Lappeenranta University of Technology
School of Engineering Science
Degree Programme in Technomathematics and Technical Physics

Kristina Nikolaenko

EXPLAINING STOCK DEVIATION CORRELATION IN NEW YORK STOCK EXCHANGE.

Examiners: Professor Tuomo Kauranne
D.Sc. (Tech.) Matylda Jabłońska-Sabuka

Supervisors: Professor Tuomo Kauranne

ABSTRACT

Lappeenranta University of Technology
School of Engineering Science
Degree Programme in Technomathematics and Technical Physics

Kristina Nikolaenko

Explaining stock deviation correlation in New York Stock Exchange

Master's Thesis

2017

48 pages, 23 figures, 5 tables

Examiners: Professor Tuomo Kauranne
D.Sc. (Tech.) Matylda Jabłońska-Sabuka

Keywords: stock market, NYSE, correlation coefficient, regression model

The development of a universal model that would allow us to describe the market sentiment, as well as identify and predict potential behavior of the economy, is very complicated. Economists, mathematicians and physicists are trying to solve this issue by applying various statistical tools, analysis and physical laws.

In this thesis, data from the New York Stock Exchange were taken as a basis for creating a simple model describing the state of the economy. A matrix of correlation coefficients between stocks was created, and further a normalization of initial data by Dow Jones and S&P500 indices was carried out on these data. A model describing the histogram data was selected and parameters for each of the histograms were constructed.

As a result, based on the obtained parameters, a simple model enabling us to identify the rise and decline period in the economy was created.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Professor Tuomo Kauranne, who supported this Master's thesis throughout the course of this study. He was always ready to give me help and advice. I also would like to thank him for his excellent guidance, patience and knowledge, this thesis would not have been without his support.

I would like to express my gratitude to Christoph Lohrmann for his productive ideas and advices, supporting this Master's thesis.

I would like to acknowledge the Lappeenranta University of Technology, the Department of Technomathematics, for hospitality and great atmosphere, and also all staff of LUT for their willingness and desire to help me in the process of my education.

I would like to acknowledge the Southern Federal University, Institute of Mathematics, Mechanics and Computer Science, for giving me this opportunity to study in Finland. In particular, I would like to express my gratitude to my Russian supervisor Associate Professor Konstantin Nadolin for his help and advice during the whole period of my study.

I would like to express my special gratitude to my lovely parents Valery and Larisa for their financial and moral support during all period of this study. Deepest thanks for their love and advice encourage me not to give up.

My special and the deepest gratitude to *моему другу сердца* Artyom Chvostov for his unbelievable patience, outstanding support, willingness to help me and desire to encourage in my joyrides. I can say with 100% probability, that this thesis as well as the whole study would not be able without him. Thank you so much, *my Dear*!

Lappeenranta, May, 2017

Kristina Nikolaenko

TABLE OF CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

ASX	Australian Securities Exchange
CSE	Chinese Stock Exchange
DFE	Degrees of Freedom for Error
DJI	Dow Jones Index
FWB	Frankfurt Stock Exchange
GEV	Generalized Extreme Value distribution
HKE	Hong Kong Stock Exchange
NASDAQ	National Association of Securities Dealers Automated Quotation
NYSE	New York Stock Exchange
OTC	Over-The-Counter Market
RMSE	Root Mean Square Error
SESDAQ	Stock Exchange of Singapore Dealing and Automated Quotation
SSE	Shanghai Stock Exchange
SSE	Sum of Squared Error
S&P500	Standard and Poor's stock market index
TSE	Tokyo Stock Exchange

1 INTRODUCTION

This research is a follow up study of the Master's thesis "Forecasting financial weather – can we foresee market sentiment? Spectrum of stock price behavior - NYSE case study" written by Alisa Zeleva. In that study statistical and topological analysis were conducted in order to examine the daily closure prices taken from New York Stock Exchange (NYSE) in the period of time from June 2006 to February 2013. For better understanding the sentiment of NYSE some basic statistics tools were applied. In particular, the correlation matrices were found by calculation correlation coefficients between all pairs of stocks. Two types of normalization by Dow Jones (DJ) and Standard and Poor's 500 (S&P500) indices were applied to initial data: by regression model and by mean and standard deviation. A metric between two stocks, called distance, was calculated. Based on it, a hierarchical cluster tree was constructed. To detect industry clusters and companies belonging to it, singular value decomposition was provided.

A number of significant findings were conducted from that research, one of which is the shape distribution of the correlation coefficient matrix, a bulk of which is shifted to the right hand side with respect to zero. Thereby, the further study is based on the research of this distribution.

This Master's thesis has the following structure. Section 2 covers the theoretical background and provides the lessons, based on the previous research on this subject. Section 3 introduces on the basic concepts of equity market, the methods of analyzing the stock prices, and it also discusses DJ and S&P500 indices and their functions.

The main Section 4 is devoted to a recalculation of results from the Master's thesis of A.Zeleva are presented, which serve as a basis for the subsequent research. This section provides the research methodology, based on which a simple model that allows to identify rise and decline periods in the economy is constructed. Section 5 gives suggestions for future work. Section 6 includes the main findings obtained during this study.

2 LITERATURE REVIEW

Comprehension of the mechanism of price changes is highly useful for decision-making, such as buying or selling an asset. This knowledge gives a perception of the price movement – up or down and at what time moment. More generally, what characteristics of a series of price increments could help to improve the forecast quality of their future development? Among a bunch of characteristics of price dynamics a promising candidate is the correlation between successive price increments.

A number of research works have focused on the studying the correlation between stocks on various trading platforms, over-the-counter markets and stock exchanges such as NYSE, NASDAQ, AMEX and CSE. Nevertheless, plenty of studies are more focused on a certain sector. As previously stated, this research is based on the Master's thesis "Forecasting financial weather – can we foresee market sentiment? Spectrum of stock price behavior - NYSE case study" written by Alisa Zeleva; papers and articles mentioned in her thesis served as a basis for this study.

Based on the results of that research, it was discovered that the establishment of a universal statistical model allowing to predict future changes in stock prices for a certain period of time with minimal risks is very complicated.

The aim of the research [1] was to perform basic statistical research for identifying possible factors that have an impact on the NYSE sentiment. The most significant findings, obtained from this study are the following:

- The shape distribution of correlation coefficient matrix, calculated from 1930 companies, taken from NYSE, is similar to a Maxwell-Boltzmann distribution.
- The NYSE is not an ergodic system. It means, that a stable mean in a short period of time is not the same as a mean in a long period.

- Singular vectors, corresponding to largest singular values, are not orthogonal over time, but they are dependent on from each other by a constant factor. This is different for various periods of time.

The Master's Thesis "Static Waves in Corporate Space: Characterizing Oscillating Trading Patterns in New York Stock Exchange" of Thacienne Uwimanayantumye was inspired by [1]. In this research, utilizing a seven year time series of stock prices used to compute S&P500 index on NYSE, a local chart to the "corporate space" was constructed in order to find standing waves and other patterns in the behavior of the stock price deviations from S&P500. The calculation of the correlation matrix normalized by S&P500 index, using a local singular value decomposition over a set of four different time windows was performed. In almost all cases, each singular vector is essentially determined by a relatively small set of companies with big positive and negative weights on that singular vector. Over particular time windows, sometimes the weights are strongly correlated with at least one industrial sector and certain sectors are more prone to fast dynamics whereas others have longer standing waves.

"Introduction to Econophysics. Correlations and Complexity in Finance " by Rosario M. Magenta and H.Eugene Stanley is substantive work that maintains research work [1] and also provides a basis for the current Master's thesis. This book is focused on the usage of statistical physics in the depiction of financial systems. The authors also construct a new stochastic model that indicates a number of statistical properties obtained from empirical data. However, the emphasis in this work is placed on the Chapters 12-13, which describe the methods used for research on the correlation and uncorrelation between pairs of stocks, and also provide the statistical tools, that are essential for market research.

"Investment Philosophies: Successful Strategies and the Investors Who Made Them Work" by Aswath Damodaran also underpins this Master's thesis. This book encompasses multiple of strategies including indexing, passive and active value investing, growth investing, technical analysis, arbitrage and other investment philosophies. In particular, in the Chapter 7 it is told about serial correlation and its significance in consecutive time periods.

“Why Stock Markets Crash: Critical Events in Complex Financial Systems” by Didier Sornette covers the territory that starts from description of how the financial market arises to the methods of crash prediction, both of the market in its entirety, and of individual financial assets. In Chapter 2 he describes trading strategies based on correlation and demonstrates examples of its realization on real data.

In the article “The Stock Market and Investment” by Warren Tease the relationship between stock prices and investments is examined, considering the question of whether investments are affected by inefficient pricing in equity markets.

The article “International Asset Allocation with Time-Varying Correlation” by Andrew Ang and Geert Bekaert solves the dynamic portfolio choice problem of a US investor encountered with time-varying investment opportunity set which maybe characterized by correlations and volatilities.

“Price-volume Multifractal Analysis and its Application in Chinese Stock Markets” by Ying Yuan is dedicated to multifractal cross-correlation analysis between stock price return and trading volume variation in Chinese stock markets. Cross-correlation between stock price and trading volume was studied using cross-correlation function and detrended cross-correlation analysis.

In the research “Autocorrelation in Global Stochastic Trend” by Anatoly Peresetskiy a new econometric model is being developed. The yield of the financial market index is presented as a sum of two independent components: a global one that depends on news that have an impact on the global financial market and a local one depending on news that are significant only for this market. The model is taking into account the non-synchronism of observation of one-day returns in different time zones, and allows to assess a global trend. It was assumed, that the increments of the global trend between closing times are independent. The author proposes a model with autocorrelation of global stochastic trends, which suggests the possibility of its increment correlation at neighboring time intervals.

The book “Risk Management: Tasks and Solutions” by George Prosvetov is devoted to theoretical questions, built on the basis of the newest empirical, financial and mathematical methods of economic analysis, and practical issues of construction of a risk management system. Particularly, Chapter 18 states about regression model, its formation, prediction and forecasting based on linear and multiple regression models.

“The Econometric Modelling of Financial Time Series” by Terence Mill and Raphael Markellos provides the techniques essential to launch the empirical analysis of financial time series. Particularly, it gives an introduction to the modern methods of econometric analysis of statistical data presented in the form of time series that take into account the possible presence of a stochastic trend in a variables under consideration.

In the literature, various methods and approaches are presented to describe the sentiment of stock prices, as well as their forecasting. However, a universal model that can fully describe such a complex, multifactorial, non-stationary system as stock exchange does not exist yet. In this work a simple model, constructed on the basis of a calculation of correlation matrix and allowing to identify periods of recession and growth, was obtained.

3 STOCK MARKETS AND THEIR BEHAVIOR

Equity markets are one of the key mechanisms for involvement of monetary resources for investments, modernization of the economy or stimulation of production growth. Nevertheless, as the experience of many decades shows, the world stock market can be a source of large-scale financial instabilities and social upheaval.

In this regards, a relevant issue with significant macroeconomic importance is the establishment of a long-term strategy for the development of the stock market based on a theoretical study of the fundamental factors affecting equity markets. Majority of publications related to the stock market are of microeconomic nature, i.e. the studies focus on individual objects and processes.

In most cases, the research do not contain a comprehensive study of development challenge and strategies from the field of stock markets, focusing only on specific aspects, strategies, groups of issues, or certain features of individual markets.

The issue of identifying and analyzing a set of fundamental factors that determine the scenario for the development of stock markets and associated with the economic and political choice of society, models of market economies, the dynamics of resource prices, long economic cycles is still open.

The aim of various studies is to establish a concept for the development of individual stock markets based on an analysis of its current sentiment, the determination of the underlying factors affecting it and the resulting causation relationships, together with the market model justification. For the analysis of equity markets and the definition of the strategy, various approaches are used, e.g. fundamental and technical analysis.

Fundamental analysis is a method of predicting future movements of securities quotations based on economic, political and other significant factors and indicators that will affect the supply and demand of securities. Fundamental analysis determines and measures factors influencing the intrinsic value of a security, for instance, the overall economic and political

environment, including factors affecting supply and demand associated with securities, goods and services [12].

Technical analysis, on the other hand, is a method of predicting price changes and future market trends by examining historical changes in the market, taking into account the prices of stocks, the volume of the open positions and other matters. Technical analysis examines what has already happened on the market, and what is not supposed to happen. It studies the price movement and the capacity of transactions with securities and builds charts according to the data, based on the actions of market participants. Technical analysis focuses on the activity of the stock market [12].

The main aspects of the stock markets are considered in this chapter, in particular, New York Stock Exchange, as the study is based on data obtained from this exchange.

The stock exchange is a special institutionally organized market where securities are traded, and transactions are performed by professional participants of the equity market.

The stock exchange provides mobilization and concentration of temporarily free cash savings and savings by selling securities to exchange intermediaries in primary and secondary stock markets, lending and financing of the state and the private sector through the purchase of their securities and resale, and lending and financing of exchange speculators through transactions, the concentration of operations with securities, the establishment of prices on them, reflecting the level and the ratio of supply and demand.

Prominent examples of stock exchanges are NYSE, NASDAQ, as well as the Canadian OTC, the Singapore stock exchange (SESDAQ), The Shanghai Stock Exchange (SSE), the Frankfurt Stock Exchange (FWB), the London stock exchange, Australian OTC (ASX), Hong Kong (HKE), Tokyo (TSE) stock exchanges.

New York Stock Exchange is one of the largest in the world. NYSE is an association, and therefore the government cannot directly control its activities. The equity market is a legal entity, and it enjoys complete independence in matters of its organization and work. The

activity of the exchange is financed by contributions from participants of the exchange that have acquired a “place” on it, annual contributions of enterprises quoting their securities on the exchange, fees from exchange transactions and other payments be exchange participants and customers (e.g. fees for issuing certificates, registering exchange transactions, for the provision of advisory intermediary, information, legal and other services by the exchange subdivisions).

Instantaneous price is published in the table of exchange rates, and is called a quotation. In order for a security to be quoted on the exchange, it must be admitted to quotation by the decision of the relevant authority of the stock exchange. Since a large number of stocks and other securities are quoted simultaneously at the exchange, prices on which are generally moving in different directions, generalized stock price indices are used to estimate the average price change, the most important of which are the DJ and S&P500 indices.

The DJ index is a price index, or an index with a weight equal to the price. Indices of this type are calculated as the arithmetic mean of the company's stocks in the basic index list. This index type also includes the index of the TSE – Nikkei.

The S&P500 index belongs to the capitalization indices. This type of index reflects the total capitalization of enterprises, i.e. the composition of the current market value of the stocks issued by the enterprise by their number in circulation, whose securities are used to calculate the index. The calculation of the index value is normalized to a certain base date.

The functioning and successful development of the securities market is impossible without the availability of information on exchange processes. In countries with developed economies, the analysis of stock market conditions, trends and business activity of the stock market is of paramount importance. There is a need for constant operational information on the market situation, on the state and changes in stock prices, etc. Key indicators of the state of the stock market, assessing its business activity are stock indices.

High costs associated with the calculation and transfer of index values over long distances, the existence of large advisory agencies and bureaus specializing directly in the development and analysis of indices, a wide variety of indices themselves indicate that stock indices occupy a significant place in the system of financial indicators of the securities market securities.

In a generalized form, the role of stock indices can be reduced to three functions:

- Diagnostic – the ability of the indices system to characterize the state and dynamics development of both the national economy as a whole and its individual as;
- Indicative - the fulfilment of indicative functions by indices implies that the existence of an objective assessment of the price situation on the stock market provides a starting point for assessing the behavior of large investment funds, individual investors and portfolio managers;
- Speculative - the usage of derivative securities for indices in the exchange game, the instant ability to respond to a wide range of economic, political and social phenomena.

4 RESEARCH METHODOLOGY

4.1 Overview of initial data

One of basic ways to study and analyze the sentiment of the stock prices is to examine the correlation between daily closure prices. The time series of a stock prices during the whole period of time is presented in the Fig.1.

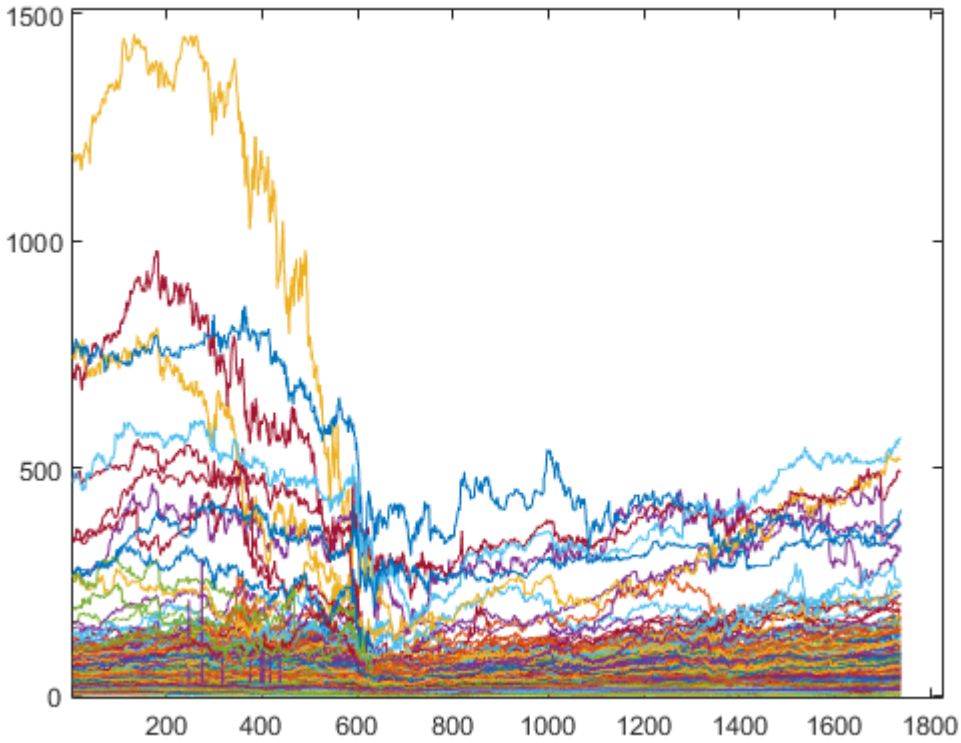


Figure 1: Time series of daily close NYSE stock prices.

In further analysis to form the matrix of correlation coefficients, the correlation between pairs of stock is calculated by the following formula:

$$\rho(ij) = \frac{\mathrm{cov}(i, j)}{\sqrt{\mathrm{cov}(i,i)\mathrm{cov}(j,j)}} \quad (1)$$

where i and j – pairs of stocks, cov – covariance matrix.

The magnitude of the correlation coefficients can vary from -1 to +1 depending on how similarly the stock prices behave. The nature of the correlation can be easily to understood from several examples:

- If two stocks have a correlation coefficient of +1 then their price fluctuations repeat each other. This is a complete positive correlation
- If two stocks have a correlation coefficient of -1 then their prices change in opposite direction. This is a complete negative correlation
- If two stocks have a correlation coefficient of 0 then their prices vary independent from each other.
- If two stocks have a correlation coefficient of +0.6 then their price fluctuations in 60% of their variation coincides, and in 40% their variation do not depend on each other. If two stocks have a correlation coefficient -0.5 then to half of their variation they have a complete negative correlation, and in the other half they are independent.

Put simply, there are three special cases:

1. $\rho(ij) = +1 -$ strong correlation between a pair of stocks
2. $\rho(ij) = 0 -$ uncorrelated pair of stocks
3. $\rho(ij) = -1 -$ strong anticorralation between a pair of stocks

By funding all correlation coefficient a square correlation matrix was obtained and it has the following properties:

1. $\rho(ij) = \rho(ji) -$ the matrix is symmetric
2. $\rho(ij) = 1 -$ the diagonal elements of the matrix are equal to 1.

For the data of n stocks the number of correlation coefficients is calculated, and is equal to:

$$\frac{n \cdot (n-1)}{2} \tag{2}$$

In this case, the number of stocks is 1930, and the number of unique correlation coefficient is equal to 1861458. Distribution of correlation coefficient matrix $P(\rho)$ is presented in the form of a normalized histogram in Fig.2.

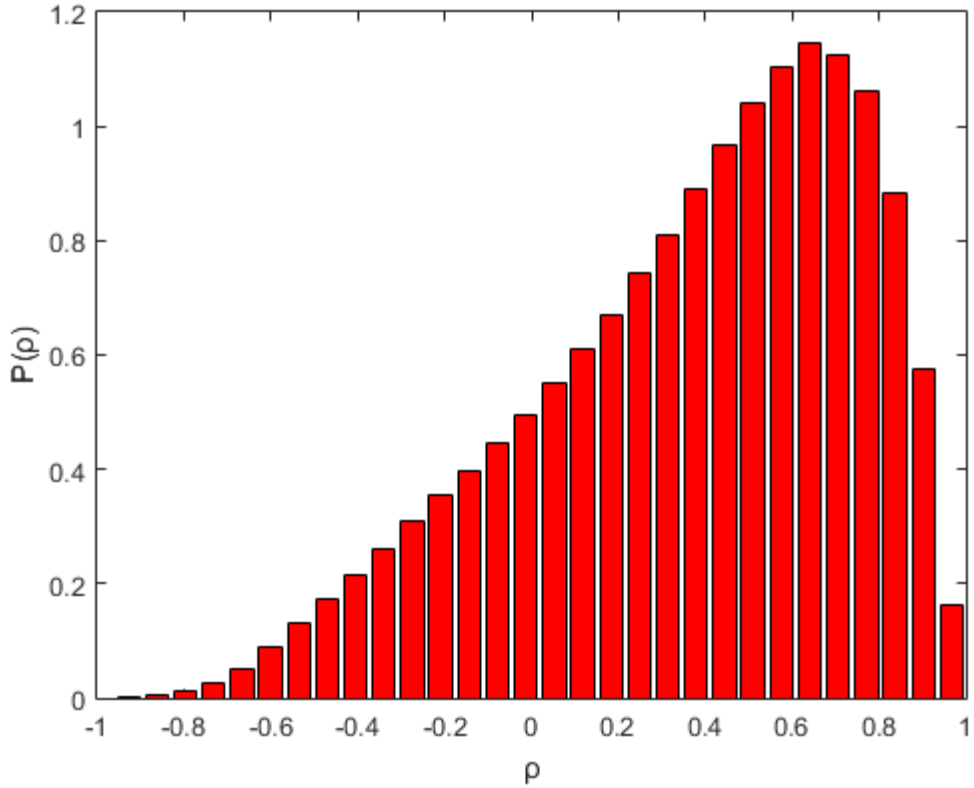


Figure 2: Distribution of correlation coefficients for the period from June 2006 to February 2013.

The normalization of histogram was implemented by the formula:

$$\text{normalized } x_i = \frac{x_i}{\sum_{i=1}^{n} x_i} \tag{3}$$

where x_i – is a value of ith calculated correlation coefficient.

The mean of correlation coefficient is equal to 0.3699. The shape of distribution P(ρ), which is shifted to the right hand side, shows that the companies mostly positively correlate with each other. The P(ρ) reaches a maximum value of 1.1407 at $\rho = 0.6275$.

In [1] it was considered that the shape of the distribution is very similar to the Maxwell-Boltsmann distribution. However, it does not give an accurate description of the distribution shape. Hence, another approach was used to identify the distribution shape in this study.

To find the distribution of shape like in the previous study (where the bigger mass is shifted to the positive side) seems relatively complicated, thus, the histogram was inverted and further we tried to identify a possible family of distribution. The inverted histogram is presented in Fig. 3.

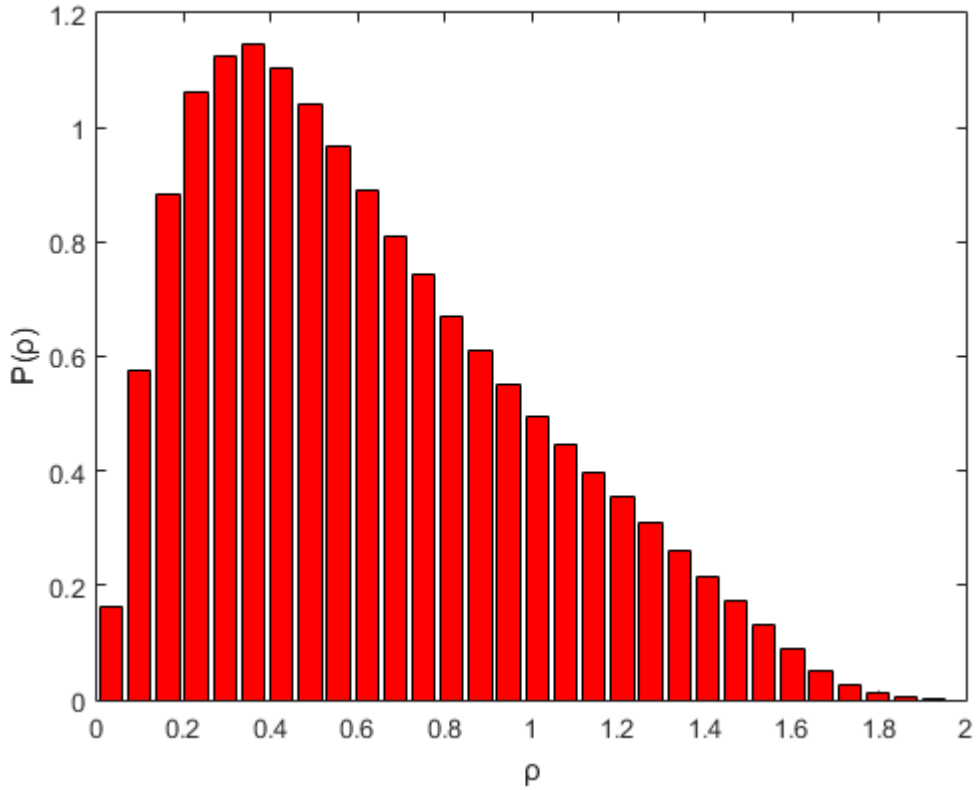


Figure 3: Inverted distribution of correlation coefficients for the period from June 2006 to February 2013.

Appropriate distributions of correlation coefficient matrix are presented in Fig.4. The distributions, describing the histogram are Weibull, Nakagami, Gamma and Rayleigh distributions. They are arranged in descending order, i.e. the most appropriate distribution is the Weibull distribution.

Weibull distribution is a continuous probability distribution with two parameters. The probability density function is:

$$F(x;\lambda;k)=\begin{cases}\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x\geq 0\\ 0 & x<0\end{cases} \quad (4)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter. If the quantity X is a "time-to-failure", the Weibull distribution gives a distribution for which the failure rate is proportional to a power of time. The shape parameter k can be interpreted directly as follows:

- $k < 1$ indicates that the failure rate decreases over time
- $k = 1$ indicates that the failure rate is constant over time.
- $k > 1$ indicates that the failure rate increases over time

In our calculations, $\lambda = 0.70802$ and $k = 1.71962$ (the failure rate increases over time).

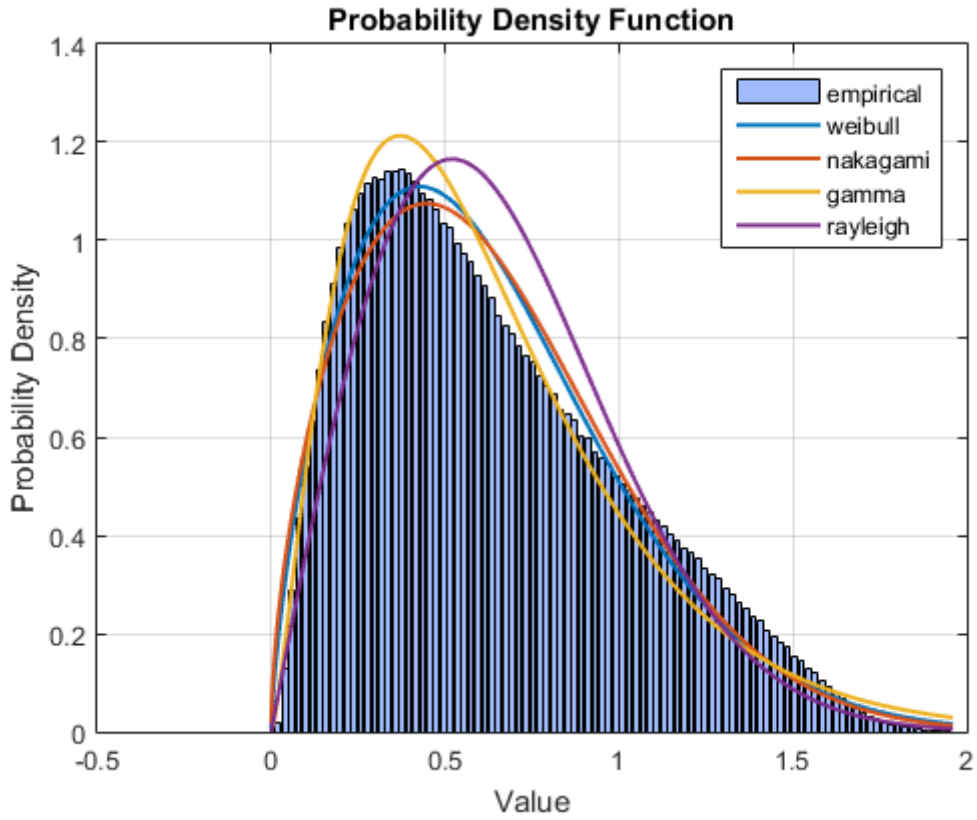
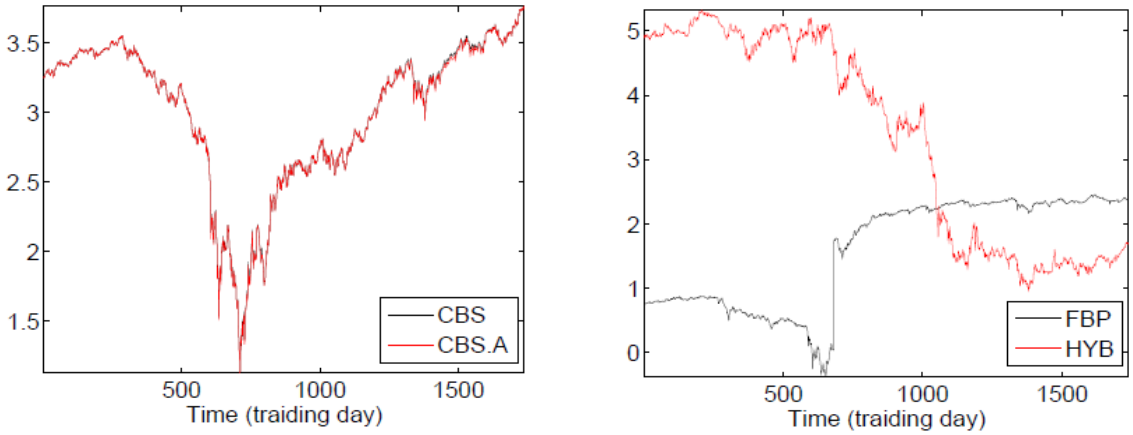


Figure 4: Distribution of correlation coefficients for the period from June 2006 to February 2013 (Inverted Histogram).

4.2 Correlation and anticorrelation between stock prices

Let us consider stocks that have the largest and smallest values of $\rho(ij)$, i.e. strong correlation and anticorrelation, respectively. The largest value corresponds to the pair of stocks "CBS" and "CBS.A", it equals $\rho(ij) = 0.99978$. This company is known as "CBS Corporation". The time series of that pair of stocks is presented in Fig.5a. The smallest value corresponds to the pair of stocks "FBP" and "HYB", it equals to $\rho(ij) = -0.96027$. The full names of the companies are "First Bancorp" and "New American High Income Fund Inc." The time series of this pair of stocks is performed in Fig.5b.



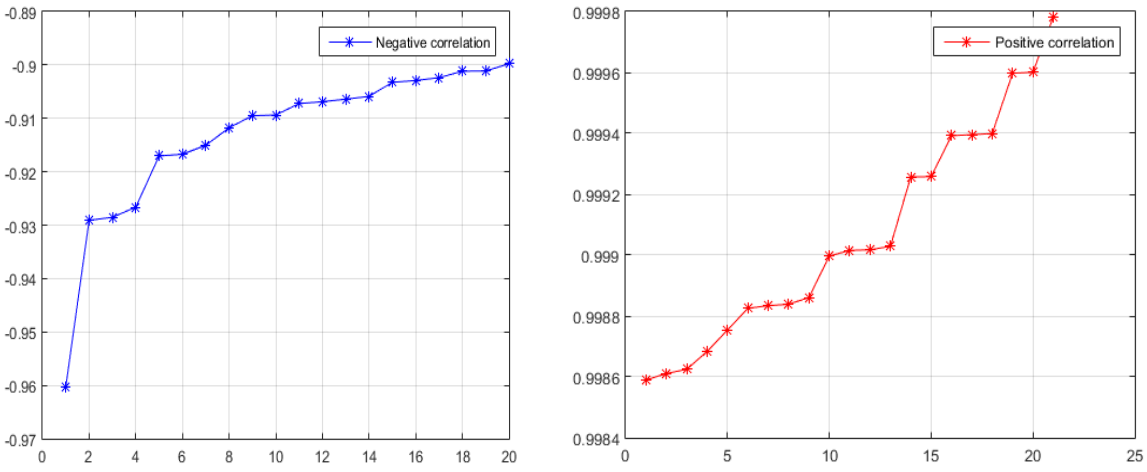
(a) The most correlated pair of stocks (b) The most uncorrelated pair of stocks

Figure 5: Pairs of stocks, corresponding to the maximum and the minimum correlation coefficient.

Matrix coefficients we transformed into a vector, i.e. the correlation coefficient matrix was reshaped into a vector and a vector was sorted from negative to positive correlation coefficient. Table 1. shows the correlation for fifteen of the most correlated and uncorrelated pairs of stock and their respective companies. Fig.6a presents a graph for twenty companies with the most negative (strong anticorrelation) coefficients, and Fig.6b provides the most positive (strong correlation) coefficients. Fig.7 shows correlation coefficients in the correlation matrix transformed into a vector for the whole period of time (from June 2006 to February 2013). It can be seen, that there are many positive coefficients, less negative ones and very few very positive correlation coefficients.

Negative Correlation			Positive Correlation		
Names of Companies		Corr. Coeff.	Names of Companies		Corr. Coeff
“FBP”	“HYB”	- 0.96027	“CBS”	“CBS.A”	0.99978
“AZO”	“NAT”	- 0.92909	“GJM”	“GKM”	0.99961
“TCI”	“THS”	- 0.92852	“STZ”	“STZ.B”	0.99959
“CHD”	“NAT”	- 0.92663	“MKC”	“MKC.V”	0.99939
“NAT”	“TDG”	- 0.91704	“SHI-F”	“SFI-G”	0.99939
“NAT”	“NEU”	- 0.91672	“BEE-B”	“BEE-C”	0.99939
“AZO”	“TCI”	- 0.91504	“RBS-M”	“RBS-N”	0.99925
“NAT”	“ROL”	- 0.91175	“SFI-G”	“SFI-I”	0.99925
“NAT”	“TYL”	- 0.90951	“SFI-F”	“SFI-I”	0.99902
“ENB”	“NAT”	- 0.90936	“SFI-E”	“SFI-F”	0.99901
“NAT”	“TJX”	- 0.90723	“GMA”	“GOM”	0.99901
“BVN”	“FBP”	- 0.90687	“GJM”	“GMA”	0.99899
“BXS”	“MCD”	- 0.90641	“RBS-N”	“RBS-P”	0.99886
“GWW”	“NAT”	- 0.90587	“GJM”	“GOM”	0.99883
“MCD”	“SKY”	- 0.90324	“IND”	“INZ”	0.99883

Table 1: 15 most correlated and uncorrelated pairs of stock and their companies.



(a) Negative correlation coefficient (b) Positive Correlation Coefficients

Figure 6: 20 companies with most negative and positive correlation coefficients.

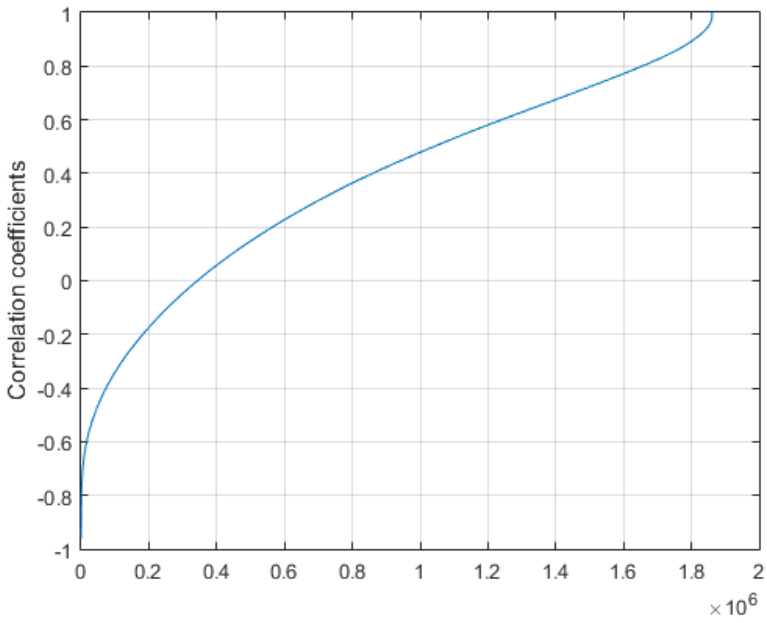


Figure 7: Correlation coefficients of the correlation matrix transformed into a vector for the period from June 2006 to February 2013.

4.3 Market sentiment

Let us look at the distribution of correlation coefficients over time and in different years. Whether a distribution of correlation coefficients would be the same as in Fig.2, i.e. shifted to the right hand side with respect to zero? In Fig.8 the correlation coefficients are shown for each year for the period from June 2006 to February 2013.

Independence between stock prices changes from year to year, and transformations in the shape of $P(\rho)$ are revealed. As it might be seen that a distribution of correlation coefficients in case of splitting data into one-year-intervals have a similar shape as in the case when a whole period is considered, however, some of the distributions look like an exponential function, e.g. 2008-2009, 2009-2010 years.

It is known, that in the year 2008 there was an economic crisis. From Fig.8 it can be seen, that the histogram of correlation coefficients of this year has a form of an exponential function. It is shifted to the right hand side, and there are more positive correlation coefficients, i.e. companies correlate between each other. Whereas in 2012-2013 (in the absence of an economic crisis and other global events, affecting the economy) a histogram

of the correlation coefficient has the form of distribution approximate to normal, and in this case the number of negative correlation coefficients is much larger.

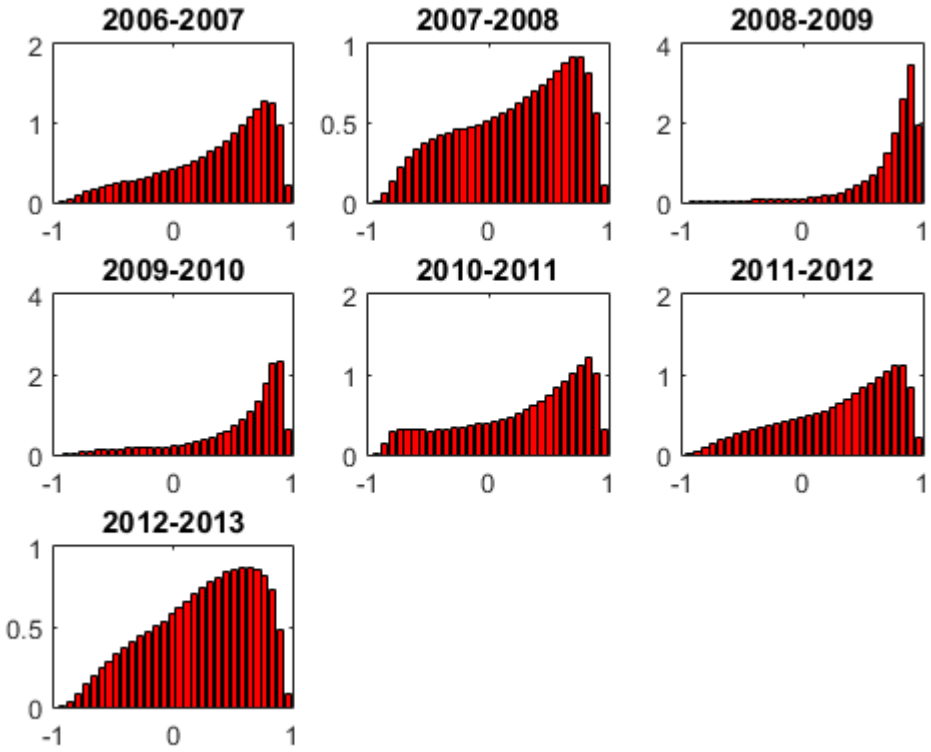


Figure 8: Distribution of correlation coefficients for different years.

Therefore, we can put across the hypothesis:

- "in bad times" the behavior of companies coincide, i.e. they tend to the same behavior, it might be called "Animal spirit"/ "Herd instinct";
- in "good times" the behavior of the companies is different, i.e. each of the company's investors move to his own individually developed strategy.

Thus, let us assume, that by the histograms' shape, we can judge the behavior of the market, in particular, the following indication of "market sentiment" can be provided:

- Pessimistic behavior
- Negative behavior
- Neutral behavior

In the case of the pessimistic sentiment, histograms, having a distribution shape similar to an exponential function, can be taken as a pattern. Histograms, whose distributions are close to normal can be considered as a case of the positive market behavior, and the rest can be attributed to neutral behavior.

4.4 Exclusion of "market trend"

The DJ and S&P500 indices characterize the level and dynamics of stock prices. The time series of these indices are presented in Fig.9. The indices behave almost identically, but the scales are different.

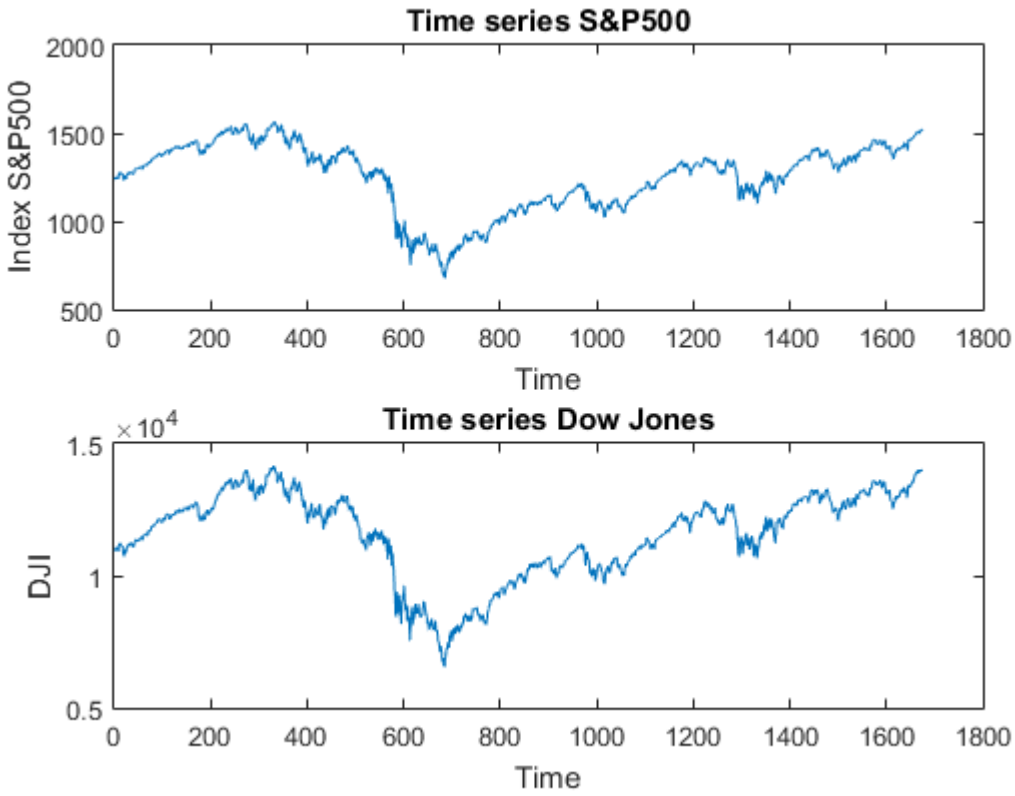


Figure 9: Stock market indexes.

It is known that the stock price is affected by a number of different factors. While it is impossible to identify the exact number as well as universal factors that can influence the sentiment of the price. However, one of such factors, affecting on the formation of stock price is the "market trend", i.e. index (in this case, DJI and S&P500). Therefore, the

"market trend" is excluded from initial stock price, in order to look how the distribution of the correlation matrix would be changed.

For that purpose, a normalization of daily closure prices as well as the normalization of indices were calculated. The normalization by mean and standard deviation was applied with the help of the following formula:

$$\text{normalized}\,\frac{s - \bar{s}}{\sigma(s)} \tag{5}$$

where s initial values, $\bar{s}$ – mean,$\sigma(s)$ – standard deviation of s.

The difference between normalized daily closure prices and indices are calculated as:

$$\text{Diff} = \text{NormPrice-Index} \tag{6}$$

The time series for both cases (S&P500 and DJ) are presented in Fig.10.

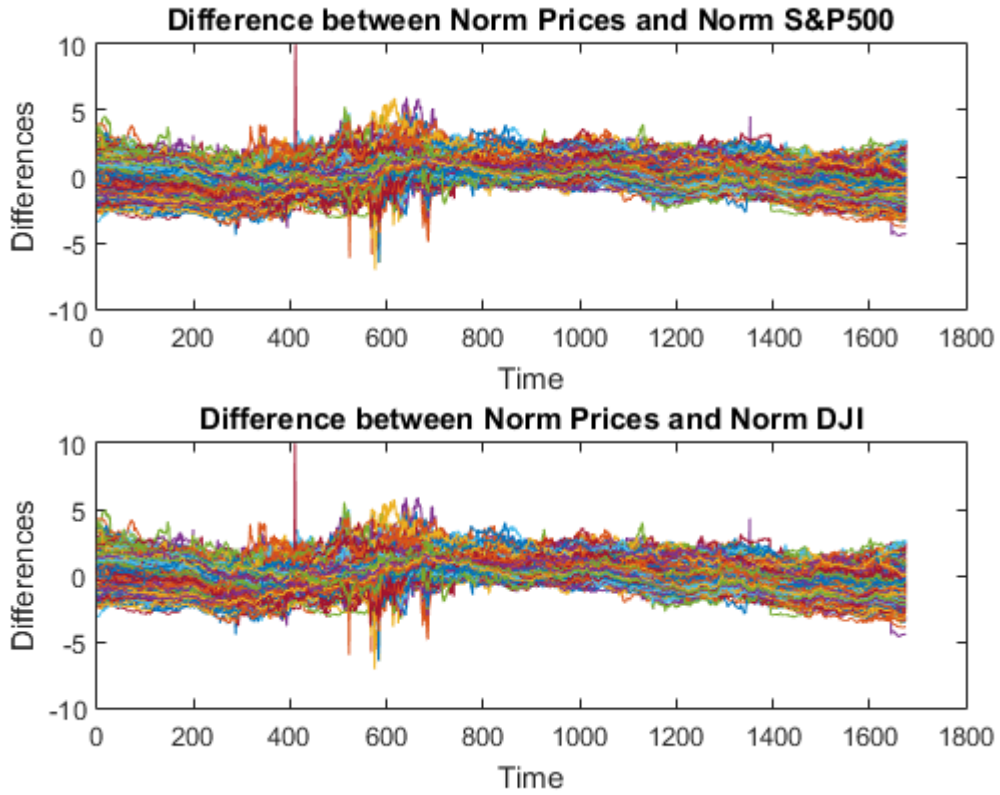


Figure 10: Difference between daily closure price normalized by mean and standard deviation.

The respective histogram for each difference is shown in Fig.11.

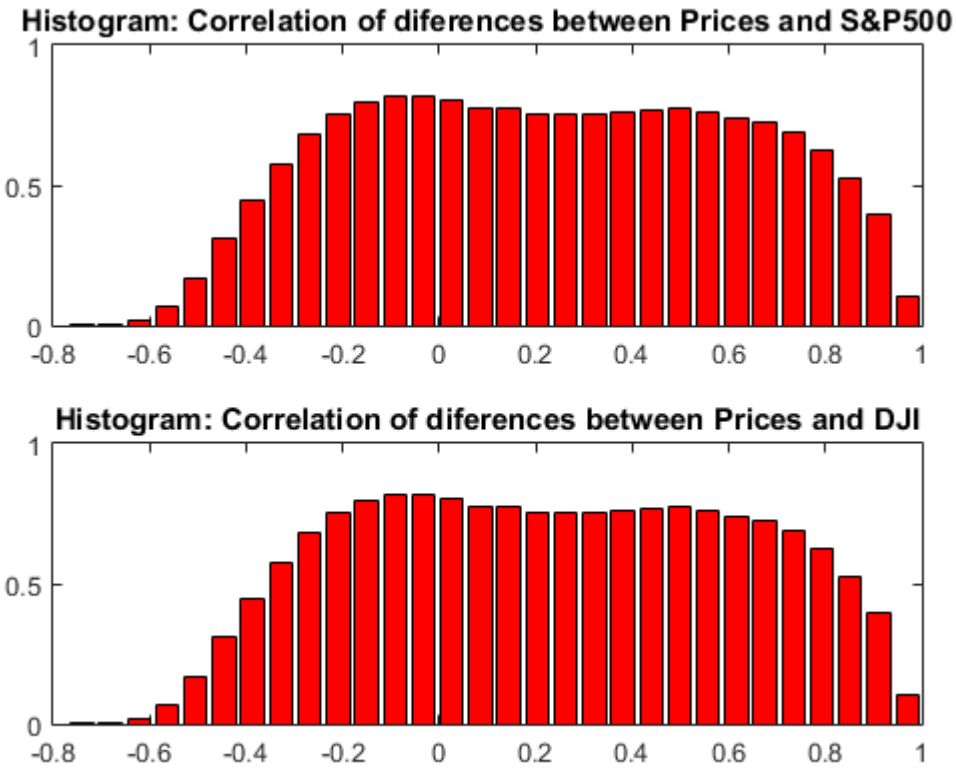


Figure 11: Histograms of difference between daily closure price normalized by mean and standard deviation.

Fig.12 presents the differences between daily closure price normalized by mean and standard deviation for two indexes (DJI and S&P500) in the form of matrix transformed into vectors from negative to positive correlation coefficients.

For this transformation the same algorithm was used as in part 4.2, i.e. matrix correlation coefficients was reshaped into vector and a vector was sorted from negative to positive correlation coefficient.

From Fig.12 it can be seen, that there are many positive coefficients, less negative ones and very few very positive correlation coefficients in each case. Thus, we can say that most of the companies correlate between each other, also in their deviation from index, i.e. from the overall market trend.

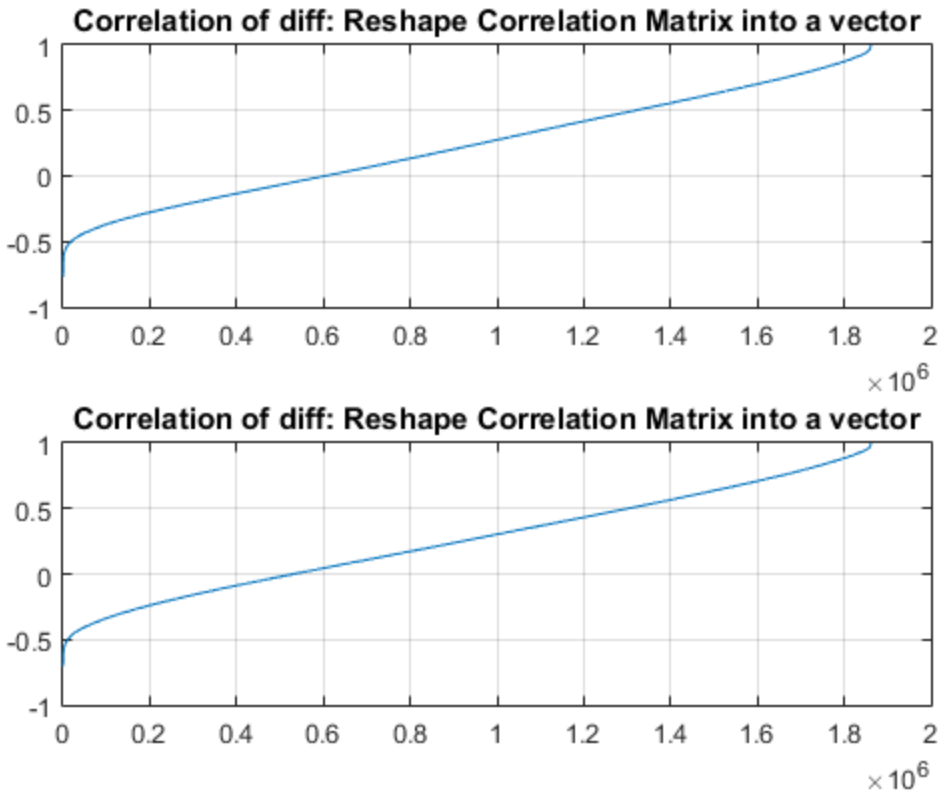


Figure 12: Correlation coefficients of differences in the form of matrices transformed into vectors.

4.5 Distribution selection

As previously stated, the exclusion of the "market trend" was needed in order to look how the distribution of the correlation matrix would be changed. We consider the distributions at one-year-intervals, which are presented in the Fig.8, i.e. the initial data was normalized by DJ and S&P500 indexes, and the same analysis, as it was presented in the part 4.1 was applied to the normalized data, in particular, the calculation of the correlation coefficient matrix.

In Fig.13 and Fig.14 matrix correlation coefficients normalized by DJI and S&P500 indices (or, it might be also called differences) are shown for each year in the period of time from June 2006 to February 2013.

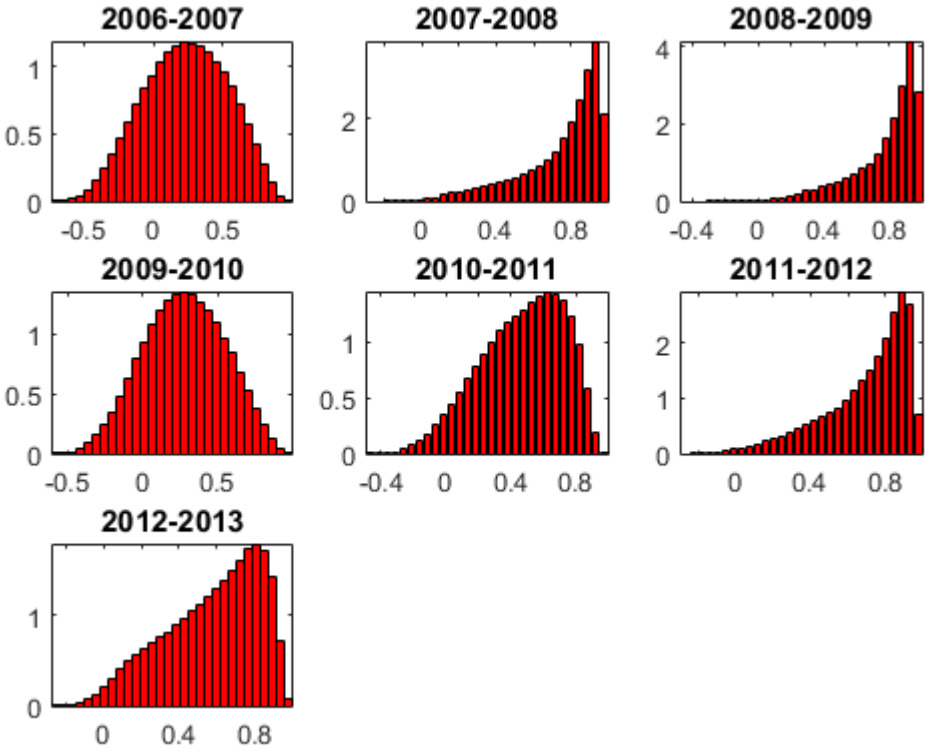


Figure 13: Histograms of correlation coefficients normalized by DJI.

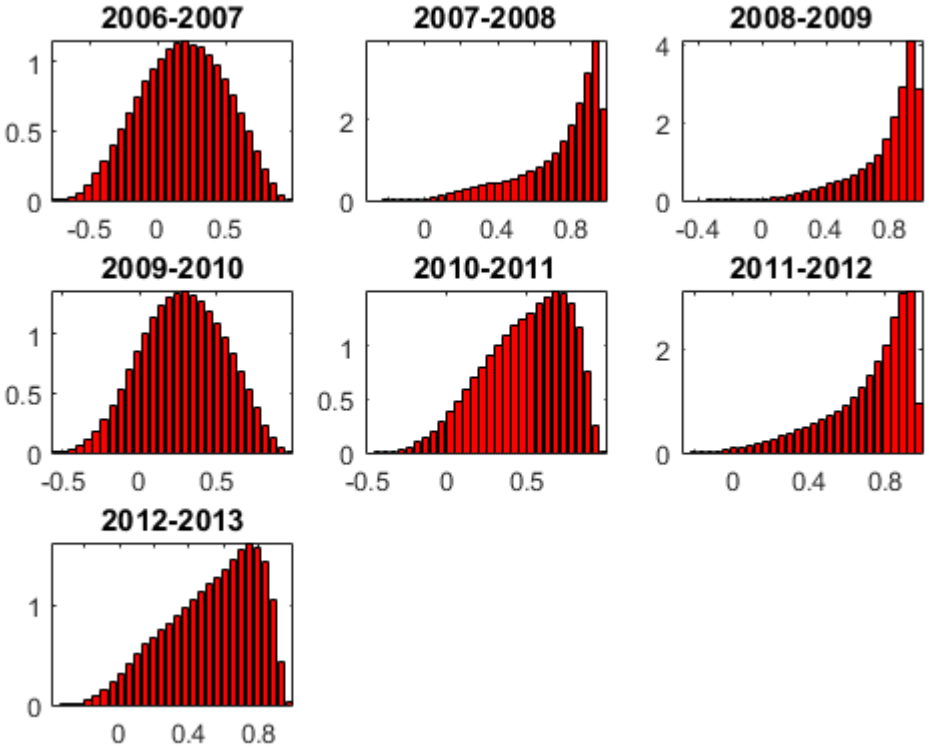


Figure 14: Histograms of correlation coefficients normalized by S&P500.

As it can be seen from Fig.13 and Fig.14, the shapes of the histograms have not changed a lot compared to histograms of initial data splitting into one-year-intervals (Fig.8). In both cases (normalization by DJI and S&P500) the histograms are almost identical, thus, for the further analysis the difference obtained by normalization the initial data by S&P500 is used.

To find the distribution of shape (the bigger mass are shifted to the positive side) seems relatively complicated, thus, the histogram was inverted, as it was done previously, and further we tried to identify the possible suitable distributions. The inverted histograms are presented in Fig.15.

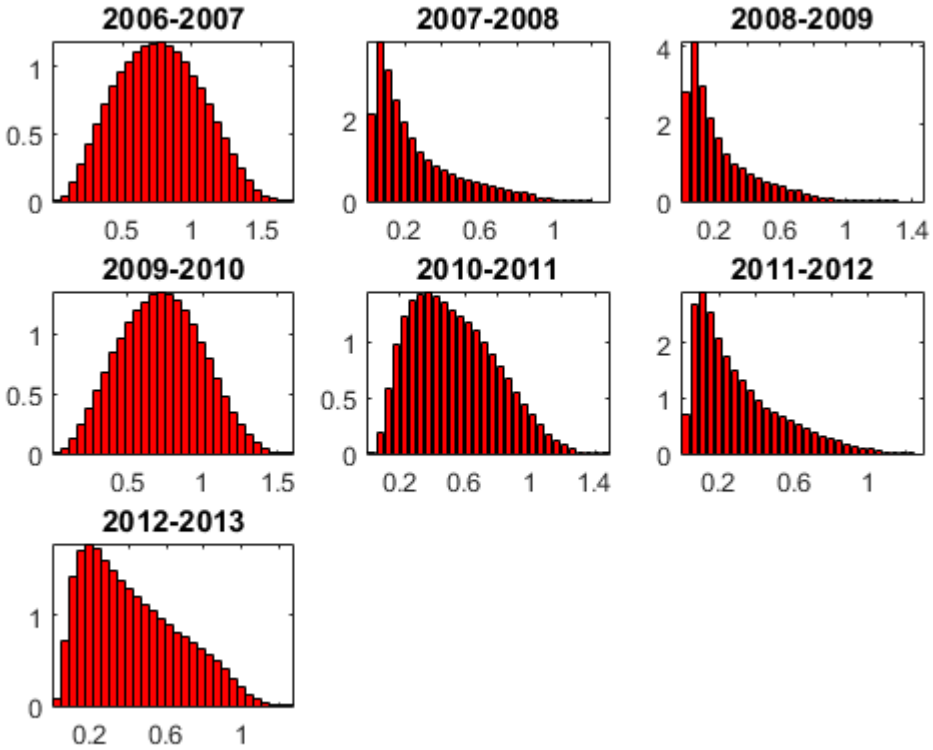


Figure 15: The inverted histograms of correlation coefficient matrix (initial data normalized by DJI).

Returning to assumptions, that by the histograms' shape, we can judge the behavior of the market, it can be seen from Fig.15, that in the case of pessimistic behavior, the sentiment of companies is still more correlated, compared to the case of positive behavior.

Let us consider the distribution for each of the histograms from Fig.15. In Fig.16 the appropriate distribution of histograms from Fig.15 is shown. We shall take a closer look at the following three candidate distributions: Weibull, Birnbaum-Saunders and Nakagami distributions.

The Weibull distribution and its parameters have already described earlier in section 4.1 of this thesis. In particular, Weibull distribution, as the most appropriate for the description of histograms is presented in Fig.16a, Fig.16d and Fig.16g. The parameters of the distributions are presented in Table 2. The years that have the Weibull distribution, might be taken as an example of positive behavior.

Year	$k-$ shape	λ-scale
2006 – 2007	2.78965	0.86296
2009 – 2010	2.94043	0.81045
2012 – 2013	1.77389	0.47822

Table 2: Parameters of Weibull distribution for different years.

The Birnbaum-Saunders distribution is a continuous probability distribution with two parameters. The probability density function is:

$$F(x;\alpha;\beta) = \Phi\left(\frac{1}{\alpha}\left[\left(\frac{x}{\beta}\right)^{0.5} - \left(\frac{\beta}{x}\right)^{0.5}\right]\right) \tag{7}$$

where α – shape parameter, β – scale parameter.

Birnbaum-Saunders distribution, as the most appropriate for the description of histograms is presented in Fig.16b, Fig.16c and Fig.16f. The parameters of the distributions are presented in Table 3. The years that have the Birnbaum-Saunders distribution, might be taken as an example of negative behavior.

Year	α-shape	β-scale
2007 – 2008	1.05853	0.15994
2008 – 2009	1.06611	0.14314
2011 – 2012	0.90002	0.21321

Table 3: Parameters of Birnbaum-Saunders distribution for different years.

The Nakagami distribution is a continuous probability distribution with two parameters. The probability density function is:

$$F(x; m; \Omega) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \exp\left(-\frac{m}{\Omega}x^2\right), \forall x \geq 0. \left(m \geq \frac{1}{2}, \Omega > 0\right) \quad (8)$$

where m – shape parameter, Ω – scale parameter.

Birnbaum-Saunders distribution appears to be the most appropriate for the description of histograms presented in Fig.16e.The parameters of the distributions are presented in Table 4. The years that have the Nakagami distribution, might be taken as examples of neutral behavior.

Year	m-shape	Ω- scale
2007 – 2008	1.14939	0.34772

Table 4: Parameters of Nakagami distribution for different years.

All of the mentioned distributions (Weibull, Birnbaum-Saunders, Nakagami distributions) are particular cases of the gamma distribution. Gamma distribution can be parameterized in terms of a shape parameter $\alpha = k$ and inverted scale parameter $\beta = 1/\theta$, called a rate parameter. If k is a positive integer, then the sum of k independent exponentially distributed random variables, each of which has a mean of θ two-parameter family of continuous probability distribution. A random variable X that is gamma-distributed with shape parameters α and β is denoted:

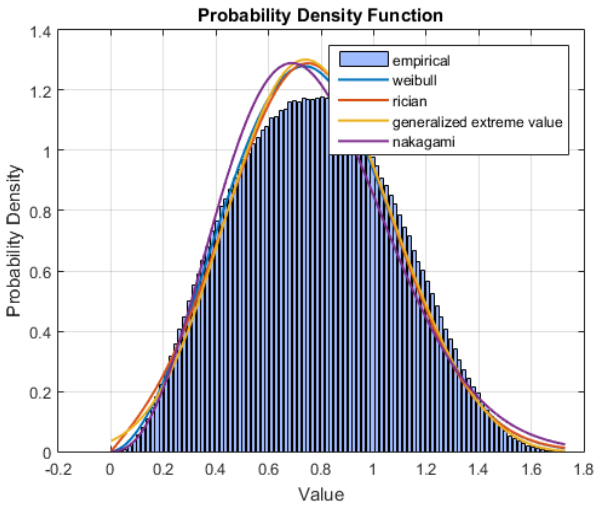
$$X \sim \Gamma(\alpha, \beta) \equiv \text{Gamma}(\alpha, \beta) \quad (9)$$

The probability density function is:

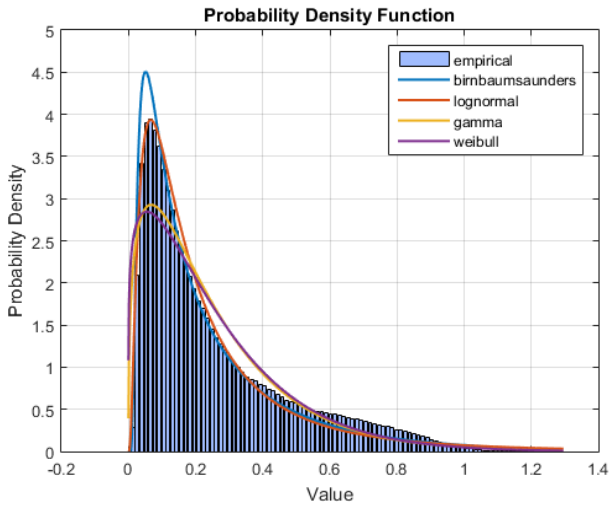
$$F(x; \alpha; \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad (10)$$

where $x > 0$ and $\alpha, \beta > 0$; $\Gamma(\alpha)$ is a complete gamma function

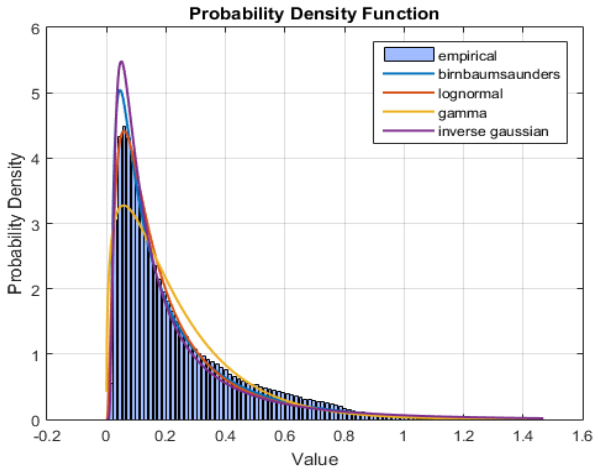
(a) 2006-2007



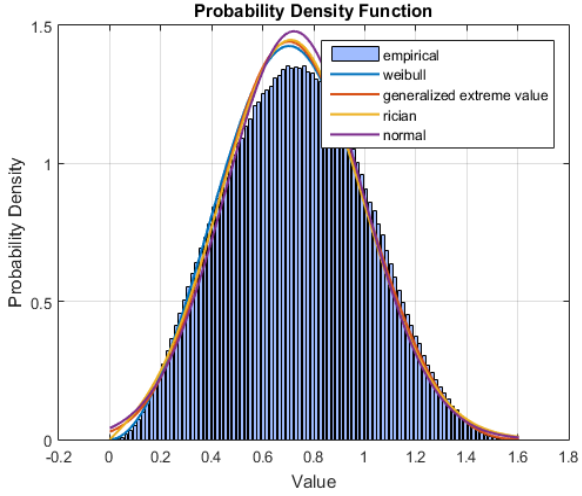
(b) 2007-2008



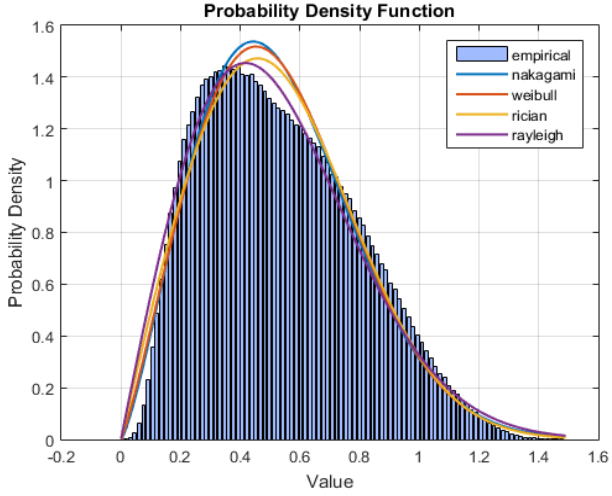
(c) 2008-2009



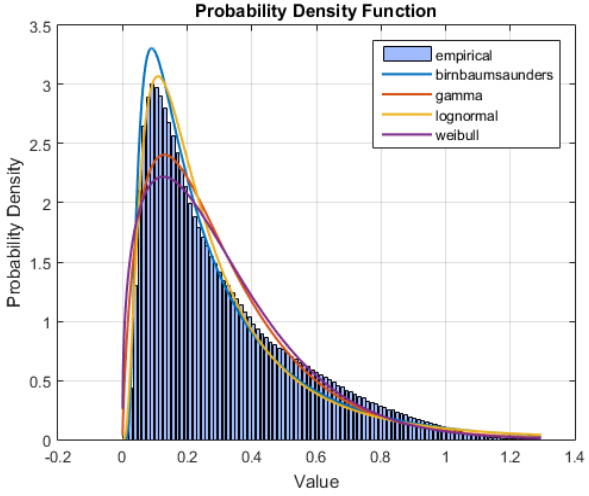
(d) 2009-2010



(e) 2010-2011



(f) 2011-2012



(g) 2012-2013

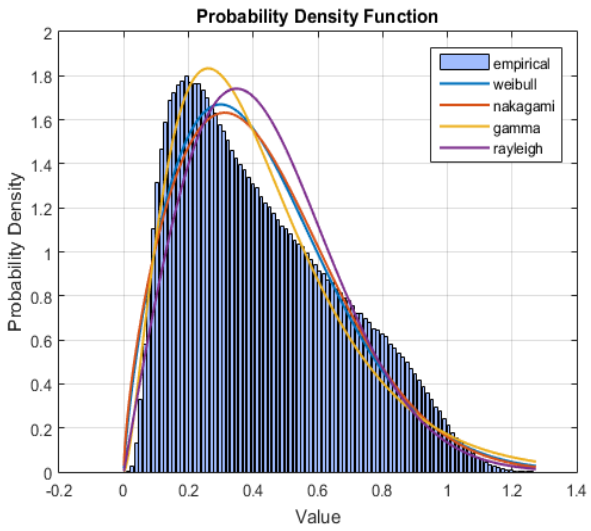


Figure 16: The distributions of inverted histograms of correlation coefficient matrix (initial data normalized by DJI).

4.6 Correlation as a function of time

Initial data normalized by S&P500 for the period from June 2006 to February 2013 are divided in such a way that 1 year, i.e. 250 days – the average number of working days of a stock exchange is shifted by 1 month, approximately, 22 days; it might be called a "moving one-year-window". In Fig.17 the histograms of all "moving one-year windows" are presented.

From Fig.17 some regularity can be seen. Earlier, it was assumed that the histograms shifted to the right characterize pessimistic behavior of the market, whereas histograms, whose distributions are close to normal can be considered as a case of positive market behavior. However, the confirmation on real data is necessary.

Thus, let us assume that the S&P500 index reflects the behavior of the market, since it consists of 500 of the largest American companies. The result is performed in Fig.18. Blue dotted lines highlight a rise period, red dotted lines show a recession period. The central parts of each of these periods indicate as a black window. We are considering the central part, when comparing the results from Fig.17 and Fig.18. A green line shows the trend (rise and fall).

Based on it, let us consider the main periods:

- Positive – increasing stock market index
- Negative – decreasing stock market index
- Neutral

In the period where the index is increasing strongly (positive period) or decreasing sharply (negative period) indicate strong market sentiment, so that buyers are bullish about buying stocks, in the first case, or, in the second case, stocks are sold off because of fear of losses. As a counterpart to that there might be more stable periods where the increase/decline is not so strong (slope is smaller in absolute terms), neutral period. When we look at bad market sentiment from 2008-2009 years and take that as a reference for extreme (pessimistic) sentiment, 2012-2013 years as an increasing (positive) sentiment, and, in the case of neutral period, 2007-2008 years can be taken as a pattern.

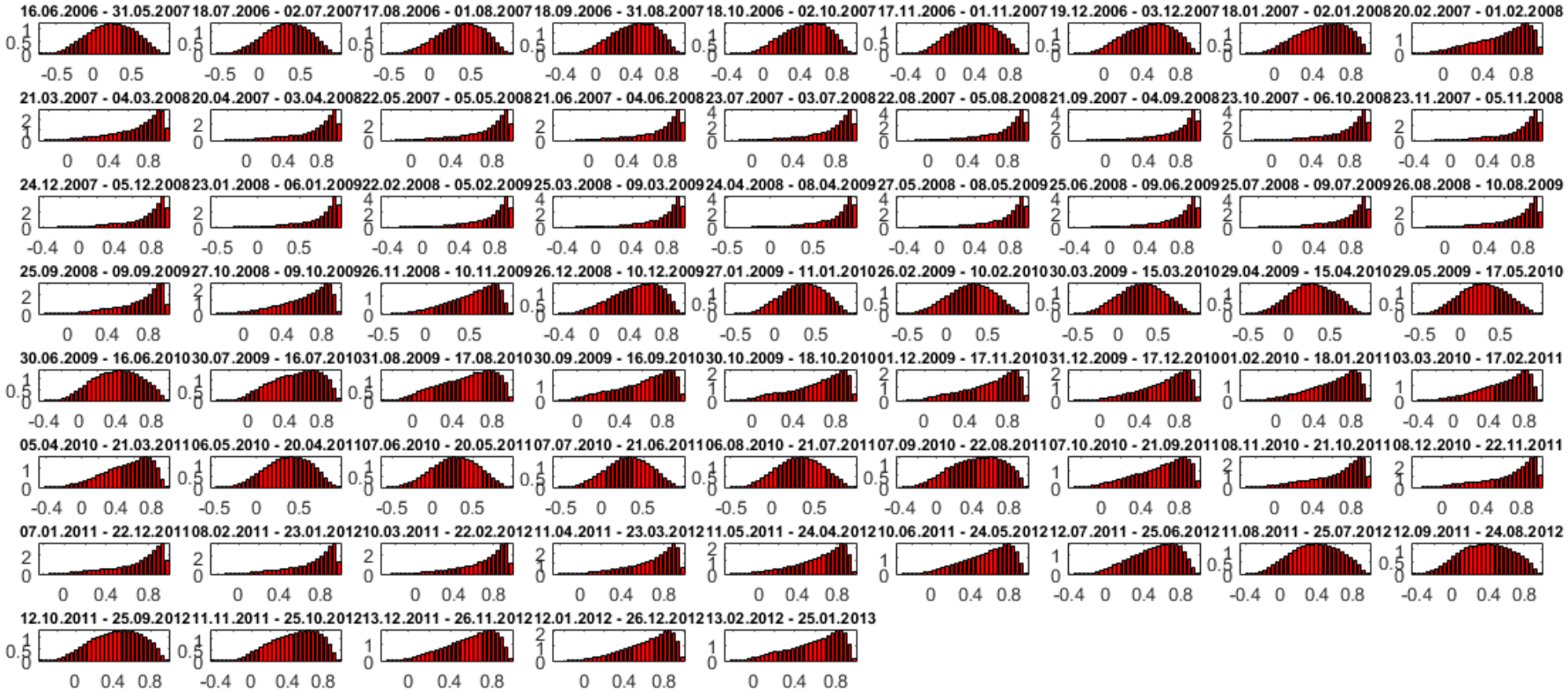


Figure 17: Histograms of a “moving one-year-window”.

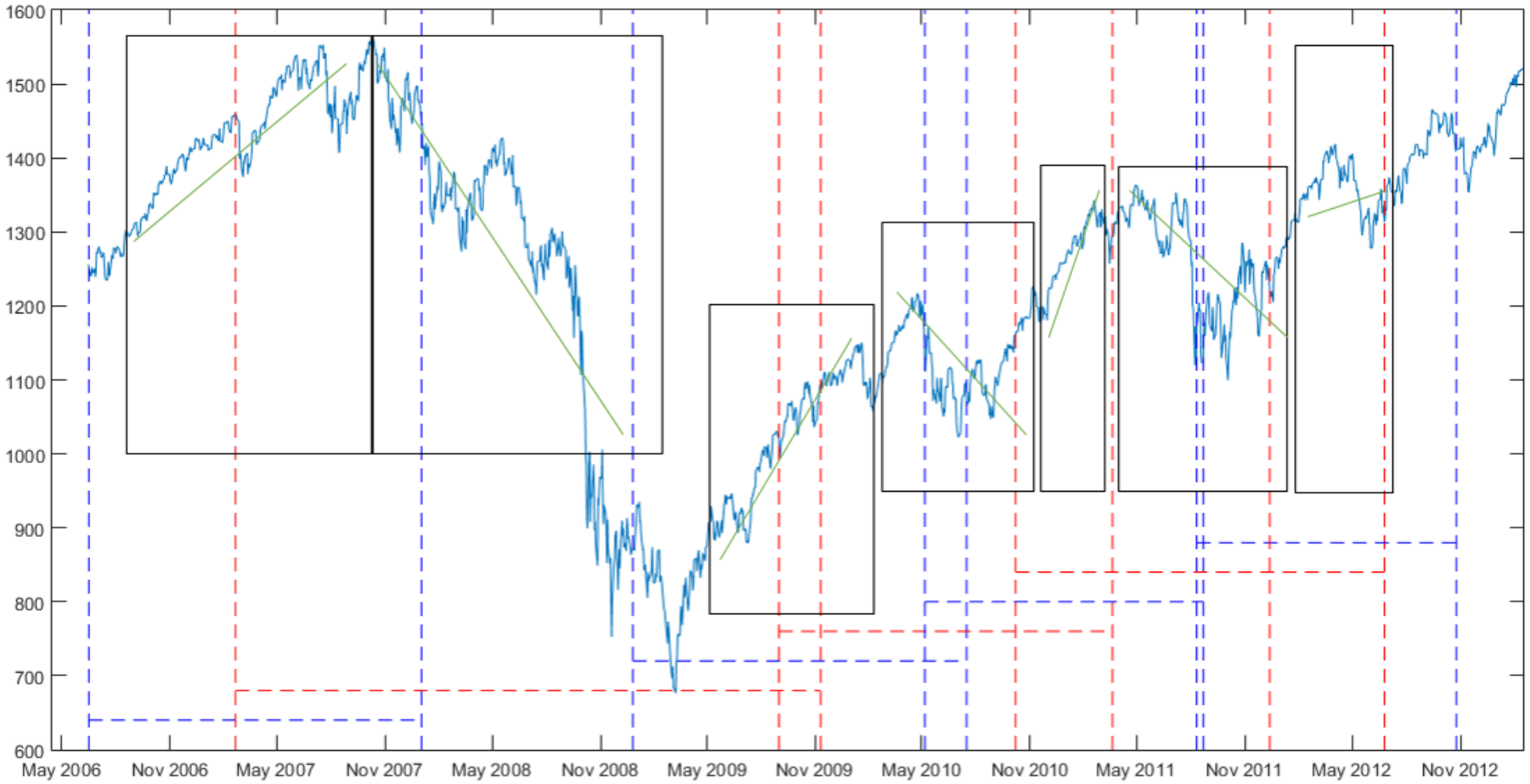


Figure 18: Trends per period.

4.7 Classifier

Based on the results shown in Fig.17 and Fig.18, we identify the rise and decline periods, i.e. the moments, when there is a positive or a negative period, respectively. As the decline period the following histograms are selected: 9 – 30, 39 – 46, 52 – 61, 66 – 68. The rest of the histograms are assigned to the neutral/positive periods.

In previous part 4.5 we chose the distribution that describes the histograms of differences, represented in the form of a correlation matrix, that is the gamma-distribution. The gamma-distribution parameters for each of the histograms from Fig.17 are calculated, i.e. each histogram has its own pair of parameters α – shape, β – scale. Let us depict the obtained results graphically in Fig.18: x-axis corresponds to parameter α, y-axis corresponds to β parameter of gamma distribution.

Based on the obtained parameters, we try to construct a "classifier", which aim is to reveal the various periods in the economy, i.e. rise/decline periods. As it might be seen from Fig.19, the result of graphically interpreted parameters of gamma-distribution is an elongated cloud, thus, the resulting data represent a statistical dependence. Therefore, it is very likely to derive some regularity/pattern, that can describe this dependence. As a model the linear regression can be constructed from the obtained data, Fig. 20.

The model of linear regression is:

$$\beta = f(\alpha, c) \tag{11}$$

where β – response, α – explanatory variable, c – model parameters.

In our case the Eq.11 looks the following:

$$\beta = c_1 + c_2\alpha \tag{12}$$

where $c_1 = 0.2036638$, $c_2 = -0.0146082$

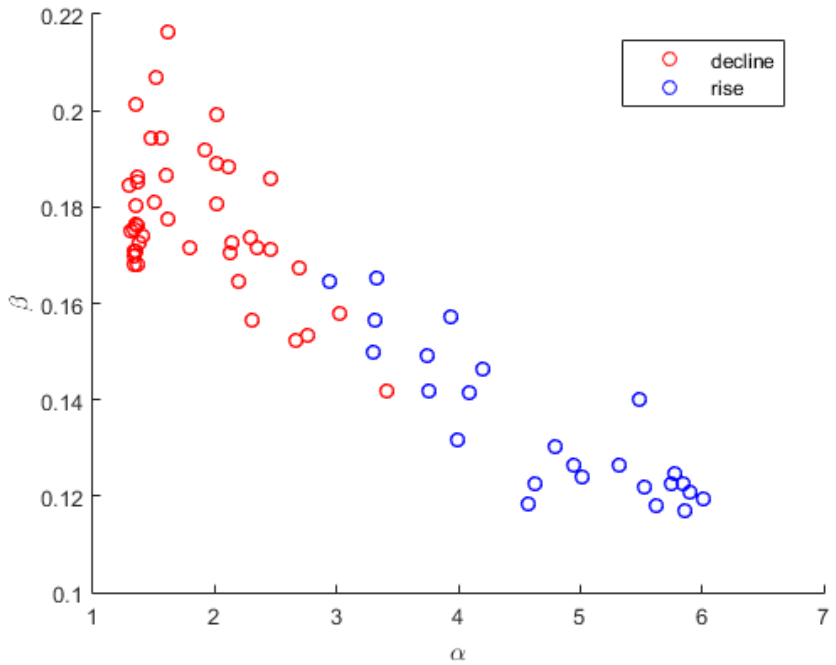


Figure 19: Parameters of Gamma-distribution from histograms.

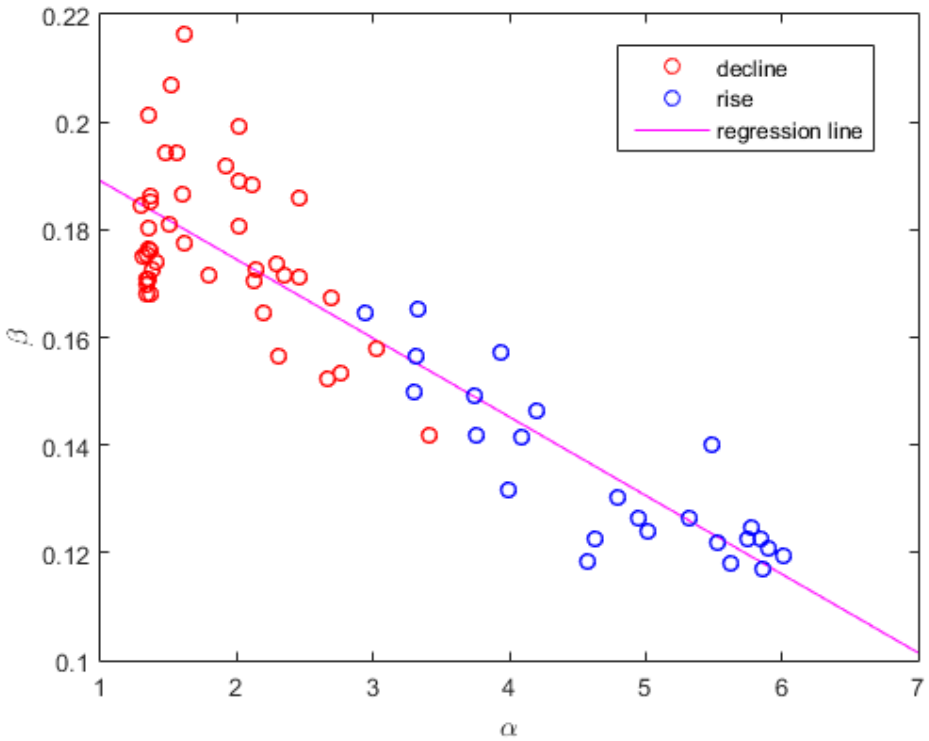


Figure 20: Parameters of Gamma-distribution from histograms and Linear regression line.

Let us look how other functions are appropriate in describing the resulting data. In Fig.21 the linear, exponential, power regression models are compared to each other. Data about model fitting is presented in Table 5.

The exponential regression model is:

$$\beta = k_1 \cdot e^{k_2 \alpha} \tag{13}$$

where $k_1 = 0.2103$, $k_2 = -0.09513$

The power regression model is:

$$\beta = k_1 \cdot e^{k_2 \alpha} \tag{14}$$

where $p_1 = 0.2027$, $p_2 = -0.2575$

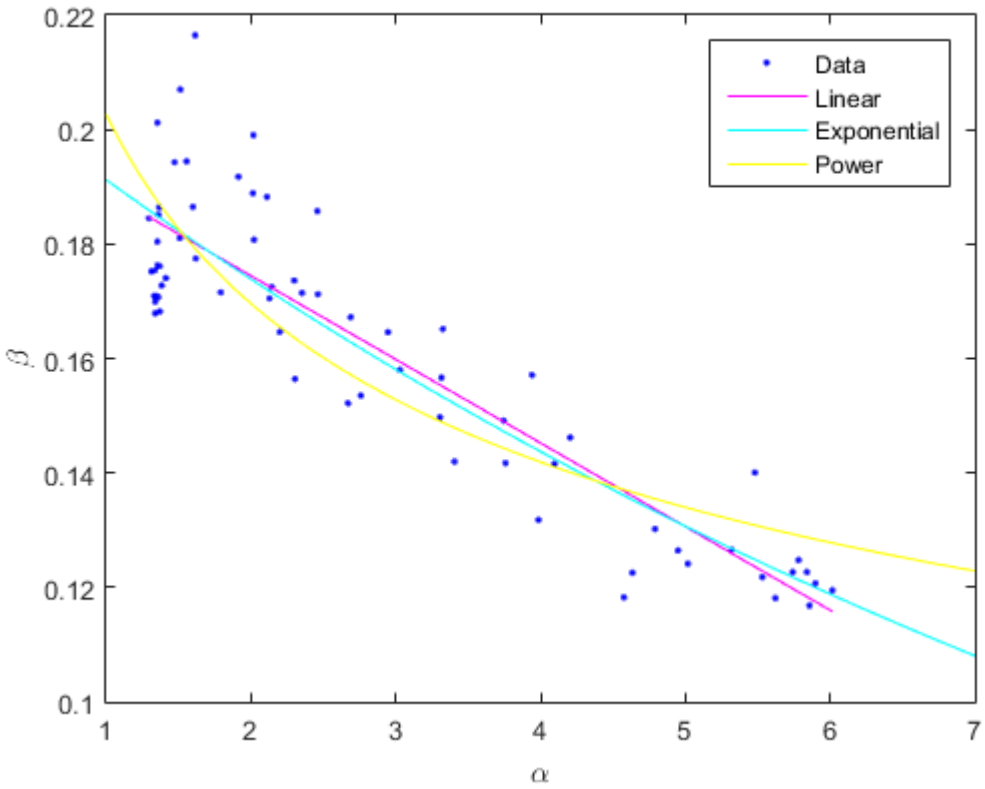


Figure 21: Various models describing the data.

	Linear	Exponential	Power
SSE	0.008255	0.008387	0.011223
R-square	0.811725	0.808736	0.744059
DFE	66	66	66
R-square (adjusted)	0.808872	0.805838	0.740181
RMSE	0.011184	0.011272	0.013041

Table 5: Goodness of Fit.

Sum of Squared Errors (SSE) is a measure of how the model fits; a poor fitting model has much larger errors and thus the larger SSE. Therefore, the power regression model has the worst fitting with the data, as its SSE is the largest compare to other models. SSE of linear regression model is the smallest one, i.e. this model has the best fit of data compared to other represented models.

Coefficient of determination (R-square) is a statistical measure of how close the data are to the fitted regression line. In the linear regression model R-square is the highest – 81% of all represented models, i.e. the linear regression model is the most consistent with the data, in comparison with other models.

In addition, the RMSE in linear regression models is the smallest one, compared to the exponential and power regression models. Thus, it might be assumed, that the linear regression model has the best fit of this data, compared to exponential and power regression models and it might be chosen for further analysis. In particular, it is used in the "classifier" construction.

Let us draw a straight line through the points in the place where there is a mixture of colors, i.e. the periods of rise and decline are connected, Fig.22.This line can be called "classifier". Thus, the points on the left side (red circles) indicated that the economy is in negative period (decline), and the points on the right side (blue circles) indicate that the economy is in the neutral or positive period. Fig.23 shows the same results, instead of circles the lines are used, the line takes the color of the point at which it goes (rise/decline).

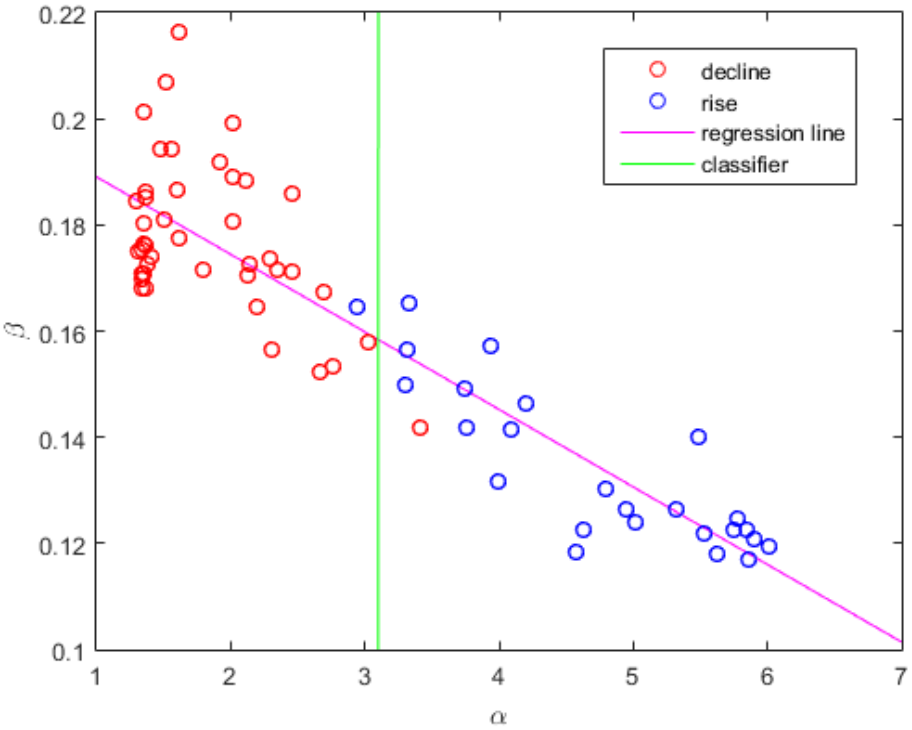


Figure 22: “Classifier” – linear regression.

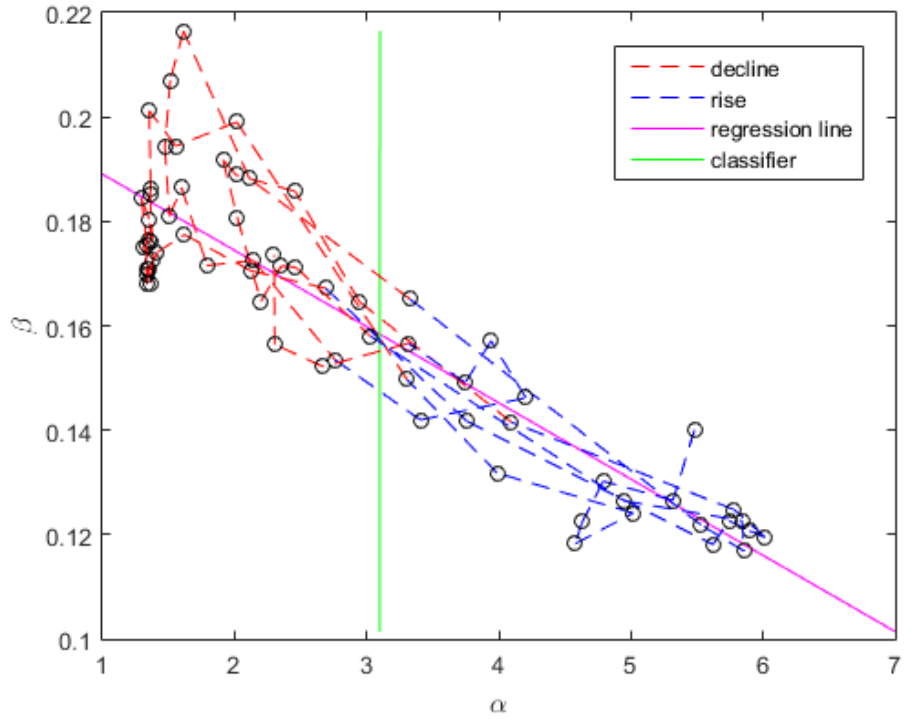


Figure 23: “Classifier” – lines.

5 FUTURE WORK

As further research, the following points can be implemented for this work:

- Distribution: instead of gamma-distribution used in this study, as a universal distribution for calculation of "classifier" parameters, it might be potential to choose another distribution or family of distributions with subsequent calculation of its parameters and realization of the performed analysis.

- "Moving periods": when dividing the differences obtained by normalizing the initial data for the period from June 2006 to February 2013 by S&P500, the term "moving one-year window" was used. It means that initial data are divided in such way that 1 year is shifted by 1 month. The magnitude of 1 month can vary, e.g. from 3 months to half a year; it may also be interesting to shift not 1 year, but e.g. 3-6 months.

- Model: as a regularity/pattern describing the data, another model whose equation would be compiled and derived independently, based on known results might be used, instead of the linear regression model.

6 CONCLUSIONS

The aim of this work is the creation of a simple model, called a "classifier", capable of describing the current state of economy, i.e. rise and decline periods, based on the data from NYSE for the period from June 2006 to February 2013, normalized by DJ and S&P500 indices.

The most significant findings, obtained from this research, are as follows:

- Distribution of correlation matrix, derived from initial data, is shifted to the right hand side with respect to the zero. The most appropriate distribution, describing the result is Weibull distribution.

- During the normalization of the initial data, i.e. exclusion of "market trend", it was concluded that the form of distributions remained unchanged.

- The resulting differences were divided into smaller periods, for each period the histograms were constructed, and for each of the histograms the corresponding distribution was identified. Based on these distributions one common distribution (gamma-distribution), describing the presented histograms, was selected.

- The consecutive shift of one year by one month, allowed us to divide data for a shorter period, and to construct the histograms, for each of which the gamma-distribution parameters were calculated.

- The model describing the obtained data was selected. Based on it a simple "classifier" enabling to identify the rise and decline period in the economy was created.

REFERENCES

[1] A.Zeleva, Forecasting Financial Weather – Can we foresee Market Sentiment? Spectrum of Stock Price Behavior – NYSE Case Study. LUT, 2015.

[2] T.Uwimanayantumye, Static Waves in Corporate Space: Characterizing Oscillating Trading Patterns in New York Stock Exchange.LUT,2016

[3] W. Huang, X.Zhaung, S.Yao, A network analysis of the Chinese stock market. Physica A: Statistical Mechanics and its Applications, pp. 2956-2964, 2009.

[4] R.N.Mantegna and H.E.Stanley, Introduction to econophysics. Correlations and Complexity in Finance. Cambridge: The Press Syndicate of the University of Cambridge, 2000.

[5] A. Damodaran, Successful Strategies and the Investors Who Made Them Work. John Wiley&Sons, Inc, pp 175-240, 2003.

[6] D.Sornette, Why Stock Markets Crash: Critical Events in Complex Financial Systems. Princeton University Press, pp. 49- 60, 2003.

[7] W.Tease, The stock market and investment. OECD Economic Studies No. 20, 2003.

[8] A.Ang, G. Bekaert. International Asset Allocation with Time-Varying Correlation. National Bureau of Economic Research, Cambridge, 1999.

[9] Y.Yuan, X.Zhuang, Z.Liu, Price-volume multifractal analysis and its application in Chinese stock markets. Physica A: Statistical Mechanics and its Applications. pp.3484-3495, 2012.

[10] A.Peresetskiy, Autocorrelation in Global Stochastic Trend. Moscow: Applied Econometrics, 2014

[11] G.Prosvetov, "Risk Management: Tasks and Solutions". Moscow: Alpha-Press, 2008

[12] T.C. Mill, R.N. Markellos, The Econometric Modelling of Financial Time Series. Cambridge: Cambridge University Press, 2008.

[13] C.Kirkpatrick, J.Dahlquist, Technical Analysis: The Complete Resource for Financial Market Technicians. John Wiley&Sons, 2011

LIST OF TABLES

LIST OF FIGURES