](#_Toc107157483)

[Introduction 4](#_Toc107157484)

[Section 2 5](#_Toc107157485)

[Section 3 9](#_Toc107157486)

[Section 4 15](#_Toc107157487)

[Conclusion 19](#_Toc107157488)

[References 20](#_Toc107157489)

Abstract*: Linear analysis and complex network analysis nowadays are some of the most useful tools to comprehend time series and extract useful information. Via visibility graph analysts have the ability to convert time series into complex networks and get even more information, visibility graphs have therefore been extended to the realm of time series analysis. Considering the daily closing price of stock as time series we can establish stock networks from those time series using the visibility graph method. In this paper, linear analysis is applied to five stocks in order to assess their performance. In addition, non-linear time series analysis tools are used such as visibility graphs to construct the complex networks and recurrence plots (RPs). We show that the proposed approach can provide useful information regarding the data set. The results could be important when considering investing in the stock market.*

*Keywords: Visibility graph, Recurrence plots, complex network, stock market*

# Introduction

*Considering the rapid development of internet finance, studying the stock market is considered to be one of the hottest topics. There are numerous methods regarding stock market analysis, such as data mining, minimum spanning tree (MST) (Boginski, Butenko et. all 2005), machine and deep learning approaches (Demetrius, Manke et.all 2005) and others. For example, complex network approaches were used to nonlinear time series and an in-depth review was provided of existing approaches of time series networks with an emphasis on recent developments by Yong Zou (Zou, Donner et.all 2018), MST was used to cluster analysis of S&P 500 index in the United States by Mantegna (1999). Based on this innovating research in 2007 Eom. C and Oh.G studied the American and Korean stock market using an improved MST (Eom, Oh et.all 2007). In recent years complex network theory, as an innovating way to comprehend the real world has attracted wide applications in many scientific fields, such as econometrics, biology, science, technology and others (Zhu, Wei 2021). In order to construct complex networks effectively based on data of stocks visibility graph method is considered a promising method (Lacasa, Luque et. all 2008). Therefore, in this paper, we conduct linear analysis in order to get a first view of the stocks and have some useful information (correlation, causality etc.). Meanwhile we conduct complex network analysis and hierarchical clustering in order to have a more in-depth analysis based on the stocks that are chosen. The data set used are five stocks mainly from the technology sector which are Advanced Micro Devices, Inc. (AMD),**NVIDIA Corporation (NVDA),**Alibaba Group Holding Limited (BABA),**Roku, Inc. (ROKU) and Unity Software Inc. (U). One year data of those stocks were collected from 21/06/2021 to 17/06/2022 from yahoofinance.com. This paper is organized as follows. In Section 2, the theory of linear analysis and the results are introduced. In Section 3, visibility graph theory and complex network analysis is presented, while in Section 4, hierarchical clustering method is used on the data set. In order to conduct our analysis two programming languages were mainly used,**MATLAB, which is a proprietary multi-paradigm programming language and numeric computing environment developed by MathWorks and Jupyter Notebook which is a web-based interactive computational environment for creating notebook documents while using Python coding language. In addition, E-views, a statistical package used mainly for time series oriented econometric analysis was used in order to conduct some tests.*

# Section 2

*In order to conduct our linear analysis, we use tools such as logarithmic returns, when trying to capture volatility in our timeseries. Logarithmic return is one of the three methods for calculating return and it assumes returns are compounded continuously rather than across sub-periods. It is calculated by taking the natural log of the ending value divided by the beginning value. Firstly diagram 1 is presented containing the original time series for our stocks.*

*Diagram 1*

Chart, line chart

Description automatically generated

*Time series of stocks*

*Watching the diagram above the first impression someone could have is that all five stocks are on a downward trend when 2022 starts and that is most probably caused due to global inflation at the time and interest rates hiking by the Federal Reserve Board (FED) in the United States at this period.*

*In the following diagrams volatility of this period is captured for every individual stock using the logarithmic returns:*

*Diagrams 2 – 6*

Chart

Description automatically generatedChart

Description automatically generated with medium confidenceA picture containing chart

Description automatically generatedChart, line chart

Description automatically generatedChart, line chart

Description automatically generated

*Volatility of stocks at current period*

*Using MATLAB’s autocorrelation, partial correlation and cross correlation functions we managed to obtain plenty of information regarding those stocks as individuals but also possible connections between them. The results though informed us that there is no big indicator showing that a linear forecasting model would not be able to obtain a precise forecast considering this data set as the indicators of autocorrelation and partial correlation functions in all those stocks were quite low, that kind of results probably suggest that white noise is strong in the data sets. While comparing the indicators we could say that they are not linearly predictable with the linear models used and with the regression conditions ranging in time lags. They are also not very important in autocorrelations and partial correlations. Considering space management, we chose to show the results of those functions for growth stock NVDA in the diagrams below, in figure 1 cross correlation for all five stocks is presented. The results though, indicated an impressive cross correlation between NVDA and AMD approximately 0.82 suggesting that the two stocks for lag=0 strongly moves in the same direction, which is presented also.*

*Diagrams 7 - 9*

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

*Autocorrelation Function Partial Correlation Function*

Chart, scatter chart

Description automatically generatedTable

Description automatically generated

*Cross Correlation Function Figure 1*

*Furthermore, using E-views we conducted granger causality tests for lags=1, 2, 3, 4 to extract information regarding the relationship between those stocks. The results indicated that none of the stocks granger causes any other stocks since the null hypothesis was not rejected in any case (Prob>0.05). Below we present the tests for all four lags that were performed.*

*Diagrams 10 – 13*

Table

Description automatically generated Table

Description automatically generated

Table

Description automatically generated Table

Description automatically generated

*Granger causality tests for Lags= 1, 2, 3, 4*

# Section 3

*The visibility graph was established to examine time series based on complex networks, in order to present it clearly, time series are plotted as vertical bars (Yun, Funito 2013).*

*For giving single variable time series , where is the value of observation at time , every node represents series data in the same order, the visibility lines between values define the links connecting nodes in the graph. If they are visible to each other in time series for the array (, ) and (, ), then, given an array, which is noted as (, ), when < < is placed between them fulfills:*

*+()*

*A set of time series and their corresponding visible diagrams are shown in diagram 14. The diagram is illustrating the vertical bars as a time series with seven data values, which was established according to the visibility method. Moreover, if a bar can be ‘’seen’’ by the top of considered one, then they will be linked, with that in mind a visibility graph network can be established whenever any visible two points in time series are connected.*

*Diagram 14*

Diagram, engineering drawing

Description automatically generated

*The visibility graph (source: Zhu, Wei 2021)*

*Visibility graphs inherits the attributes of time series and extract all the important information of the initial time series. Ultimately, every time series can be transformed into a different type of network by using that method, which can be distinguished between broad classes of dynamical systems (Bezsudnov et. all 2012).*

*Analyzing stocks traditionally is usually qualitive in terms of economics, although in this paper a quantitative method is given using the visibility graph method. Given the fact that we obtained closing prices of the stocks for each trading day we consider the closing price as a classical time series and each time a trading day represents a node, a link between each pair of node is decided by the relationship of the bars. Therefore, those connections indicates that the previous daily closing price has an impact on the next one.*

*Furthermore, in order to establish our complex networks, we considered the closing prices of our chosen five stocks with 253 daily closing price data and converted them into histograms that are shown in diagram 15 in order to comprehend the data set and have some starting clues to what we expect to see in the visibility network of each stock:*

*Diagram 15*

*Chart, histogram

Description automatically generated*

*Histograms of the selected stocks*

*Then, we created the recurrence plots for each stock, a recurrence plot (RP) is an advanced technique of nonlinear data analysis, basically it is a visualization (or a graph) of a square matrix, in which the matrix elements correspond to those times at which a state of a dynamical system recurs. The recurrence plot or adjacency matrix is a square matrix , The element of is equal to 1 if the distance between points and in phase space does not exceed some predetermined value ε, in the opposite case is equal to 0.*

*The adjacency matrix is an indicator of the way the nodes are connected to each other. By observing the matrix structure, we can see transition area displayed corresponding to transitions in the behavior of the dynamical system. While examination, we can identify dynamical changes when comparing to the original time series. A simple example is shown in diagram 16 and the recurrence plots are presented in diagrams 17-21.*

*Diagram 16*

*Chart, scatter chart

Description automatically generated*

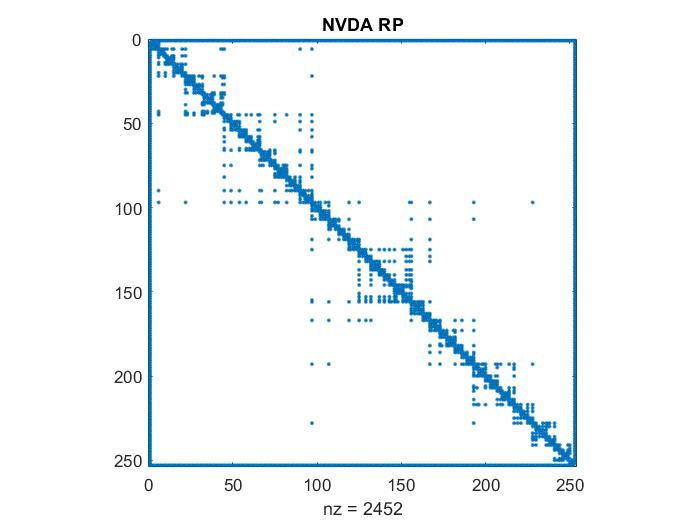
*Source: Avraam Xarokopoulos ‘’data clustering’’ lecture*

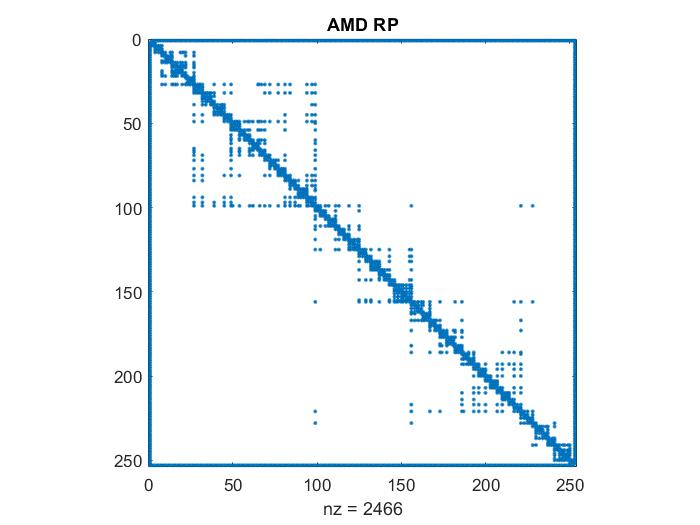
*Diagrams 17 – 21*

*Chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated*

**

*Adjacency matrix of each network*

*Using the open-source network analysis and visualization software package Gephi we managed to establish by visibility graph the network graphs for each stock we are examining presented below:*

*Diagrams 17 – 21*

*BABA AMD NVDA*

*A picture containing shape

Description automatically generatedA picture containing accessory, necklet, colorful

Description automatically generatedA picture containing smoke, accessory, colorful

Description automatically generated*

*ROKU U*

*A picture containing map

Description automatically generated A picture containing accessory, necklet, vector graphics

Description automatically generated*

*Network graphs (Modularity class)*

*Some of the topological metrics that were used are the following:*

*Average Degree: The degree of a graph is a fairly important local property of each node, this represents the number of links that the node has for the other nodes, in undirected graphs.*

*Clustering Coefficient: This measure (C), measures the total number of closed triangles in the graph, that is, it measures the degree to which the nodes in a graph tend to cluster. It is calculated by the ratio of the number of closed triangles to the number of possible triangles, that is, the amount of connected triplets of nodes.*

*Modularity: Measures how good the division of the graph is in specific communities, that is, how different are the different nodes, belonging to different communities, from each other. A high modularity value, (Q), indicates a graph with a dense internal community structure, that is, with many edges between nodes within communities and sparse connections between nodes of different communities.*

*Based on the topological metrics we can summarize that each line indicates a correlation between the connected elements, and the thickness of the lines indicates the strength of the correlation, while the size of every node represents the number of edges connected to it (usually referred to as degrees).*

*More specifically by observing the network graphs we can identify communities with large number of connected nodes in different colors, those communities correspond to different areas of our time series and each one of them are unique as they exhibit the same characteristic behavior.*

*There are more or less connections between any two elements, a few elements occupy the center of the topological structure, while some are at the edge of the structure (see NVDA, AMD). Many elements act as bridges between the elements present in the middle and the elements at the edges, which present a particular community feature. Each community feature is presented by a different color.*

# Section 4

*Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects (in our case prices of the time series) into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the prices within each cluster are broadly similar to each other. We decided to use the knee point method in order to achive the most accurate number of clusters to use for our data set, doing so, we used the k means clustering algorithm. When using the k means clustering algorithm, we had to specifically define k, or the number of clusters we wanted the algorithm to create. Rather than selecting an arbitrary value, such as the number of clusters you want for practical purposes, there’s a specific way to do the selection of the optimum k.*

*Usually the most accurate number of clusters is identified visually using a data visualisation tool known as the elbow plot. The elbow plot is generated by fitting the k means model on a range of different k values (in our case from 1 to 20) and then plotting the SSE (Sum of squared errors) for each cluster. The elbow or knee represents the point at which a higher k, or additional clusters, stop adding useful information and make the clusters harder to separate. However, on some datasets, this inflection point is not always easy to spot and its up to the analyst to choose which one in his opinion represents the most efficient one. Considering our data set we ran some tests based on this algorithm via python in order to try to comprehend the number of clusters would fit best for our data set, the elbow plot indicated that k=3 or k=4 would fit best as diagram 22 indicates although, examining visually diagram 23 by running the algorithm for both our choices we decided k=3 would be best for our data set.*

*Diagram 22*

*Chart

Description automatically generated*

*Diagram 23*

*Chart, scatter chart

Description automatically generated*

*As we can see in the diagram above we have three similar groups (clusters) of our stocks indicating relationship between sets of data. Furthermore, to have a better understanding of our data set we used a dendrogram based on our current analysis to identify which set of data match. Below diagram 24 represents the dendrogram for our stocks, while diagram 25 shows a dendrogram based on daily prices for all five stocks.*

*Diagram 24*

*Chart, histogram, box and whisker chart

Description automatically generated*

*The key to interpreting a dendrogram is to focus on the height (Euclidean Distance) at which any two stocks are joined together. In our case, we can see that NVDA and AMD are most similar, as the height of the link that joins them together is the smallest. The next two most similar stocks are BABA and ROKU. In the dendrogram above, the height indicates the order in which the clusters were joined and reflects the distance between the clusters as is shown above. The diagram shows that the big difference between clusters is between the cluster of BABA,ROKU versus that of U,AMD and NVDA.*

*Diagram 25*

*Chart, histogram, box and whisker chart

Description automatically generated*

*Based on these diagrams we can identify the relationship between ROKU and BABA is strong and represents our first group, while AMD and NVDA also have a strong relationship and are connected with U.*

*Cutting the dendrogram by a threshold value, we have chosen a cut-off of 1 in which a vertical line is drawn. For a chosen cut-off/threshold value, we can always simply count the number of intersections with vertical lines of the dendrogram to get the number of formed clusters. In our analysis we decided that 3 clusters are more efficient so 3 clusters should appear in our dendrogram.*

*Diagram 26*

*Chart, histogram

Description automatically generated*

# Conclusion

*In this paper, five stocks mainly from the technology industry are considered for a linear analysis and from the perspective of complex networks. We use various functions to extract information from the linear analysis and the visibility graph method to construct networks. More specifically, ACF, PACF and CCF were used as well as granger causality tests for our linear analysis, recurrence plots were employed along with complex network time series analysis while hierarchical clustering method was finally used along with dendrograms. The data set is considered during a possible economic recession period and the results indicated correlations and relationships to some degree between our time series, useful information was extracted also from complex networks which helped us visualize the data set and extract information related to the topological metrics, hierarchical cluster analysis also had a big part in our analysis in which indicated strong connections between the data set. Knowledge of dynamic characteristics combined with all the previous information could establish strong investing strategies, risk management strategies and could provide some insight into the stock market.*

# References

***[1]*** *C. Eom, G. Oh, S. Kim, Topological properties of a minima spanning tree in the Korean and the American stock markets, J. Korean Phys. Soc. 2007*

***[2]*** *L. Demetrius, T. Manke, Robustness and network evolution–an entropic principle, Physica A 346 (3–4) (2005)*

***[3]*** *I.V. Bezsudnov, S.V. Gavrilov, A.A. Snarskii, From time series to complex networks: the dynamical visibility graph, Physica A 414 (2012)*

***[4]*** *L. Lacasa, B. Luque, F. Ballesteros, J. Luque, J.C. Nuno, From time series to complex networks: The visibility graph, Proc. Natl. Acad. Sci. 105 (13) (2008)*

***[5]*** *R.N. Mantegna, Hierarchical structure in financial markets, Eur. Phys. J. B 11 (1) (1999)*

***[6]*** *V. Boginski, S. Butenko, P.M. Pardalos, Statistical analysis of financial networks, Comput. Statist. Data Anal. 48 (2) (2005)*

***[7]*** *W. Yun, M. Funito, An analysis of backbone of outer spatial morphology of campus, Commun. Comput. Inf. Sci. 369 (2013)*

***[8]*** *Jia Zhu, Daijun Wei, Analysis of stock market based on visibility graph and structure entropy, Physica A: Statistical Mechanics and its Applications, 2021*

***[9]*** *Yong Zou, Reik V. Donner, Norbert Marwan, Jonathan F. Donges, Jürgen Kurths, Complex network approaches to nonlinear time series analysis, Physics Reports 2019*

***[10]*** *Avraam Charakopoulos, Theodoros Karakasidis, loannis Sarris, Pattern identification for wind power forecasting via complex network and recurrence plot time series analysis, Energy Policy 2019*

***[11]*** *Michele La Rocca, Francesco Giordano, Cira Perna, Clustering nonlinear time series with neural network bootstrap forecast distributions, International Journal of Approximate Reasoning 2021*

***[12]*** *Hailin Li, Ruiying Jia, Xiaoji Wan, Time series classification based on complex network, Expert Systems with Applications 2022*