



Artificial Intelligence in Finance:

Forecasting Stock Market Returns Using Artificial Neural Networks

Alexandra Zavadskaya

Department of Finance

Hanken School of Economics

Helsinki

2017

HANKEN SCHOOL OF ECONOMICS

Department of: Finance	Type of work: Thesis
Author: Alexandra Zavadskaya	Date: 29.09.2017
Title of thesis: Artificial Intelligence in Finance: Forecasting Stock Market Returns Using Artificial Neural Networks	
Abstract: <p>This study explored various Artificial Intelligence (AI) applications in a finance field. It identified and discussed the main areas for AI in the finance: portfolio management, bankruptcy prediction, credit rating, exchange rate prediction and trading, and provided numerous examples of the companies that have invested in AI and the type of tasks they use it for.</p> <p>This paper focuses on a stock market prediction, and whether Artificial Neural Networks (ANN), being proxies for Artificial Intelligence, could offer an investor more accurate forecasting results. This study used two datasets: monthly returns of S&P500 index returns over the period 1968-2016, and daily S&P 500 returns over the period of 2007-2017. Both datasets were used to test for univariate and multivariate (with 12 explanatory variables) forecasting. This research used recurrent dynamic artificial neural networks and compared their performance with ARIMA and VAR models, using both statistical measures of a forecast accuracy (MSPE and MAPE) and economic (Success Ratio and Direction prediction) measures. The forecasting was performed for both in-sample and out-of-sample. Furthermore, given that ANN may produce different results during each iteration, this study has performed a sensitivity analysis, checking for the robustness of the results given different network configuration, such as training algorithms and number of lags.</p> <p>Even though some networks have outperformed certain linear models, the overall result is mixed. ARMA models were the best in minimizing forecast errors, while networks often had a better accuracy in a sign or direction prediction. Artificial neural networks outperformed respective VAR models in many parameters, but the difference of this outperformance was not significant, measured by S-test.</p> <p>Moreover, all model produced significant results in accurate stock index direction predictions, as 75-80% of the time the models predicted correct direction change in S&P500 return.</p> <p>Furthermore, combining Artificial Intelligence with the concept of Big Data, Google Trends search terms were used as a measure of a market sentiment. This study finds the use of Google trends search terms valuable in the modeling of S&P 500 returns.</p>	
Keywords: forecasting, stock returns, S&P500, market sentiment, ANN, artificial neural networks, artificial intelligence, forecasting accuracy, forecast evaluation	

CONTENTS

1	INTRODUCTION	8
1.1	Purpose of the study	9
1.2	Study limit.....	9
1.3	Contribution.....	10
1.4	Thesis organization.....	11
1.5	List of abbreviations	12
2	BACKGROUND INFORMATION	13
2.1	What is AI	13
2.2	Applications of AI in finance	14
2.2.1	Anomaly detection.....	14
2.2.2	Portfolio management and robo-advisory	14
2.2.3	Algorithmic trading	15
2.2.4	Text mining.....	16
2.2.5	Market, sentiment or news analysis	17
2.2.6	Credit evaluation and loan/insurance underwriting.....	17
2.3	Types of Artificial Intelligence (AI)	18
2.4	Artificial Neural Network (ANN).....	18
2.4.1	Types of ANN.....	18
3	THEORY REVIEW	22
3.1	Traditional Investment Theories	22
3.2	Behavioral finance	22
3.3	Adaptive Market Hypothesis (AMH).....	24
3.4	Chaos Theory	24
3.5	Time-series momentum theory.	25
3.6	Forecasting methods.....	25
3.6.1	Linear vs. nonlinear methods.....	26
3.6.2	Benefits of ANN	26
3.6.3	ANN drawbacks.....	27
3.6.4	Forecast Combinations.....	27

4	LITERATURE REVIEW	29
4.1	Bahrammirzaee's (2010) review of previous studies of ANN applications	29
4.2	Individual papers review	30
5	DATA.....	38
5.1	Dataset 1 description.....	38
5.1.1	Variables	39
5.1.2	Descriptive statistics	44
5.1.3	Correlations	51
5.2	Dataset 2 description	53
5.2.1	Variables	53
5.2.2	Descriptive statistics.....	57
5.2.3	Correlation analysis.....	59
6	RESEARCH METHOD	61
6.1	Research Question and Research Hypothesis:	61
6.2	Traditional long-term forecasting methods.....	61
6.2.1	ARIMA.....	61
6.2.2	VAR.....	63
6.2.3	In sample forecasting versus out-of-sample	64
6.3	Research method for ANN.....	64
6.3.1	Process of constructing and utilizing ANN	64
6.3.2	NAR network	67
6.3.3	NARX network.....	69
6.4	Method for comparing the model's forecast accuracy	70
6.4.1	Statistical loss functions	70
6.4.2	Financial/economic loss function	72
6.5	Statistical hypotheses.....	75
7	EMPIRICAL RESEARCH	77
7.1	ARIMA modeling for monthly dataset	77
7.1.1	ARMA (5,2) model	79
7.2	ARIMA modelling for daily dataset	81
7.2.1	ARMA (4,5) model.....	82
7.3	NAR network	83

7.3.1	NAR network for monthly data set.....	83
7.3.2	NAR network for daily data set	86
7.4	VAR model.....	88
7.4.1	VAR modelling for monthly data set	88
7.4.2	VAR modelling for the daily dataset.....	91
7.5	NARX network.....	94
7.5.1	NARX network for monthly dataset.....	94
7.5.2	NARX network modelling for the daily dataset	96
8	MODEL DIAGNOSTICS.....	98
8.1	Model diagnostics summary	98
8.2	Model diagnostics ARMA (5,2).....	101
8.2.1	Model stationarity	101
8.2.2	Residual diagnostics	101
8.2.3	Normality test.....	102
8.2.4	Heteroscedasticity test	103
8.3	Model diagnostics ARMA (4,5).....	103
8.3.1	Model stationarity	103
8.3.2	Residual diagnostics	103
8.3.3	Normality test.....	104
8.3.4	Heteroscedasticity test	104
8.4	Network diagnostics NAR (2), LM training, monthly dataset.....	104
8.4.1	Test for autocorrelation of the residuals	105
8.4.2	Normality test.....	105
8.5	Network diagnostics of NAR (12), BR training, monthly dataset.....	106
8.5.1	Test for autocorrelation of the residuals	106
8.5.2	Normality test.....	107
8.6	Network diagnostics NAR (2), LM training, daily dataset	107
8.6.1	Test for autocorrelation of the residuals	107
8.6.2	Normality test.....	108
8.6.3	Correlation between series at time t and error at time t.	109
8.7	Model diagnostics VAR (2) monthly dataset	109
8.7.1	Test for VAR stability/Inverse roots of AR characteristic polynomial.....	109
8.7.2	LM test for serial correlation.....	110

8.7.3	Normality test	110
8.8	Model diagnostics VAR (2) daily dataset	110
8.8.1	Test for stability of VAR model	110
8.8.2	Normality test	111
8.8.3	Portmanteau test for autocorrelation	111
8.8.4	LM test for serial correlation	111
8.8.5	Heteroscedasticity test	112
8.9	Model diagnostics VAR (6) daily data set	112
8.9.1	Test for stability of VAR model	112
8.9.2	Normality test	112
8.9.3	Portmanteau test for autocorrelation	113
8.9.4	LM test for serial correlation	113
8.9.5	Heteroscedasticity test	113
8.10	Network diagnostics NARX (2), LM training, monthly dataset	113
8.10.1	Test for autocorrelation of the residuals	114
8.10.2	Correlation between input and error	114
8.11	Network diagnostics NARX (2), LM training, daily dataset	115
8.11.1	Test for autocorrelation of the residuals	115
8.11.2	Normality test	115
9	RESULTS	117
9.1	In sample fit	117
9.2	Out of sample prediction accuracy	119
9.2.1	Artificial neural networks versus traditional models	119
9.2.2	Picking the best model	122
9.3	Sensitivity analysis	122
9.3.1	Sensitivity of results given a number of iterations	123
9.3.2	Sensitivity to number of lags	123
9.3.3	Sensitivity to training algorithms	123
9.4	Effect of market sentiment variables on stock return	125
9.4.1	Market sentiment, monthly dataset	126
9.4.2	Market sentiment, daily dataset	126
9.5	Effect of explanatory variables on S&P500 return	129

10	DISCUSSION.....	131
10.1	Answering the Research question.....	131
10.2	What to learn from companies that implement AI.....	132
10.2.1	Credit card, banks and major financial institutions.....	132
10.2.2	Portfolio management firms	133
10.2.3	Investment and trading companies.....	133
10.3	Choosing the best model.....	134
10.4	Importance of forecasting.....	134
10.5	Usefulness of Big Data in predicting stock returns	135
10.6	Non-linear vs linear	136
10.7	Caution in network configuration.....	136
10.8	Statistical versus Economic measures	137
11	CONCLUSION.....	138
11.1	Implications	138
11.2	Limitations.....	139
11.3	Suggestion for further research	140
	REFERENCES	143

APPENDICES

Appendix 1	Full correlation matrix for daily dataset	151
------------	---	-----

TABLES

Table 1	Summary of previous research on ANN forecasting returns	29
Table 2	Summary of previous literature individual papers	30
Table 3	Potential independent variables	39
Table 4	Descriptive statistics summary	45
Table 5	Variables for the first dataset	52
Table 6	Correlation summary	52
Table 7	Summary of potential factors useful for predicting the S&P500 return.....	54
Table 8	Descriptive statistics summary for the daily dataset	58
Table 9	Correlation matrix of variables for daily dataset	60

Table 10	Table of statistical hypotheses	76
Table 11	ACF and PACF illustration for monthly dataset	78
Table 12	Summary of information criteria coefficients, monthly dataset.....	79
Table 13	ARMA (5,2) model output, monthly dataset.....	80
Table 14	ACF and PACF, daily dataset	81
Table 15	ARMA (4,5) model output, daily dataset	82
Table 16	Lag length selection using information criteria, monthly dataset.....	88
Table 17	Lag exclusion test, VAR (2), monthly dataset	89
Table 18	VAR (2) model output, monthly dataset.....	90
Table 19	Lag length selection using information criteria, daily dataset.....	91
Table 20	Lag exclusion test, VAR (2), daily	92
Table 21	VAR (2) model output, daily dataset.....	93
Table 22	ACF and PACF of the residuals of ARMA (5,2), monthly dataset.....	101
Table 23	ACF and PACF of residuals of ARMA (4,5), daily dataset	103
Table 24	Normality test of VAR (6), daily dataset	113
Table 25	Comparison of in-sample forecasting accuracy	117
Table 26	In-sample fit vs. out-of- sample forecasting	118
Table 27	Forecasting out-of-sample accuracy results	119
Table 28	S-test pair-wise comparison of forecasts' performance significance.....	121
Table 29	Range of values for ANN iterations for in-sample fit.....	123
Table 30	Robustness of ANN given different network configuration.....	124
Table 31	Granger causality test, monthly dataset.....	126
Table 32	Granger causality test, VAR (2), daily dataset	127
Table 33	Granger causality test, VAR (6), daily dataset	128

FIGURES

Figure 1	Neural Network structure of MLP N^{p-q-1} (Source: Khashei & Bijari, 2010).....	19
Figure 2	Elman Network structure (Source: Gómez-Ramos & Francisco Venegas-Martínez, 2013)	19
Figure 3	Support Vector Machine structure (Source: OpenCV, 2014).....	20
Figure 4	Self-organizing map structure (Source: Kriesel, 2007).	21
Figure 5	S&P500 price in 1968-2016 period.....	46

Figure 6	S&P500 return in 1968-2016 period.....	46
Figure 7	Term spread in 1968-2016 period.....	47
Figure 8	10 year Treasury bond and 3 month Treasury note yields, 1968-2016.....	47
Figure 9	Credit spread for 1968-2016 period	48
Figure 10	S&P500 index and unemployment rate, 1968-2016	48
Figure 11	S&P 500 index price level series and market sentiment index	49
Figure 12	Changes in crude oil price during 1968-2016.....	50
Figure 13	Changes in gold price during 1968-2016.....	50
Figure 14	S&P 500 return on change in VIX.....	56
Figure 15	S&P 500 return and Google trends search on term “debt”	57
Figure 16	NAR architecture.....	68
Figure 17	NARX network structure (Source: MATLAB, 2016)	70
Figure 18	NAR training and validation graph	84
Figure 19	Correlation coefficient of neural network	84
Figure 20	Training and validation summary NAR 12 lags monthly	85
Figure 21	Correlation coefficient NAR 12 lags monthly dataset	86
Figure 22	Training and validation NAR 2 lags daily dataset.....	87
Figure 23	Correlation coefficient of NAR 2 lags daily dataset.....	87
Figure 24	Training and validation image of NARX 2 lags monthly	95
Figure 25	Explanatory power NARX 2 lags monthly	95
Figure 26	Training and validation illustration NARX 2 lags daily	96
Figure 27	Explanatory power NARX 2 daily dataset.....	97
Figure 28	Model diagnostics ARMA (5,2)- plot of residuals	102
Figure 29	Distribution of the residuals.....	102
Figure 30	The distribution of residuals from ARMA (4,5) daily dataset	104
Figure 31	Autocorrelation of the residuals.....	105
Figure 32	Error distribution	106
Figure 33	Error autocorrelation NAR 12 lags	106
Figure 34	Residuals distribution NAR 12 lags.....	107
Figure 35	Test of autocorrelations in the residuals	108
Figure 36	Residuals distribution NAR 2 lags daily dataset	108
Figure 37	Network diagnostics: Correlation between Input and Error.....	109
Figure 38	VAR stability test	110

Figure 39	Illustration of VAR stability test for daily data set	111
Figure 40	Illustration of VAR stability test for daily data set	112
Figure 41	Autocorrelation of the residuals NARX 2 lags monthly	114
Figure 42	Correlation between input and error NARX monthly.....	114
Figure 43	Autocorrelation of the residuals NARX 2 daily	115
Figure 44	Residuals distribution NARX 2 daily	115
Figure 45	Impulse response of R to Google search “S&P500 index”	128
Figure 46	Impulse response of S&P500 return to Google search of “S&P500 index”	129

1 INTRODUCTION

The topic of predicting returns, stocks, bonds, indexes or market movements – is very important for the investors, as an accurate forecast will allow to make appropriate investment decisions and realize an attractive return. Therefore, forecasting financial time-series gained its high popularity.

Traditionally, investors utilized various asset pricing models for the estimation of the return for one time period ahead, and different autoregressive models (ARIMA and Vector autoregressive models), random walks, exponential smoothing and moving averages for long-term forecasting. Recent development of technology, access to the Big Data, data mining and advances in computers, allowing for the creation and computation of complicated algorithms, opened new doors for the investor in the appearance of Artificial Intelligence (AI). Even though artificial intelligence is still in its early stage of development, it is gaining increasing popularity in the professional world of traders, investment firms, portfolio managers and bankers. The discovery of artificial neural networks as method for forecasting and latest creation of even more complex networks such as deep learning, has received a lot of deserved attention due to its arguably superior forecasting results (Nomura Research Institute, 2015).

Senior managing director of Accenture Finance and Risk Services, Steve Culp (2017), predicts that AI will be a disruptive force in the banking industry, as it will restructure the operating model and processes. IDC Research (2016) suggests further that banking will be able to adjust by becoming the biggest user of AI solutions. Furthermore, recent survey of companies' executives and industry experts infers that AI will grow with a compound annual growth rate (CAGR) of 55.1% over the 2016-2020. (IDC Research, 2016).

There is an increasing number of companies in the finance field that are implementing various types of artificial intelligence, for example, in automated portfolio management (Betterment (2017), Schwab Intelligent Portfolios (2017)), algorithm trading (Renaissance Technologies (2017); Walnut Algorithms (2017)), loan and insurance underwriting, fraud detection (algorithms scan for abnormal activity), news analysis and others.

Therefore, it becomes apparent that AI will drastically change the finance industry, as it will be able to offer solutions that are superior in performance to the traditional methods and human mental capacity. Due to the before mentioned growing trend of utilization of AI in

the finance area, this study is dedicated to research recent developments in this topic, trends, applications, as well as to examine what does AI hold for the investment finance, and could it offer a better and more accurate predictability tool that will allow investors to realize abnormal return.

1.1 Purpose of the study

The purpose of the study is to examine the phenomenon of Artificial Intelligence (AI) in finance, research its applications in the business world and determine whether the early stage of AI technology, such as artificial neural networks, is able to solve time-series prediction problem with higher accuracy than traditional forecasting methods, such as ARIMA and Vector Autoregressive models.

This thesis is particularly focuses on the data from the US market, as it is one of the most reliable and easily accessible. It also provides with large enough time frame to test different time frequencies. Specifically, it uses two datasets: firstly, monthly S&P500 index return and its explanatory variables over the period of 01/1968-09/2016 and secondly, daily S&P500 index return with its explanatory variables from 22/05/2007 – 24/05/2017.

For the empirical methods ARIMA models and Vector Autoregressive models will be used as the proxies for “traditional” and linear forecasting methods, while artificial neural networks will be used as proxies for AI. The measurements of the forecasting accuracy are statistical loss functions such as MSE (Mean Square Error) and RMSE (Root mean Square Error) for the in-sample fit, MSPE (Mean Square Prediction Error) and MAPE (Mean Absolute Prediction Error) for out-of-sample fit, and economic loss functions, such as percentage of correct sign predictions and percentage of correct direction prediction. To measure the significance of one forecast over the other one, S-test from Diebold and Mariano (1995) and Direction Accuracy test (Pesaran and Timmermann (1992)) will be used.

1.2 Study limit

Study of Artificial Intelligence is complicated due to the lack of data available, firstly, as the phenomenon is very new and, secondly, finance companies and hedge funds that utilize this strategy are often privately held and are not obliged to disclose the returns. Besides, companies and funds that are implementing AI often lack the desire to publish extensive information about the technology they use or realized return, which could be explained by their fear of a competition (high return attracts competition, while low return scares the

investors). Furthermore, even if all finance companies that employ AI would publish their results, due to the small number of those companies worldwide, the quantitative research on this topic would be limited and possibly biased.

For these reasons, this study is limited to the investigation of the performance of neural network application in forecasting of financial time-series. It is important to note that the results of this research could not be extrapolated to the artificial intelligence at wide, as each AI type differs from one another, therefore, the conclusion can only be made for the performance of one application of artificial intelligence (neural networks).

The research is limited to linear methods used, such as ARMA, VAR and non-linear such as neural networks. Thus, the comparison on the forecasting accuracy can only be made on those models/networks.

This study is limited to USA market; hence the results might be different if a similar study is conducted on different markets.

1.3 Contribution

Even though researches on Artificial intelligence (particularly on artificial neural networks) have been around for the last two decades, their results were not conclusive, therefore, calling for a further study. Furthermore, artificial neural network (ANN), given its structure and complexity, does not give the same result, as the way the information is learnt by the system might be different during each time it trains (learns). Moreover, in addition to various types of ANNs, each of them can be further fine-tuned by varying the training process choice, number of hidden layers and number of neurons specifications, all of which can create a unique system, worth investigating. For example, Niaki and Hoseinzade (2013) point out that since the ANN produces different results every time when initialized with different random values, it is therefore recommended to run the initialization multiple times in order to obtain statistically supportable conclusions. However, not so many previous research papers check for the robustness of the results, given different network configuration. Therefore, this study intends on tackling this problem by including a wide sensitivity and robustness analysis by checking for the results given different network configuration, thus offering a new evidence to the existing research. It further checks for robustness by using two separate data set with different time length and time-series frequency.

Additionally, the topic of Artificial Intelligence is often linked to the Big Data phenomenon and data mining. This is due to the fact that A.I. system can process large amount of data much more efficiently than humans. “Big data” takes its roots from human interactions in the Internet and various social media, allowing for the investigation of the behaviour of market participants and, consequently, market sentiment analysis. For instance, researchers show that certain finance-related Internet page visits and search statistics tend to correlate with market movements (see e.g. Moat et al., 2013, Preis et al., 2013). These results underline one of the biggest strengths of artificial neural networks – the ability to detect patterns in huge amount of data that fluctuate rapidly over short periods of time. While a number of recent studies have investigated the predictive power of artificial neural networks on the US stock markets (see e.g. Sheta et al., 2015, Niaki and Hoseinzade, 2013, Kumar, 2009), none of these studies considered the aspects of sentimental market analysis. Therefore, this research would provide an interesting and distinguishing perspective from the existing studies by incorporating market sentiment and Big data. University of Michigan’s Consumer Sentiment Index will be used as the proxy for market sentiment in monthly frequency data set, and Google Trends search terms for the daily data set. To the author’s knowledge, this will be the first study including these variables in the artificial neural network.

Therefore, for the reasons mentioned above, this research will provide a further insight into this new area in finance as well as enlighten its readers by providing a good analysis of the previous research, current trends, and some technicalities behind the trending topic of “Artificial Intelligence”.

1.4 Thesis organization

The thesis is structured as follows. In chapter two, relevant background information on the subject is explained and the reader is introduced to AI terminology and applications. Chapter three discusses relevant theories, and chapter four describes literature review and previous studies. Chapter five will present information on the data and descriptive statistics. Chapter six discusses research methods used, states research question, research hypothesis and statistical hypotheses. In the following chapter seven empirical study is described, followed by the model diagnostics in chapter 8. Results are summarized in chapter 9, discussed in chapter 10. This paper is concluded with chapter 11 summarizing the research, stating the implications, study limits, and suggestions for future research.

1.5 List of abbreviations

AI – Artificial Intelligence

ANN (or NN) – Artificial Neural Network

ARIMA - Autoregressive Integrated Moving Average

BP – Back Propagation (learning algorithm)

BR – Bayesian Regularization

BPNN – ANN that is trained with BP

LM - Levenberg–Marquardt (learning algorithm)

NAR - Nonlinear autoregressive network

NARX - Nonlinear autoregressive network with exogenous inputs

SCG – Scaled Conjugate Gradient (learning algorithm)

SOM – Self-organizing maps

SVM – Support Vector Machine

VAR – Vector Autoregressive

2 BACKGROUND INFORMATION

2.1 What is AI

Artificial Intelligence is a field of study that aims at replicating and enhancing human intelligence through artificial technologies to create intelligent machines (Zhongzhi, 2011; Russell & Norvig, 1995). Some researchers suggest that AI should be able to act and think rationally, while others suggest slightly different definition of its ability: to act and think like humans (Russell & Norvig, 1995).

AI is a young field of study that was researched for the last 50 years, however, it has its foundations in the sciences, ideas and techniques from the long-established fields, such as philosophy (ideas of reason, logic and mind), mathematics (which gave theories of logic, deduction and induction, probability, decision making and calculation), psychology, linguistics and computer science (Russell & Norvig, 1995).

One of the major keystones in the setting up the field for the AI development was Alan Turing's (1912-1954) introduction of Turing Machine (1937), (a model of ideal intelligent computer), and his development of the automata theory. The first commonly regarded AI study was the research of Walter Pitts and McCulloch (1943), that had developed the MP neuron, which would set up the beginning of Artificial Neural Network research. (Zhongzhi, 2011; Russell & Norvig, 1995). Since then the researchers were interested in the creation of a "thinking machine" that could imitate the process of human brain. In response, Alan Turing proposed *Alan Turing test* (1950) to determine the level of intelligence exhibited by machines (AI). If the machine is able to demonstrate human-level performance in all cognitive tasks so that it tricks the investigator to believe that the communication was held by a person, not a machine, then it passes the test, meaning it has a thinking and information processing capacity of a human (through an ability to store information provided before and during the interview, use that information to respond to the questions and make new conclusion, as well as being able to adapt to new context, identify and extrapolate patterns). (Russell & Norvig, 1995).

Research of AI has experienced three distinctive waves. First one as described above was facing limitations of computational capacity and processing power of computers and could not progress much further. Second wave appeared in 1980s, with the further development of artificial neural networks that were imitating human brain functions, machine learning

systems and enhanced computational capacity of the computers. (Zhongzhi, 2011). The world is currently experiencing a third wave of AI research, driven by the development of the **deep learning**, which allows for the creation of much more complex neural system and a more efficient application in the real world. The distinctive difference between it and the previous types of AI is the lack of human intervention in the system. While the programmers were required to prepare training data, train, formulate analytical logic and evaluate the system's results and accuracy for the ANN, they are not needed for training of deep learning system, as it trains itself. (Nomura Research Institute, 2015).

Most common types of Artificial Intelligence are neural networks (including deep learning) and machine learning. Artificial Intelligence has progressed dramatically and nowadays most popular areas of its application are speech recognition, natural language processing, image recognition, sentiment analysis, autonomous driving, and robotics. (Pannu, 2015)

2.2 Applications of AI in finance

In the last few years Artificial Intelligence was able to make many advances that enabled to create applications for finance professional that could (or, arguably, will) disrupt the finance industry. It is therefore hypothesized that AI not only might be able to replace (fully or partly) human capital, but also to improve the performance beyond the human benchmark. Some of its applications and companies that already implement AI will be discussed below.

2.2.1 Anomaly detection

Firstly, AI is utilized in **anomaly detection**. Pattern recognition helps to identify behavior that deviates from standard patterns. For example, AI can be used in identifying money laundering, security threats, illegal financial schemes and illicit transactions and issuing an alert. (Nomura Research Institute, 2015). One of the companies that already uses this technology is MasterCard (2016) that in the end of 2016 introduced Decision Intelligence service which uses artificial intelligence for fraud detection, particularly it uses algorithms for identifying normal and abnormal shopping patterns for the clients, location of the purchases, time, typical price range and other factors which helps to quickly identify the abnormal behaviour and block further uses of the account until further clarification.

2.2.2 Portfolio management and robo-advisory

Secondly, it is used in establishing **optimal investment strategies**. Nowadays there has been an increasing number of robo-advisory services that make automated recommendation

for **portfolio management** to individual investors. (Faggella, 2016; Nomura Research Institute, 2015). Automated portfolio management companies can charge lower fees, while providing at least equally good results (if not better). For example, **Betterment** (2017) and **Schwab Intelligent Portfolios** (2017) are portfolio management companies that base their model and analysis with AI to provide automated optimal portfolio choice and automated rebalancing, asset portfolio selection and equity selections.

2.2.3 Algorithmic trading

Third application of AI in finance is algorithmic trading, which is the systems that uses proprietary algorithms to incorporate knowledge about changing market conditions and price level and make automated very fast trades. Often, the trades are made so quickly that gained the term “high-frequency trading”. For obvious reasons the human cannot process information in such short period of time, therefore losing to the algorithmic system.

Some companies that have already implemented AI and perform successful algorithm trading are **Renaissance Technologies** (2017) and **Walnut Algorithms** (2017). Walnut Algorithms, the finalist in the Innovation in Investment management Fintech Award 2016, use AI in the form of sophisticated trading models used for identifying patterns in financial markets, self-adapt to changing market condition and implement trading.

Other examples of the companies that use this type of AI application are two competing San-Francisco based firms Cerebellum Capital and Sentient Investment Management. **Cerebellum Capital** (2017) is a hedge fund management firm that implements artificial intelligence solutions for investment decisions. It uses machine learning, which means that it is a self-learning algorithmic system that makes its own predictive models on where the markets will go, tests them, corrects and updates its algorithms. People behind the company have 30 years of research of statistical machine learning and time-series prediction fields. **Sentient Investment Management** (2017) is an investment company that just like Cerebellum Capital uses artificial intelligence in the form of machine learning to develop investment tactics and proprietary quantitative trading. Additionally, it also uses newest form of AI – deep learning and image recognition - for scanning through all the information, trying to find relevant for the investment decision.

2.2.4 Text mining

Fourth utilization of AI is in text mining, news and semantic (syntax) analysis. AI is used to automatically “read” and analyse text (reports, news, social media activity and content). (Nomura Research Institute, 2015). This will be very important in the future development of investment services, when AI machine will read all the relevant information and news in just a few seconds, while humans would need plenty of hours for this task and still would not be able to cover ALL information that could be affecting the particular stock performance. Data mining method helps with analysing market data, forecast activity and price level. It can also incorporate predictions of regulatory and institutional changes and simulate its results. (Nomura Research Institute, 2015).

One example of this technology implemented in practice is **AlphaSense** company. AlphaSense (2017) is a search engine for finance professionals, which implements linguistic search algorithms to help a person to find desired information in far less time than in other search engines. It utilizes natural language processing to find most relevant information, and their algorithms are self-learning, i.e. they learn from their mistakes and make next search more efficient. The company was founded by Jack Kokko, who previously worked as an analyst at Morgan Stanley after graduating from the Helsinki School of Economics.

Another example is **Dataminr**. Dataminr, founded in 2009, uses algorithm to scan through Twitter and other posts on public social media platforms to pick up on the news and send actionable alerts to the subscribed investors before the news are reported. (Dataminr, 2017).

Furthermore, company called **Kensho** (2017) is an example of data mining technology being implemented in the advisory services. It uses natural language processing (NLP) systems that can read posted questions and scan through appropriate information to make quick recommendations or answers. The algorithm can identify association between certain events and stock prices and recommend investment decisions that people might not have observed or contemplated. The clients could be individual investors or investment firms themselves. Kensho is suitable for a younger generation that does not require a human advisor, as people are perfectly fine corresponding with an automated system. The companies that would be most hurt by this technology are, for example, current financial advisors at big banks, that require \$350k salary and 40hours of work for something that Kensho can do in the matter of minutes, without entailing salary, bonuses, sick days, or maternity/paternity leaves.

2.2.5 Market, sentiment or news analysis

Use of AI application in market sentiment analysis is steaming from the previous application (text mining) and is often considered as its subset. Market sentiment is gaining more popularity recently with the wide development of social media platforms and creation of big pull of data. This increasing mass of “big data” resulting from human interactions in the Internet and various social media, offer intriguing new ways of investigating the behaviour of market participants. For instance, Ceron et al.(2015) find that aggregated sentiment analysis of social media content improves forecasting accuracy of upcoming results in several elections. Similarly, some other studies connected the number of online searches for a specific topic with a fore-coming economic activity (Wikipedia and Google Trends searches with a consequent Dow Jones Index price change). (Moat et al., 2013, Preis et al., 2013).

The support from the practical field comes in the form of a numerous of finance companies that try to implement exactly that. For example, **iSentium** (2017) uses own algorithms to scan through millions of messages and posts on social media platforms in order to understand **consumer sentiment** and make a prediction about future economic activity. iSentium is often used by investment banks and hedge funds.

2.2.6 Credit evaluation and loan/insurance underwriting

Further successful application of AI could be in **credit evaluation** (credit risk analysis, credit rating and scoring, bond rating etc.). Bahrammirzaee (2010) finds several studies that conclude that corporate finance will be enhanced with the use of artificial neural networks, as ANN has better accuracy of credit evaluation and bankruptcy prediction.

ZestFinance, founded by former CIO of Google, Douglas Merrill, created machine-learning system that allows for a smarter and more efficient connections between borrower and lender (2017). Traditional lending schemes have not changed in 50 years, and they still use only few data points (less than 50) and make (often biased) decisions. ZestFinance’s creation of **ZAML** (Zest Automated Machine Learning) is destined to find millions of new borrowers by using wide range of data points from Big data and also to keep bias out of the credit analysis. Based on years of their research they argue that there is not any single deal-breaking point in the credit analysis, thus ZAML by using thousands of data points will help to more accurately identify good borrowers. Additionally, it also helps to eliminate bias. As the founder points out, when a person runs late on his payment and calls for more time, he is not always a bad borrower and they were surprised to know that other factors were behind

the probability of a person defaulting. This is especially relevant for the young people with little or no previous borrowing history, that traditional underwriters usually try to avoid. Furthermore, in order to break the “black box” of using machine learning, the founders also offer ZAML explainability tools that will in a simple way describe how the result was derived and provide legally required information for the applicants in the adverse action.

2.3 Types of Artificial Intelligence (AI)

Artificial intelligence comprises of data-driven methodologies, such as artificial neural networks, machine learning, genetic algorithms, probabilistic belief networks and fuzzy logic (Binner et al., 2004). This paper will mainly focus on artificial neural networks.

2.4 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a collection of interconnected simple parallel processing components (nodes or units), imitating the structure and functionality of a human neuron. These imitated neurons are the processing elements of the artificial neural network. The processing capacity of the network is stored in the linking weights, which are obtained by a process of learning from a set of training patterns (Oztekin et al., 2016). Neural networks differ greatly in their structure, and researches identify the following main types of ANNs: 1) Feedforward Networks (Multilayer Perceptron (MLP) as the most common type), 2) Recurrent Networks, 3) Support Vector Machine, 4) Modular Networks and 5) Polynomial Networks. (Gómez-Ramos & Francisco Venegas-Martínez, 2013).

2.4.1 Types of ANN

Perhaps the most common group is **Feedforward Networks**. Their main characteristic is that they only establish connections with neurons in the forward positions and do not communicate with the neurons on the same or previous layers. Some of the networks with such specifications are Radial Basic Function, Probabilistic Neural Network and Multi-Layer Perceptron (MLP). Multi-Layer Perceptrons (MLP) are the most prevalent neural network models (Oztekin et al., 2016; Hornik, Stinchcombe, & White, 1990), as they are able of learning complex non-linear functions with significant accuracy rates. MLP is a feed-forward system, with supervised learning that needs a desired output for the learning process. It consists of an input layer (where number of neurons equal to the number of variables), a number of hidden layers, and an output layer (with the response to the problem, having as many neurons as the number of quantities computed from the inputs) (Oztekin et al., 2016).

The example of MLP architecture is presented in Figure 1.

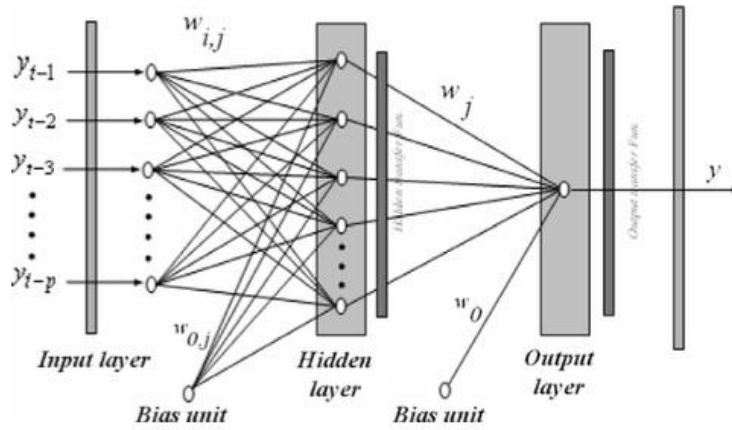


Figure 1 Neural Network structure of MLP N^{p-q-1} (Source: Khashei & Bijari, 2010)

Second group of neural networks is called **Recurrent Networks**. The difference between Feed-forward networks and Recurrent Networks is that a latter has a cycled connectivity between its nodes, as they communicate to the neurons in all layers of the network and store the information. (Gómez-Ramos & Francisco Venegas-Martínez, 2013). These networks are characterized by the dynamism of their connectivity and are often suggested for the time-series problems (MATLAB, 2017). Most distinguished representatives of this group are Elman Networks (ELN) (depicted on Figure 2) and Autoregressive Networks (ARN).

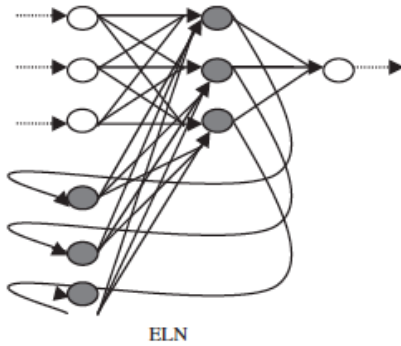


Figure 2 Elman Network structure (Source: Gómez-Ramos & Francisco Venegas-Martínez, 2013)

Third group of neural networks is **Support Vector Machine** type. It works as a so-called “hyperplane” that act as a decision surface and aims at maximizing the margin of separation (margin of training data). It utilises supervised learning. It often works best in a classification problem. The example of SVM is shown in Figure 3.

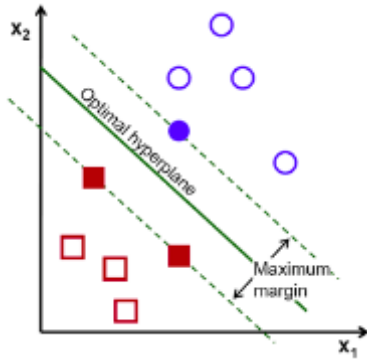


Figure 3 Support Vector Machine structure (Source: OpenCV, 2014)

Another type of the neural network is a **Modular Networks** type. Modular Networks combine many other networks (modules) in one, allowing it to be solving tasks separately and then linking the answers together.

Additionally, Gómez-Ramos and Francisco Venegas-Martínez, (2013) discuss another group of neural networks -**Polynomial Networks**, that possess a more efficient processing qualities of polynomial input variables. It could be used in both classification and regression problems. Examples of such networks are Function Link Networks and Pi-sigma networks.

Furthermore, **Deep learning** is another neural network, that was recently developed by Hinton and Salakhutdinov in 2006 and received much of attention. It contains several hidden layers and utilizes unsupervised learning, which allows for a much more sophisticated network and better results. The creation of deep learning has sparked a third wave of research into neural networks because it provided world-record results in many benchmarked classification and regression problems.

Another neural network that used unsupervised learning is a **Self-Organizing Map (SOM)**, where all neurons are interconnected with their neighboring neurons. It does not have any hidden layers, and resembles a decision-making surface (illustrated in Figure 4). It was developed by a Finnish researcher Teuvo Kohonen. (Kriesel, 2007).

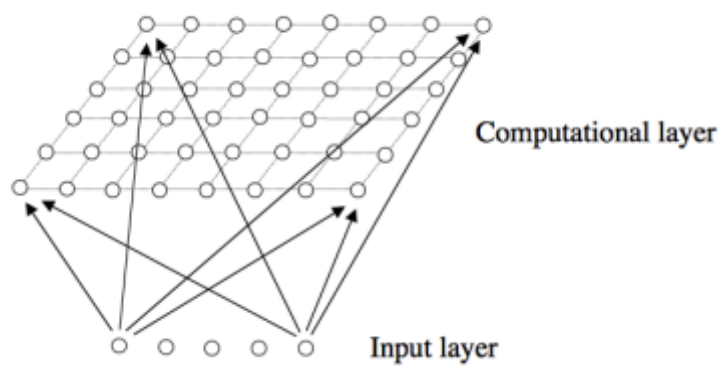


Figure 4 Self-organizing map structure (Source: Kriesel, 2007).

3 THEORY REVIEW

3.1 Traditional Investment Theories

Traditional finance is based on multiple theories and principles, where the investor is assumed to behave rationally. One of the most important theories that claim to explain the investment decision-making is **Efficient Market Hypothesis (EMH)**, developed by Fama (1970). The definition of the efficient market as the one where the price of the asset reflects all available information implied that successive changes in the price are independent; hence accurate forecasting of future values in order to realize a high return on investment is impossible. Thus, forecasting using random walk (the best estimate for the value at time $t+1$ is equal to the value at time t) received a theoretical foundation.

However, Efficient Market Hypothesis received a lot of criticism. Grossman (1976) and Grossman & Stiglitz (1980) argued that perfectly efficient markets, reflecting all available information, are an impossibility. They base it on the logic that if markets are perfectly efficient, there is no financial gain in collecting information, giving no purpose to trading and causing markets to collapse. Therefore, markets need a certain degree of inefficiency to compensate the investors for collecting the information. (Grossman, 1976; Grossman & Stiglitz, 1980). Furthermore, many studies have re-examined random walk hypothesis and found evidence against it (Lo, 2004), concluding the dependence of stock returns on its past lags, leading to the rejection of a random walk hypothesis. This implies that future returns could be predicted using its historical values.

Another theory that is based on the assumption of a rational investor is **expected utility theory (EUT)** (Von Neumann & Morgenstern, 1944) that proposed that investors can rationally evaluate all the alternatives and the associated risk on the basis of their utility and make an investment decision. However, Kahneman and Tversky (1979) found that this assumption is not valid in the empirical test, and EMH and EUT do not hold through. Their research has ignited the study of behavioral finance as a new concept that combines psychology and investors biases in the economic and financial decision-making.

3.2 Behavioral finance

Kumar and Goyal (2014) in their extensive literature review of behavioral biases found the following four most common **behavioral biases** that affect the investment decision-making process: herding, overconfidence, home bias/familiarity bias and disposition effect.

Firstly, one of the most common biases in the investment world is known as **herding**, wherein investors irrationally follow the decision of the other investors without fully understanding the reasons for it. There are various reasons for it, some arise from the study of the group psychology, that finds that in a group decision-making, the minority feels pressured and is more willing to change its mind (even if individual's own calculations differ from others, he/she will start to think that if so many other people agree on this answer, it must be right) (Zarnoth & Snizek, 1997). Other reasons could be steaming from a poorly structured performance measurement of fund managers, when the fund manager's compensation is tied to the benchmark of other funds' performances (if he underperforms together with others, he is not penalized, however if he pursues individualistic decision and happens to be wrong, his compensation would be negatively affected) (Cheng et al. 2015).

Second widely recognized bias is **overconfidence** – appears when investor is overconfident about his/her ability to make a profitable investment. Given his past success, the investor projects the success to continue, believing in his exceptional ability to forecast the stock performance better than others, which results in him often neglecting the risks and making irrational investments. (Odean, 1999).

Another commonly exhibited behavioral bias is **home bias/familiarity bias** which emerges when investors tend to hold excessively more domestic securities than the rational investor would. It is often refereed to equity home bias puzzle because the returns on international diversification of the portfolio are higher than the returns on domestic portfolio. This bias is studied in a great detail by Coval and Moskowitz (1999).

Lastly, a **disposition effect** has been considered as a common investor bias. It materializes when an investor is selling winning stocks and is holding onto the loss-making assets (could be due to tax motivations) (Odean 1998).

Behavioral biases are only attributed to the human behavior, as Artificial Intelligence is free from irrational decision-making. Moreover, the superiority of AI (being algorithm and pattern recognition based) comes in its capacity to identify a pattern of irrational decision-making made by human investors and to make the bets against it, realizing an abnormal return.

3.3 Adaptive Market Hypothesis (AMH)

Andrew Lo (2004) proposed a concept that reconciles market efficiency with behavioural finance by using biological arguments, such as relating the characteristics of evolution (natural selection, adaptation and competition) to financial transactions. He suggests that behavioural biases (overconfidence, herding, loss aversion and overreaction) should not be regarded as irrational as they stem from the heuristics of an evolutionary context. Instead of regarding them as counterexamples to economic rationality, one could better refer to them as maladaptive.

He discusses market efficiency on the example where he uses market participants as species, competing for the resources (for example, assets on the stock market). He suggests that when there is a lot of competition for the resource (either there are lots of market participants desiring it, or the resource is scarce), the market becomes highly efficient (both on pricing and information availability). However, when there is a small number of species competing for an abundant resource (for example, oil paintings), the market becomes less efficient. Therefore, market efficiency is context dependent. Andrew Lo (2004) further discusses that financial markets (similarly to ecological forces) have cycles, meaning that the risk preference of the individuals tends to be time-varying and dependant on its historical values. This, in consequence, leads to time-varying risk premium (which is shaped by the forces of natural selection, such as number of participants and number of profit opportunities available). Thus, AMH, in comparison to EMH, provides more complex market dynamics, with cycles, trends, crashes, panics and bubbles, - all of which exist in the biological ecosystem. AMH also allows for the time-varying arbitrage opportunities to exist, which, coupled with market dynamics, give motivation for active management, that can quickly adapt to changing market conditions. This theory gives an additional reason for studying forecasting of stock market return, its historical values, cyclicity and trends.

3.4 Chaos Theory

Another theory that is involved in many financial time-series problems is the Chaos theory. The theory that is based on the notion of non-linear dynamics suggests that small changes in the beginning of the time series might have a significant effect later on, hence what at first might have seem random, could have complex underlying patterns. Hence, it gave researchers the inspiration to take this theory to the finance field to analyse and explain stock market prices and returns with the function of non-linear dynamics of investors behaviour (Olsen,

1998). This theory is especially appropriate for this thesis that uses neural networks as non-linear methods, with the purpose of finding non-linear dynamics and improving a forecast performance in a time series question.

3.5 Time-series momentum theory.

Furthermore, the research by Moskowitz et al. (2012) provides additional evidence and arguments for forecasting stock returns. They research time-series momentum (which is different to cross-sectional momentum researched by Fama) and find that there is a persistence of the returns among many different assets classes for the 12 months after which, the reversed returns are expected. This suggests that the investors can realize substantial abnormal returns by predicting future prices, given current and historical prices. This undermines Random Walk Hypothesis that says that future prices are unrelated with historical returns. This brings an important implication for the field of forecasting as predicting future stock prices and stock returns now receives a theoretical background. This prediction, if executed with sufficient accuracy, can result in abnormal returns for the investor. Furthermore, for the field of artificial intelligence it implies that having an algorithm that identifies the cycles of each company (if the company going through a profitable or unprofitable stage and in which profitability stage is it heading) and makes automated trades, will likely yield a high return.

3.6 Forecasting methods

Accurately predicting stock market return is quintessential for the investor. Conventionally, previous studies used linear models for forecasting time-series, such as ARIMA, exponential smoothing and moving average. ARMA approach is very popular given its beneficial statistical properties and the well-known Box–Jenkins methodology. (Zhang, 2003; Box & Jenkins, 1970). However, there are many other forecasting methods available for the time series prediction. In addition to the “classic” models that have been around for decades such as moving average and exponential smoothing, there are also relatively new models that steam from the development of neural networks.

Andrawis et al. (2011) considered the following forecasting methods as the most popular and tested them and their variations in their research: ARMA (with a variety of order specifications), multiple regression, AR (with the order specified using the BIC), simple moving average, Holt’s exponential smoothing (with parameters obtained by maximum

likelihood method), Gaussian process regression, multilayer neural network and Echo state network. In their research, they find that the combination of different forecasts provides better accuracy (Andrawis et al., 2011). Some other methods include multivariate forecasting methods (VAR), discriminant analysis, as well as non-linear models, such as regime-switching models and threshold models (smooth transition autoregressive (STAR)) suggested by Teräsvirta et al., (2005) (Gooijer & Hyndman, 2006).

The choice of the forecasting model depends on many factors, such as data characteristics (noise, distribution), length of the forecast, type of the financial series and theoretical support (from the previous research) for the choice of the model.

3.6.1 Linear vs. nonlinear methods

The researchers distinguish two main groups of forecasting methods: *linear* and *non-linear* (Clements et al., 2004). Though, there is no clear evidence that linear model is better fitted for estimating the return, as the assumption of linearity may not be accurate for solving complex real-world problems (Zhang, 2003; Hwarng, 2001). Clements et al. (2004) argue that a lot of financial series follow non-linear patterns, for example, there could be few regimes (expansion, and recession), or periods with different financial variables (periods of low and high volatility), or situations when certain variables are applicable only in certain conditions (for instance, only when the oil price reaches certain high ceiling, it then affects the GDP level).

The number of non-linear financial series observed in economics and finance field sparked the research and application of non-linear forecasting methods, such as regime switching models, neural networks and genetic algorithms, which are more sophisticated forecasting methods with arguably better accuracy. (Clements et al., 2004). They state that there is no clear solution to the question of which model to use and in which types of problems to use it, as the answer varies from one dataset to another. Overall, they suggest using non-linear methods, if one suspects the nonlinearity to occur.

3.6.2 Benefits of ANN

Out of various forecasting methods many researchers found that **ANN is often superior in forecasting** due to some of its favorable features: firstly, its **ability to capture non-linear relationship**; secondly, its numeric nature allows for the **direct data input** (without possible incorrect transformation and inappropriate data intervals), thirdly, **no**

assumption of the data distribution is needed; fourth, some ANNs **allow new information to be entered** in the network and provide new results without reprocessing the old ones (whereas in the regressions new and old information would be combined in a batch and produce one result), lastly, it **does not require a model estimator**. The system allows for the data interaction without the model specification from the user, while using hidden layers allows for more complex communication. Additionally, **ANN can also combine the capabilities of different systems**. (Bahrammirzaee, 2010; Zhang, 2003; Hwarng, 2001; Khashei & Bijari, 2010). This suggests that **ANN could be better suited for some problem that has no theoretical background offered on the underlying data generating process**. (Zhang, 2003).

3.6.3 ANN drawbacks

However, artificial neural networks methods received some criticism. Firstly, Brooks (2008) criticizes the use of ANN as no diagnostics or specification tests available to check the adequacy of the model. There is also a lack of theoretical foundations for determining the optimal network parameters, including learning rate, momentum, dimensions of input vectors and number of hidden nodes/layers. This drawback is going to be accounted for in this study by introducing a sensitivity analysis for robustness of the forecasting results given different network configurations. Secondly, Brooks (2008) argues that coefficient estimates have no real theoretical interpretation. However, the purpose of this study is ability to deliver an accurate forecast, and not correct interpretations of the coefficients of the variables. Thirdly, he states that model can produce excellent fit-in forecast, but poor “out-of-sample” forecasts. (This appears when ANN models follow the sample data very closely, hence could not be adaptive to a different sample (newer data)). Nevertheless, some solutions to this problem have been suggested. For example, de Matos Neto et al (2017) suggested using *perturbative approach* that provided accurate fit out-of-sample. Additionally, another solution to this could be achieved by performing *pruning*, i.e. removing some part of the network, which should allow for a better accuracy for out-of-sample forecasts.

3.6.4 Forecast Combinations

The optimal forecast is the one that minimizes the loss function (often measured by mean squared error (MSE))(Elliott & Timmermann, 2016). However, Elliott and Timmermann (2016) also find that there is no single forecasting approach that uniformly dominates all

other alternatives, thus giving a strong motivation for forecast combinations. Empirical research concludes that forecast combinations provide superior forecasting accuracy than individual forecasting methods (Timmermann, 2006).

There are several methods for forecast combinations. One of them is **Bayesian model averaging** perspective where forecasts are averaged according to the likelihoods of the underlying models (Andrawis et al., 2011). Other popular methods for the forecast combination are **linear forecasting schemes** such as $f^c = \omega_0 + \omega_1 f_1 + \omega_2 f_2 + \dots + \omega_n f_n$ (where $\omega_1, \omega_2 \dots \omega_n$ are respective weights; while $f_1, f_2 \dots f_n$ are different forecasts). The weights are estimated by OLS or by weighting systems (for example, the inverse of the MSE of the predictions relative to the average MSE). (Elliott & Timmermann, 2016). However, many researchers found that very often **simple combination schemes (equal weight combination)** outperform the sophisticated forecast combinations methods. (Timmermann, 2006)

4 LITERATURE REVIEW

This chapter portrays previous studies that are applicable for this research. It will firstly present the description of Bahrammirzaee's extensive review of studies on ANN, followed by the summaries of the additional individual papers on relevant to this study findings.

4.1 Bahrammirzaee's (2010) review of previous studies of ANN applications

There is a wide range of the papers that test ANN in time-series prediction problem. Bahrammirzaee (2010) compiled an extensive study consisting of many previous researches that were made on ANN. He finds that *credit evaluation* is a very common application area of ANN, and that neural networks often outperformed the forecasting of traditional methods (multiple discriminant analysis, logistic regression) in this problem type. Secondly, ANN seemed to surpass its traditional counterparts (multiple discriminant analysis, logistic regression and random walk model) in the *portfolio management* area. Third, ANN is more efficient in *bankruptcy prediction* (compared to logistic regression, linear and multiple discriminant analysis). (Bahrammirzaee, 2010). Lastly, Bahrammirzaee (2010) finds that there are certain previous researches that have found ANN to be superior in forecasting stock returns. The most relevant studies that are cited in his paper are summarized in Table 1.

Table 1 Summary of previous research on ANN forecasting returns

Authors	Research area	Methods	Results
Liao & Wang (2010)	Forecasting of global stock indexes	ANN (with BP stochastic time effect) compared with numerical experiment on the data of SAI, SBI, HSI, DJI, IXIC and SP500, and the validity of the volatility parameters of the Brownian motion	ANN performs better
Faria et al. (2009)	Forecasting Brazilian stock market	ANN (BP) with adaptive exponential method	ANN performs better
Chen et al. (2003)	Forecasting and trading the Taiwan Stock Index	ANN (PNN) with B&H strategy, random walk model and the parametric GMM models	ANN performs better
O'Connor & Madden (2006)	Forecasting stock exchange movements	ANN (BPNN) with four simple benchmark functions	ANN performs better

(Source: Bahrammirzaee (2010))

4.2 Individual papers review

The following section will describe most relevant previous researches, their data, methods and results. The summary of the papers that this section describes is shown in Table 2.

Table 2 Summary of previous literature individual papers

Authors	Research area	Methods	Results
Niaki and Hoseinzade (2013)	Forecasting S&P500 index	ANN and design of experiments comparing with logit	ANN performs better
Sheta et al. (2015)	Forecasting S&P500 index	ANN, SVM and regression	SVM performs better
Kumar (2009)	Forecasting S&P 500 and Hang Seng Index	ANN (SCG) compared with ARIMA	ANN performs better
Oztekin et al. (2016)	Forecasting daily returns of BIST 100 Index	ANN (BPNN) compared to fuzzy inference system and SVM	SVM performs better
Olson & Mossman (2003)	Forecast of Canadian stock returns	ANN (BPNN) compared to OLS and logit	ANN performs better
Maia & Carvalho (2011)	Forecasting stock prices of 15 different companies	ANN compared with Holt's exponential smoothing and hybrid	Hybrid outperforms ANN performs better than exponential smoothing
Zhang (2003)	Forecasting of British pound/ US dollar exchange rate	ANN compared to ARIMA, random walk and a hybrid	Hybrid outperforms ANN performs better than ARIMA
Khashei & Bijari (2010)	Forecasting of British pound/ US	ANN compared to ARIMA, random walk and a hybrid	Hybrid outperforms ANN performs better than ARIMA
Teräsvirta et al.(2005)	Forecasting macroeconomic time series (e.g. CPI, interest rate and unemployment rate)	Linear AR model, LSTAR (logistic, smooth, transition autoregressive), and ANN (trained with BR)	LSTAR and ANN (BR) outperformed linear models

Niaki and Hoseinzade (2013) “Forecasting S&P 500 index using artificial neural networks and design of experiments”

First paper that will be discussed is Niaki’s and Hoseinzade’s (2013) research on stock market prediction. Their main research task was to perform daily forecast of S&P 500 index direction. They used feed-forward artificial neural network and logit model in their methodology.

The dataset in their research consisted of daily returns of S&P500 index as a dependent variable (output) and 27 other financial and economic indicators as independent variables that previously have been known to help in the forecast of stock returns (such as lags (3) of the series, oil price change, gold price change, change in the market yield of US Treasuries of 3 month, 6 month 1 year, 5 year 10 year maturities, change in currency rates with other countries, change in market yield of Moody’s AAA and BBB corporate bonds and last day returns of the big US companies and main other indexes). Using the design of experiments (DOE) the researchers identified variables that had a stronger influence on the output. They found exchange rates to have a strong association with returns, while the lags of S&P500 (lags of dependent variable) were only confusing the network. The time frame was from 1994 to 2008, which compiled to 3650 days in total out of which 80% was used for training purpose of the artificial network, 10% for verification and additional 10% for test performance. ANN was trained until the early stopping rule terminated the training process. They had feedforward ANN with one hidden layer and Back Propagation training algorithm, and also performed DOE on the other network design features, such as number of hidden nodes (which varied from 5 to 60).

Niaki and Hoseinzade analysed a financial performance of ANN networks by simulating an investment strategy that consisted of the following: total capital of \$100,000 all of which is used to buy an index when the forecast shows there will be an up movement, and sells all when it indicates a down movement (sells made at the end of the day with its closing price). Transaction costs are \$8 per trade and are deducted from the separate account at the end of the period. They compared this investment strategy (mean of 100 final capitals) with buy-and-hold strategy (under which all capital is used to buy the index on the first day and liquidate it on the last). They found that ANN significantly increased the final capital in comparison with buy and hold strategy, with the result being significant at 5% significance level.

Additionally, the authors compared the forecast of ANN with logit model and found that ANN significantly outperformed the logit model, hence suggesting that the relationship between input and output has non-linear features that were not captured by the linear model (logit), thus justifying the use of non-linear models such as ANN.

Sheta et al. (2015) “A Comparison Between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index”.

Sheta et al. (2015) have predicted S&P500 index using artificial neural networks, support vector machine (SVM) and multiple linear regression. They used the same 27 variables that have been used in research of Niaki & Hoseinzade (2013) as independent variables (input) and S&P 500 daily index over the December 2009- September 2014 period amounting to 1192 days of data. ANN configuration was multilayer perceptron (MLP) network with one hidden layer that was trained with back propagation learning algorithm. In their research, they found that it is optimal to have 20 neurons in the hidden layer.

They have used MAE, RMSE, Relative absolute error and Root relative squared error as the measure of the performance and found that SVM had outperformed multiple linear regression and multilayer perceptron.

Kumar (2009) “Nonlinear Prediction of the Standard & Poor’s 500 and the Hang Seng Index under a Dynamic Increasing Sample”

Kumar (2009) had an objective to forecast the next day return of S&P 500 index and HIS index in Hong Kong using artificial neural network and comparing the performance to ARIMA model. The timeframe was from November 1928 to December 2008, amounting to 20,132 daily observations. The dependent variable was the first difference of the natural logarithm of the daily index price, and independent variables were previous days’ returns.

Network configuration was MLP with one hidden layer with sigmoid activation function, trained with Scaled Conjugate Gradient (SCG). He found that best neural network architecture had 3 lags of S&P500 return as an input and 4 hidden neurons in one hidden layer. Kumar used ARIMA (2,1,2) and ARIMA (1,1,1) models as they suggested from AIC and SBIC information criteria.

In order to evaluate the value of predictability the author used trading simulation (\$100 that is used to buy stock index funds when the forecast predicted an increase in price or sell if

there is a decline predicted) and Hit ratio (HR) that measures the proportion of the time that the forecast predicated the right direction of the index over the whole time period (which for the useful models should be higher than 50%).

He finds that HR ratio for the ANN model was higher than 51% in 8 out of 12 out-of-sample periods in 2006, 9 out of 12 in 2007 and 7 out of 12 in 2008, whereas for ARIMA the results were moderately worse: ratio over 51% was achieved in 6 out of 12 months in 2006, 5 in 2007 and 6 in 2008. Trading experiment showed similar results: monthly gain from ANN prediction was positive in 9 out of 12 months in 2006, 7 in 2007 and 2008, whereas for ARIMA gains were achieved in 6 out of 12 months in 2006, 3 in 2007 and 6 in 2008. This concludes that there was some non-linear noise in the data that ANN was able to capture, hence outperforming the forecasting ability of ARIMA.

Zhang (2003) “Time series forecasting using a hybrid ARIMA and neural network model”

In his article Zhang (2003) compares the performance of ARIMA, ANN and a hybrid model combining the properties of ARIMA and ANN in forecasting 1 step-ahead. Zhang argues that real world problems are rarely strictly linear or non-linear, hence the model that incorporates the features of linearity and non-linearity could be most appropriate. For ARIMA method, Zhang follows the Box Jenkins methodology and finds that random walk ARIMA model fits best for the exchange rate data-set. For the ANN, Zhang uses single hidden layer feed-forward network, as he finds it most widely used.

He defines the output (y_t) and the inputs ($y_{t-1}, y_{t-2}, \dots, y_{t-p}$) to have the following relationship:

$$y_t = \alpha_0 + \sum_{j=1}^q \alpha_j g \left(\beta_{0j} + \sum_{i=1}^p \beta_{ij} y_{t-i} \right) + \varepsilon_t$$

where α_0 is a constant, α_j ($j=0,1,2,\dots,q$) and β_{ij} ($i=0,1,2,\dots,p; j=1,2,\dots,q$) are model parameters (so-called interconnection weights), p is the number of input nodes, while q represents the number of hidden layers.

He uses logistic transfer function in the hidden layer (as it is commonly used by other researchers). It is given below:

$$g(x) = \frac{1}{1 + \exp(-x)}$$

This implies that ANN is equivalent to a non-linear autoregressive model (as it is nonlinear function of incorporating past observations of y into the future value of y_t).

He suggests that often simple ANN as this (with not so many neurons in the hidden layer) can perform better in out-of-sample forecasting, perhaps, due to the problem of over-fitting that complex ANN models often face.

Zhang discusses that there is no theoretical guidance on the sufficient number of hidden layers, neurons and lags, and suggests conducting few experiments varying those parameters in order to find the one that minimizes MSE. As for the training of the system he uses generalized reduced gradient (GRG2) nonlinear optimization algorithm.

Research of Denton (1995), cited in Zhang's article, showed that given nonlinear feature of ANN, ANNs could outperform linear forecasting methods when there are outliers or multicollinearity present in the data.

For the hybrid methodology, Zhang first uses ARIMA to analyze linear part of the problem, then he uses ANN to model the residuals from ARIMA. Given ARIMA's inability to explain non-linear function of the data, the residuals from its regression will contain non-linear information, which will be consequently modeled by ANN, that will be able to explain the non-linear part of those residuals. Hence, the hybrid model incorporates both linear and non-linear functions. However, Zhang suggests that given the models' ability to explain different patterns, it could be perhaps even more beneficial to model the ARIMA and ANN separately and then combine the forecasts, as he presumes that by doing so, it will provide higher accuracy forecast.

Zhang uses three datasets: British pound/US dollar exchange rate (733 observations: weekly for 1980-1993), Wolf's sunspot data (288 observations of yearly number of sunspots) and Canadian lynx data set (number of lynx trapped per year in Canada for the period of 1821-1934).

He finds that ANN and hybrid model surpass the accuracy of the random walk model, and hybrid model outperforms both of the methods.

Maia & Carvalho (2011) “Holt’s Exponential Smoothing and neural network models for forecasting interval-valued time series”

Maia and Carvalho (2011) analyzed the performance of Holt’s exponential smoothing forecast with the accuracy of neural network method in a 5 step-ahead forecasting stock market time series. Additionally, in accordance with Zhang (2003), they also introduced a hybrid model (of exponential smoothing and ANN). First, for the neural network method they used multilayer perceptron (MLP), with two-feed-forward layers (one hidden layer and one output layer) and trained the network by Conjugate Gradient Error Minimization. For the second approach, they used Holt’s new model of exponential smoothing where smoothing parameters are estimated using non-linear optimization methodology. Lastly, they used a hybrid of MLP and Holt, which merged liner structure and non-linearity in the residuals. However, quoting Zhang (2003), they also suggest that it is not necessary to perform forecast combination by creating a hybrid, as it is sometimes more beneficial to perform a linear model forecast and a non-linear forecast separately, and then to combine the forecasts in order to improve the accuracy of the modeling. They used **stock prices for 15 companies from different markets and industry segments** (for example, Coca-Cola (2004-2008 period), Microsoft (2003-2008), Apple Inc.(2005-2008), Google Inc.(2006-2008), IBM (1987-2008) and others). They used *interval U of Theil statistics* and *Interval average relative variance* as the measure of performance.

They find that the hybrid performs better than MLP and Holt.

Olson & Mossman (2003) “Neural network forecast of Canadian stock returns using accounting ratios”

Olson & Mossman (2003) compared the performance of ANNs to the forecast of ordinary least squared (OLS) regression and logistic regression (logit) and found that ANN was superior in forecasting accuracy than traditional forecasting techniques. They used market adjusted abnormal return as dependent variable, and 61 accounting ratios (such as current ratios, valuation ratios, debt-equity ratios, return on assets etc.) as explanatory variables over the period of 18 years (1976-1993). They found that ANN (Back propagation neural network) had better forecasting accuracy than OLS and Logit.

Oztekin et al. (2016) “Data analytic approach to forecasting daily stock returns in an emerging market”.

Oztekin et al. (2016) used ANN, fuzzy inference system (FIS) and support vector machines (SVM) to predict daily returns of the BIST 100 Index (trading on Istanbul stock market).

Oztekin et al. (2016) note that ANN specification varies from one study to another because of the changes in the problem and data type, as well as network type, search type, weight type and other variations in the algorithm. For their data type and a question (forecasting daily stock returns), after the trial and error experiments, they found that back-propagation algorithm with the MLP (feed forward) gave the most favorable results with the [6-n-m-1] architecture (first layer of input consisted of 6 neurons representing 6 variables; 2 hidden layers and 1 layer of the output). Fuzzy inference system used fuzzy logic instead of Boolean logic.

In their question of predicting stock performance, Oztekin et al. gave the output layer a log-sigmoid function, which takes a numerical value between 0 and 1. If the output is higher than 0.5, there will be an up movement, if it is lower than 0.5 – down movement.

They find that SVM performs better than other models.

Khashei & Bijari (2010) “An artificial neural network (p,d,q) model for time series forecasting”

In order to update Zhang’s (2003) results, Khashei & Bijari (2010), compared the performance of the ANN that had twelve inputs, four hidden and one output neurons (12-4-1), with ARIMA model (random walk)(using the same data sets as in Zhang (2003) and following closely his methodology). They additionally introduced their own hybrid model that had better results than that of Zhang (2003). They used MSE and MAE as the measure of forecast accuracy. Using the same three dataset as Zhang (2003), they find that their hybrid had the best forecasting performance, followed by ANN.

Hwarng (2001) “Insights into neural-network forecasting of time-series corresponding to ARMA (p, q) structures”

Hwarng used back-propagation neural network (BPNN) for forecasting one-period-ahead, and compared the performance to ARMA model of Box Jenkins methodology. Hwarng tested the performance using many time-series dataset (stock company return, exchange

rate, lynx data and sunspot data as in Zhang (2003). He used RMSE and MAPE in order to evaluate the performance and finds that BPNN performs significantly better.

He found that ANN performs better than the corresponding ARMA model.

Teräsvirta et al.(2005) “Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series. A re-examination”

Teräsvirta et al. used linear AR model, (logistic) smooth transition autoregressive ((L)STAR) model, neural network models specified with Bayesian regularization and autoregressive single hidden layer feedforward neural network (AR-NN) for predicting **monthly macroeconomic variables, such as unemployment rate, short-term interest rate, consumer price index inflation** and others, using only lags of the series as input variables. For evaluation of forecast accuracy, they used root mean-squared forecast error (RMSFE) and mean absolute forecast errors (MAFE). They found that both LSTAR and NN with Bayesian regularization outperform linear models. Additionally, they also found that forecast combinations produced higher accuracy in forecasting than individual forecasts.

In summary

- Previous researchers found ANN to be frequently superior to traditional forecasting methods in various financial applications: credit evaluation, bankruptcy prediction, portfolio management, exchange rate prediction and stock market prediction
- This suggests that, while algorithmic and data mining methods are not based on traditional finance theory and often considered as “black-box” methods, they often produce better forecasting results.

5 DATA

This chapter describes the data used in the research. This study uses two datasets: a monthly dataset of the S&P500 index returns and explanatory variables over the time period from January 1968 to September 2016, and a second dataset, consisting of daily S&P500 returns with explanatory variables, over the period of 24th May 2007- 22th May 2017. This chapter is, therefore, divided into two parts, each of them representing the respective dataset, describing it in a detail, discussing the variables of interest and rationale for their inclusion in the dataset, and providing descriptive statistics and correlation analysis.

5.1 Dataset 1 description

Time period and frequency

First data sample consists monthly data of market stock returns from January 1968 till September 2016 (inclusive), which incorporates 585 months over almost 50 years. This is a significantly long period, which allows the forecaster to learn from the moderately long historical performance the patterns of the returns and forecast them in the future. It is not advisable to use older observations (before 1926-1960s) due to the poor recording and unreliable results. Monthly frequency has been chosen because one can argue that predicting stock returns or stock market movements one or few months in advance is more desirable for an investor than predicting its returns or movements only a few days in advance.

Market

The market returns are chosen on US stock market, due to the fact that this market is the one with the most reliable and widely available historical results, as well as the fact that several previous researchers have performed studies on US market (Niaki & Hoseizade, 2013; Sheta, 2015; Kumar, 2009), so it would be interesting to compare the accuracy. Additionally, this choice is also supported by the idea that predicting US stock market is arguably very important for any reader or finance professional, as it has large influence on the rest of the world.

5.1.1 Variables

Dependent variable choice

The dependent variable is **S&P 500 index return**, measured as $\frac{P_{t+1}-P_t}{P_t}$, where P_{t+1} is the closing price at time $t+1$, and P_t is the closing price at time t . The prices were retrieved from Robert Shiller database. The choice to use the return series over S&P500 price series is given by the necessity to have stationarity in the variables in ARMA and VAR models.

Independent variables

Several variables have been found in certain previous studies to have explanatory power in stock return forecasting. The variables are listed in Table 3.

Table 3 **Potential independent variables**

N.	Variable	Description
1	S&P500return_n	Lags of the dependent variable, where n=1,2,3...n number of lags
2	Term spread	Difference between 10 year Treasury bond and 3 month T- bill yields
3	Credit Spread Aaa	Difference between 10 year Treasury bond and Moody's Aaa bond yields
4	Credit Spread Baa	Difference between 10 year Treasury bond and Moody's Baa bond yields
5	Corporate Spread	Difference between the yields on Moody's Aaa and Moody's Baa corporate bonds
6	USD_GBP	Change in the currency exchange rate of U.S. Dollar per British Pounds
7	USD_JPY	Change in the currency exchange rate of U.S. Dollar per Japanese Yen
8	USD_DEM_EUR	Change in the currency exchange rate of U.S. Dollar per German Deutschemark till 1999 and U.S Dollar per Euro from 1999 onwards
9	Real Dividend	S&P500 index's dividend paid out adjusted by Consumer Price Index
10	Real Earnings	S&P500 index's earnings adjusted by Consumer Price Index
11	Dividend yield	Difference in the S&P500 dividend yield
12	Market Sentiment	Consumer confidence index measured by University of Michigan
13	CPI	Consumer Price Index, change in prices paid by urban consumers for goods and services (natural logarithm of the series)
14	Unemployment	Percentage of unemployed people in USA (16+ years), at time $t-1$
15	Gold	Change in gold price
16	Crude oil	Change in crude oil price
17	Money Supply	Change in funds that are easily accessible for spending (M1 Money Stock in FRED)

Lags of S&P 500 return

Starting from building the simplest model, independent variables would be given by the **lags of a dependent variable**. For example, given the findings of Moskowitz et al. (2012) of persistence of stock returns for the 12-month period, it is likely that the lags of the S&P500 index returns would have an explanatory power over the current value of the index return in this research as well. However, Kourentzes et al. (2014) point out that adding several lags significantly affects the training of the neural network (due to the loss in degrees of freedom), resulting in a worse model performance. Therefore, this research will test the forecasting accuracy with different number of its lags.

However, many researches have discussed that there are other factors that explain the stock returns, therefore, encouraging to include them and building a more sophisticated forecasting model. Refer to the Table 3 for the short overview of the below described potential independent variables.

Interest rate specific variables

First independent variable for the inclusion in a more sophisticated forecasting model (in addition to the lags of the dependent variable) is **Term Spread**, which is the difference between **10-year US government bond** yield and **3-month US Treasury bill** yield. Both 10-year T-bond and 3-month T-bill were downloaded from FRED (given by Federal Reserve Bank of St. Louis (2017)). (The choice of this variable is in accordance with in accordance with Goyal & Welch (2008), Campbell & Thompson, (2008), Niaki & Hoseinzade (2013) and Sheta et al. (2015)). Some researchers have found that T-bill, by itself and with a few of its lags, has proven to predict stock return (Pesaran & Timmermann, 1995), therefore, it will be used as an independent variable in this research as well. The expectations are that positive yield curve positively correlates with positive stock returns whereas inverted yield curve (negative yield) is a sign of the fore-coming recession. These expectations are in accordance with Wang and Yang findings (2012).

Second independent variable is **Credit Spread_Aaa**, which is the difference between 10-year Treasury bond yield and Moody's Aaa bond yield. Similarly, there is also **Credit Spread_Baa** variable which measures the difference between 10-year Treasury bond yield and Moody's Baa bond yield. Given the wide acceptance in the previous literature of the credit spread measured by difference between T-bond and Aaa bonds, it will be used in the

paper as well. Both Moody's corporate bonds yields (Aaa and Baa) were downloaded from Federal Reserve Bank of St. Louis (2017). The choice of this variable is in accordance with in accordance with Goyal & Welch (2008). Fama & French (1993), in their research of risk factors of stock and bonds, have included term spread and default spread and concluded that it offers an interesting time-series variation in expected stock and bond returns. Therefore, my expectations are that credit spreads are useful in forecasting future level of economic activity and investment growth, which would, consequently, translate into future stock market returns. Thus, credit spread should have a predictive power over stock returns.

Thirdly, another possible independent variable in this category is a **Corporate Spread**. It is measured as the difference between the yields on Moody's Aaa and Moody's Baa corporate bonds. (In accordance with Goyal & Welch, (2008)). The expectations are the same as for Credit Spread.

Exchange rate specific variables

Next, following Oztekin et al. (2016), Sheta et al. (2015) and Niaki & Hoseinzade (2013), this study includes various exchange rates: exchange ratio of **U.S. Dollar per British Pounds**, **U.S. Dollar per Japanese Yen** and **U.S. Dollar per Euro**. All except for U.S. Dollar per Euro are available in Factset database for the whole period, while and **U.S. Dollar per Euro** is available from late 1999 (given the late birth of Euro currency). For the preceding periods (1968-1998) **U.S. Dollar per German Deutschemark** is used as the proxy. German Deutschemark was downloaded from Factset. (Respective symbols for the variables are USD_GBP, USD_JPY, and USD_DEM_EUR). Currency rates have been found to have a relation to stock market returns (Niaki & Hoseizade, 2013). The expectations are that the relationship between currency depreciation and stock market is positive: if US dollar depreciates (USD_GBP increases), it may be because of an economic downturn, subsequently, resulting in a decline in stock market returns.

Index specific variables

Additionally, other potential independent variables which are relevant for S&P 500 index are **Real dividend** (measured as dividend paid out, adjusted by Consumer Price Index), **Real earnings** (measured as earnings, adjusted by Consumer price Index), and **Dividend yield**. The variables are provided in Robert Shiller online database (2017) (where monthly earnings and dividend values were calculated from the S&P four-quarter sums for the

quarter starting from 1926, with linear interpolation to monthly numbers). The choice of these variables is in accordance with Goyal & Welch (2008). Furthermore, previous findings of Hodrick (1992) and Fama & French, (1988) indicate that changes in dividend yield predict substantial changes in expected stock returns. Therefore, the expectation is that dividend yield has a predictive power over the returns.

Macro-economic specific variables

Furthermore, this study includes monthly **unemployment rate**, measured as the percentage of the American population aged 16 and over that are not employed in the previous month. Monthly unemployment rate was downloaded from U.S. Bureau of Labor Statistics (2017). The lag of the variable is explained by the fact that unemployment rate of a particular month is available during a next month (delay in data publishing), thus for the prediction of the return in September, only observations up to August could be used. One could hypothesize that high rates of unemployment are during economic downturn, during which the stock returns are negative, therefore, the expectation is that the relationship between unemployment rate and stock return is negative.

Another macro-economic variable is the change in **money supply**. Money supply are funds that are easily available in US economy for spending. It was downloaded from FRED database as M1 Money stock monthly, not seasonally adjusted, time series. Previous researches have different opinions on the effect of money supply on stock returns. According to Sellin (2001), the effect will only be visible when the change in money supply is unexpected or significant. If there is a big jump in money supply, people could expect tightening monetary policy in the near future which would lead to the increase in bidding for bonds and consequent increase in current interest rates. This will lead to the increase in discount rates and the decline of present value of earnings, therefore, the decline in stock prices. Other economists (real activity theorists) argue that if the positive change in money supply is not overly dramatic, it should lead to the increase in stock prices. (Maskay, 2007). In his research Maskay found evidence supporting real activity theorists, - that positive money supply leads to the increase in stock prices. Based on these findings, the expectation is to have a positive relationship between the change in money supply and stock returns.

Further, in accordance with Oztekin et al. (2016), Pesaran and Timmermann (1995), Goyal and Welch (2008), and Campbell and Thompson (2008), this study will also include

Consumer Price Index (CPI), which is a proxy for the inflation. It is available for download from Robert Shiller online database (2017). In order to achieve stationarity in the series, natural logarithm of series was taken. Consumer Price Index (All Urban Consumers) is published monthly from the Bureau of Labor Statistics. However, due to the fact that the information is released only in the following month, only past month value can be used for prediction (e.g. if predicting the returns for April, one can only use March or earlier data). This methodology is in accordance with before-mentioned scholars. Previous researches have argued that high inflation leads to the drop in real activity and, hence, to the decline in stock prices (Fama 1981), however, in practice these findings have been inconclusive. The expectation is that there is a negative relationship between change in money supply and stock returns.

Commodities

Another independent variable is a relative change in **Gold** price. Gold price is recorded in US dollars per ounce. Gold/USD exchange ratio was downloaded from Factset database from 1974-2017, and from Kitco Metals Inc. (2017) from 1968 to 1974, as that period was not available in the Factset database. (The choice of this variable is in accordance with Oztekin et al (2016), Sheta et al. (2015) and Niaki & Hoseinzade (2013)). One can expect that in the times of the recessions (negative stock returns), the investor exhibit “flight-to-safety” behavior, by selling risky assets (stocks) (thus driving the price even lower) and investing in safer assets such as gold (thus driving the prices of the gold higher) (Baele et al. 2014). This would mean an inverse relationship between the price of gold and stock index.

Another variable that is known to have some explanatory power on stock returns is **crude oil** price change. Monthly crude oil prices (in US dollars per barrel) were downloaded from FRED database (2017) (Spot Crude Oil Price: West Texas Intermediate (WTI) series). One could expect that when oil prices increase, the costs for many businesses and consumers will consequently increase, which translates into reduced corporate earnings, and decrease of stock prices. Therefore, the expected relationship between change in oil price and stock returns is negative.

Market sentiment variables

One distinctive variable that is used in this study and will provide an interesting addition to the previous research is the measurement of market sentiment. This study uses Consumer Sentiment Index, retrieved from the University of Michigan, as the proxy for a market sentiment. The index measures consumers' confidence and can signal whether consumers feel comfortable in the current economic setting, what changes in the economic conditions do they anticipate in the future, how do they compare current financial situation to the one during the previous years. It is indexed to 1966 Q1=100 and is recorded quarterly till 1978 and monthly after that. Brown and Cliff (2005) finds that the above-mentioned index has been a good predictor of long term stock markets at time horizons of up to two or three years. The authors point out that excessive optimism tends to drive prices above intrinsic values, which results in a consecutive reverse trend of low returns as market prices reverts to their fundamental values. These results are further supported by Harmon et al. (2015), who find a dramatic increase in market mimicry before each market crash during the past 25 years. Therefore, I expect higher consumer confidence to occur in good economic environment, when typical stocks have a positive return, hence having a positive relationship.

5.1.2 Descriptive statistics

Summary of the descriptive statistics for the variables in the dataset is presented in Table 4.

S&P500 return

S&P500 return has an average of 0.6% a median of 0.8% for the whole sample, with the minimum value of -20.4% and maximum value of 12.0%. The variance is not surprising, given the large time horizon with many crashes and booms. S&P500 price, ranges from \$67 to \$2170, given its dramatic historical growth. Graphical representation of S&P500 price and return time series are shown in Figure 5 and 6.

It appears that that S&P500 price time series follow a random-walk process, while the returns depict white noise process (mean around 0, no trend and no covariances) (Brooks, 2008).

Table 4 Descriptive statistics summary

	Mean	Median	Min.	Max.	Std.Dev.	Skew.	Ex. Kurt.
S&P500 return (%)	0.6	0.8	-20.4	12.0	3.6	-0.7	3.0
S&P500 price	667.1	412.6	67.1	2170.9	597.0	0.7	-0.7
Term Spread (%)	1.7	1.8	-2.7	4.4	1.3	-0.5	-0.2
Credit_spread_Aaa (%)	1.1	1.0	-0.2	2.7	0.6	0.3	-0.5
Credit spread_Baa (%)	2.2	2.1	0.9	6.0	0.7	1.3	3.7
Corporate_spread (%)	1.1	1.0	0.6	3.4	0.4	1.8	4.2
1 month T-bill (%)	0.4	0.4	0.0	1.4	0.3	0.5	0.4
3 month_T-bill (%)	4.9	5.0	0.0	16.3	3.3	0.5	0.4
10 year T-bond (%)	6.6	6.5	1.5	15.3	2.9	0.5	0.1
Aaa bond yield (%)	7.7	7.5	3.3	15.5	2.5	0.7	0.4
Baa bond yield (%)	8.8	8.3	4.2	17.2	2.7	0.9	0.8
Gold (\$/oz)	485.5	379.2	25.2	1828.5	409.8	1.5	1.4
Gold change (%)	0.8	0.2	-28.4	50.4	6.2	1.2	9.4
Crude oil price	33.2	22.5	3.1	133.9	28.2	1.3	1.0
Crude oil price change (%)	0.8	0.0	-32.7	134.6	9.2	5.3	75.9
CPI	136.1	139.7	34.1	241.4	64.5	0.0	-1.2
Unemployment rate (%)	6.2	5.9	3.4	10.8	1.6	0.6	-0.2
Money Supply	1016.3	951.0	183.0	3316.6	739.0	1.2	1.1
Money Supply change (%)	0.5	0.6	-4.2	7.8	1.6	0.0	0.5
Market Sentiment	84.7	87.1	51.7	112.0	12.3	-0.3	-0.5
USD_GBP change	-0.1	-0.0	-12.1	13.4	2.8	-0.1	5.0
USD_JPY change	0.3	0.0	-10.6	16.8	3.2	0.6	5.4
USD_DEM_EUR change	0.7	0.6	0.2	1.6	0.4	0.5	-1.2
S&P 500 Dividend	22.9	21.3	16.1	45.5	6.5	1.7	2.6
S&P 500 Earnings	52.7	42.6	7.9	108.7	22.6	0.9	-0.3
S&P 500 Dividend yield	3.0	2.9	1.1	6.2	1.2	0.5	-0.6
Change in dividend yield	0.0	-0.3	-12.9	24.9	3.8	1.1	7.6

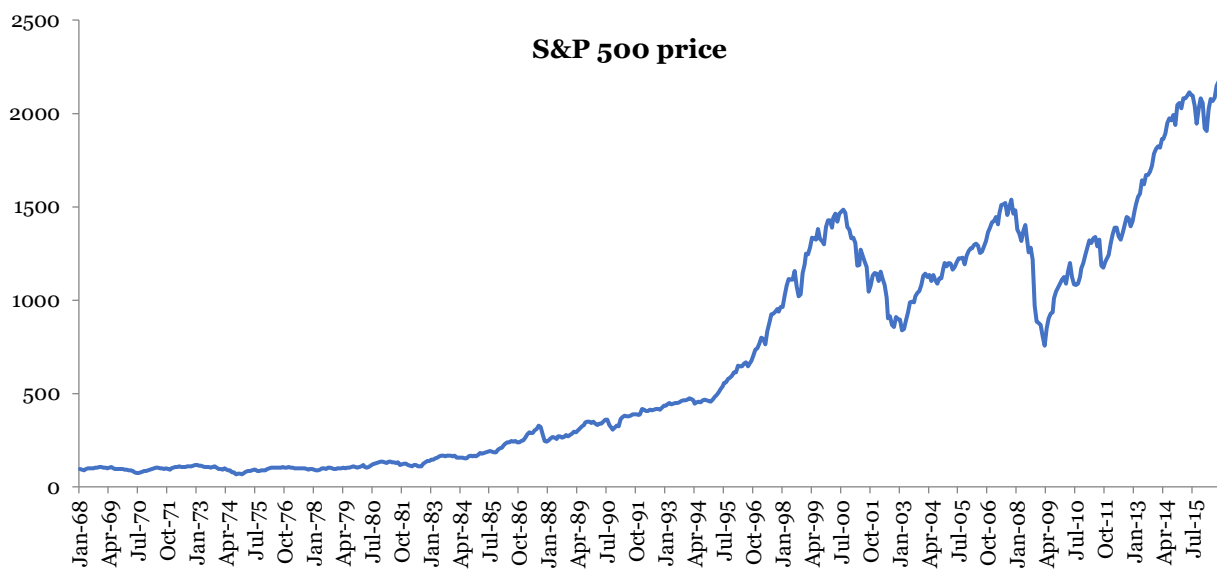


Figure 5 S&P500 price in 1968-2016 period

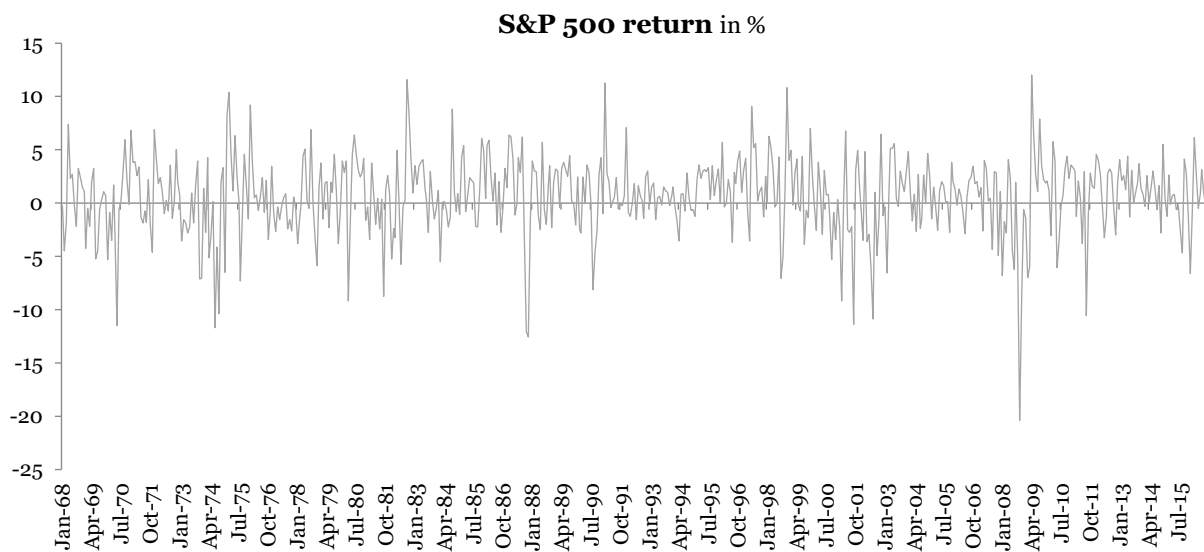


Figure 6 S&P500 return in 1968-2016 period

Interest rate related variables

Term Spread average value for the period is 1.7% and it varies from -2.7% to 4.4%, which corresponds to the times of high and low liquidity.

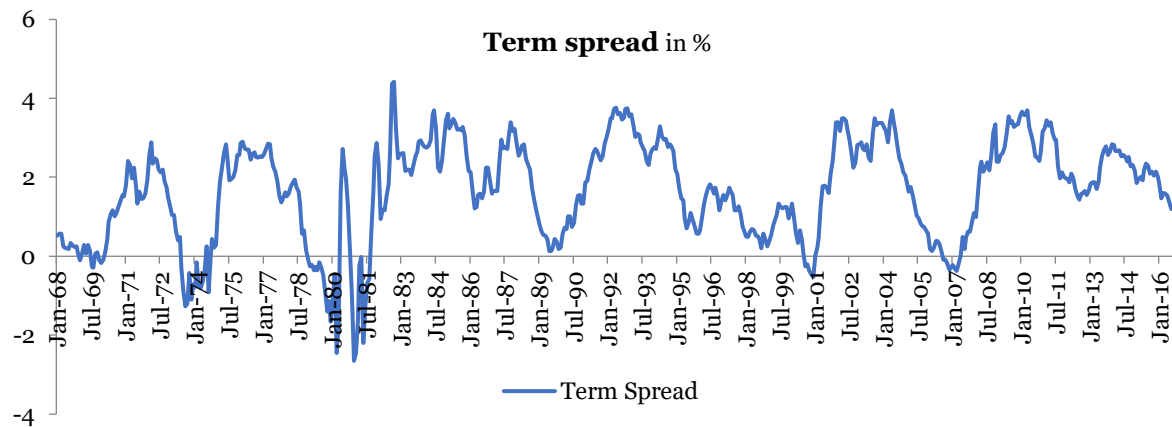


Figure 7 Term spread in 1968-2016 period

Treasury 10-year bond yield got an average value of 6.6%, while ranging from 1.5% to 15.3%. 3-month T-bill has the mean of 4.9%, varying from 0.01% to 16.3%. It is noteworthy that the maximum value of a 3-month T-bill is higher than that of the 10-year bond. From analyzing the time series, there are clear time periods identified, when the inverted yield curve is present (when the yield of short term security is higher than the yield on the long-term security), as depicted on the Figure 8. Several researchers found that inverted yield curve is a predictor of a forthcoming recession (Wang & Yang, 2012). This is evident here as well, as the inverted yield curve is observed during March-April 1980 before 80s Recession, August-December 2000 before the burst of a tech bubble, and August 2006-April 2007 before the subprime mortgage crisis.



Figure 8 10 year Treasury bond and 3 month Treasury note yields, 1968-2016

Credit spread (using Aaa and government bond) has the mean of 1.1%, and a range of -0.2% to 2.7%. The performance of the time-series is depicted in Figure 9.

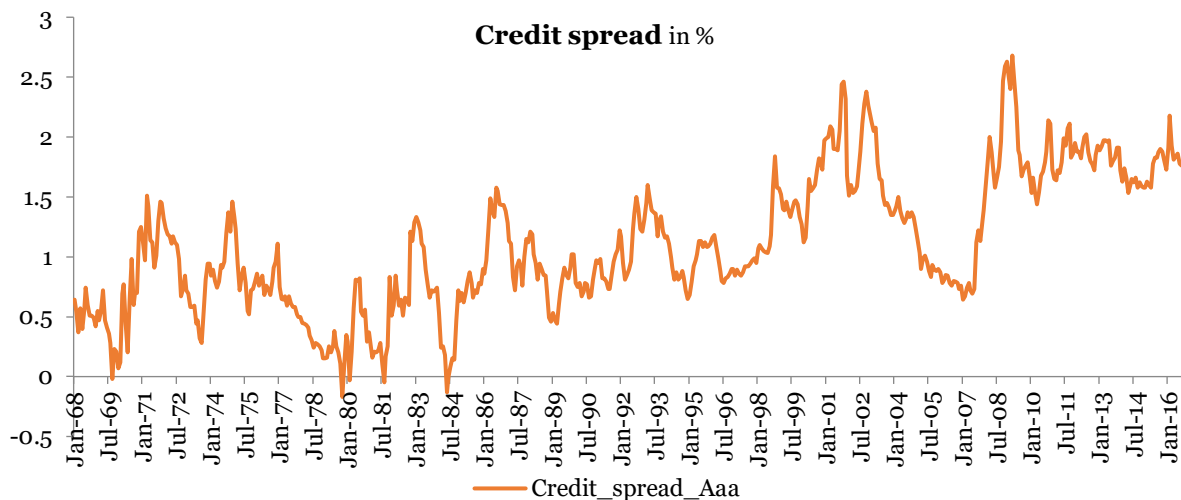


Figure 9 Credit spread for 1968-2016 period

Macroeconomic variables

Unemployment rate appears to have a negative relationship with S&P500 price index especially evident in the time period 1993-2016, as depicted in Figure 10. Unemployment rate changes also appear to be slightly ahead of stock index price changes, which gives hopes for its explanatory power in predicting stock prices and returns.

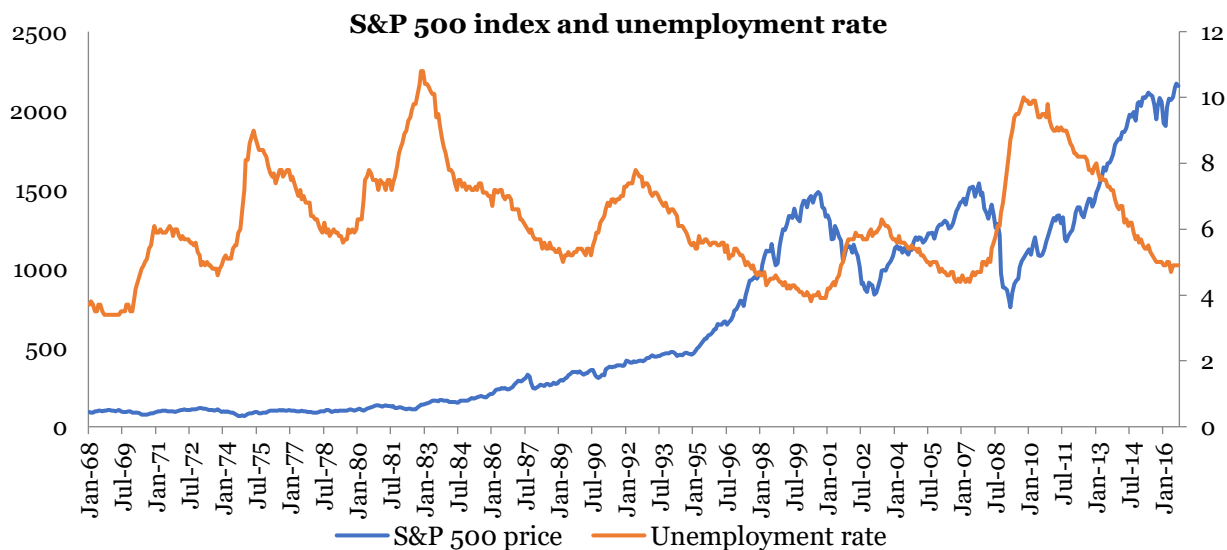


Figure 10 S&P500 index and unemployment rate, 1968-2016

Money supply has been increasing from its lowest \$183 billion to its highest value of \$3317 billion.

Consumer Price Index varies from 34 to 241, with the average value of 136.

Market Sentiment variables

Consumer sentiment index appears to have a positive relationship with S&P500 price index as depicted in Figure 11. It is especially evident for the most recent data 1997-2016.

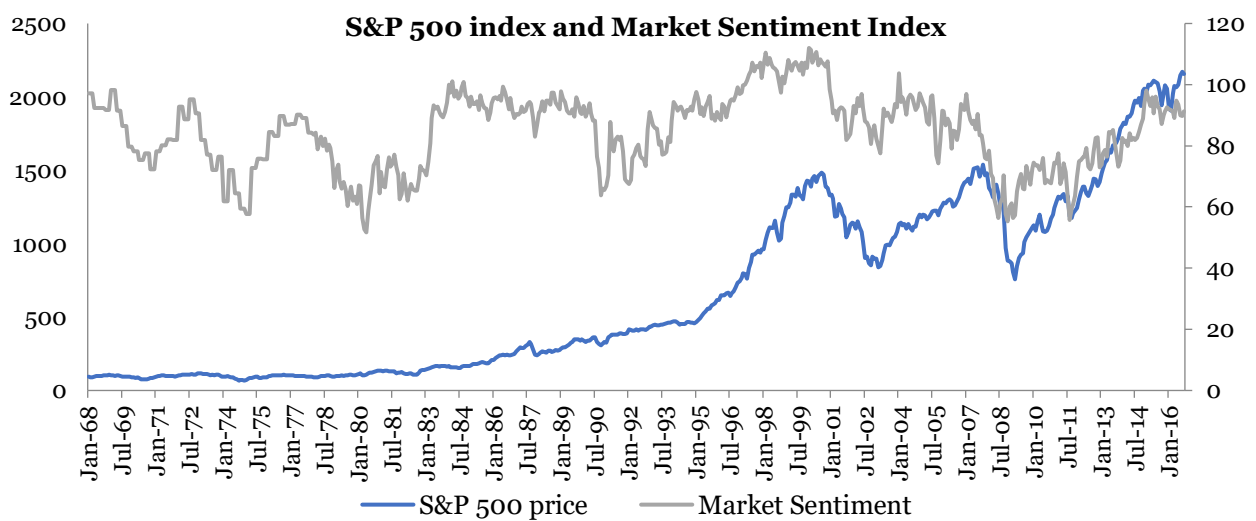


Figure 11 S&P 500 index price level series and market sentiment index

Commodities

Gold prices have seen a dramatic change from its minimal value of \$25.2/ounce in 1968 to record high of \$1828/ounce in August 2011, with an average value for the whole period being \$485/ounce as depicted on Figure 13. Similarly, price for Crude Oil has increased dramatically from \$3/barrel in 1968 to a record high of \$134/barrel in June 2008, with the average value of \$33/barrel, illustrated on Figure 12.

Normality of the data can be discussed using Skewness and Excess Kurtosis metrics. Data is normally distributed if it has 0 skewness (it is not skewed not to the left, nor to the right), as well as an excess kurtosis of 0 (no peakiness or fat tails in the data). For example, market return is only slightly skewed to the left side of the distribution. Change in crude oil variable

appears to have the most peakedness in it with fat tails, meaning high risk of the extreme values occurring (For instance, during the crude oil price shocks in 1974 (OPEC Embargo), 1986 (dramatic price decline following the decline in consumption) and 1990 (First Persian Gulf War)). Gold price change variable shows similar patterns of high peakedness. Change in currencies rates appear to be not-normally distributed, given the excess kurtosis.

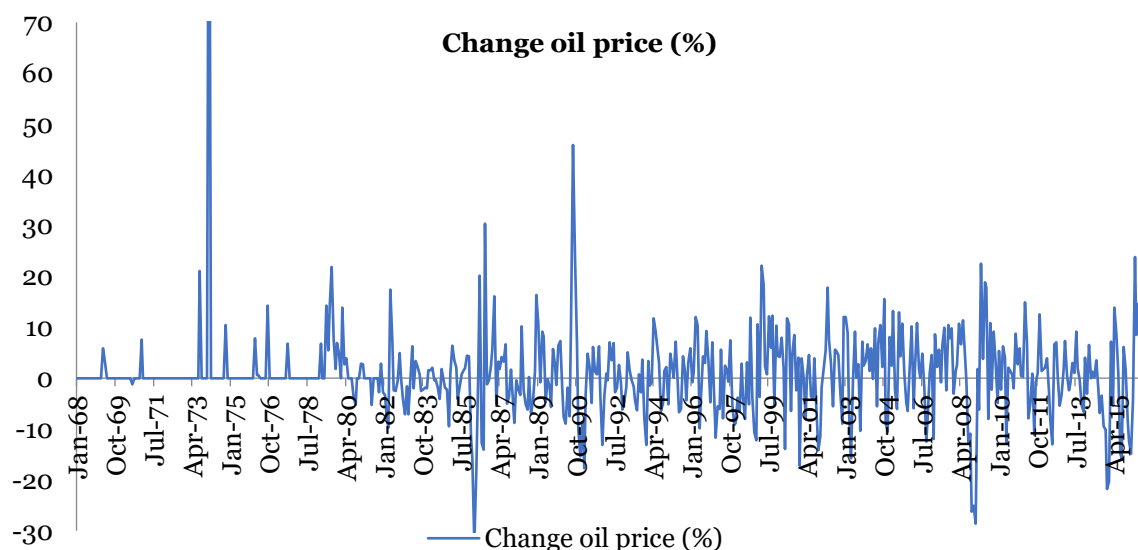


Figure 12 Changes in crude oil price during 1968-2016

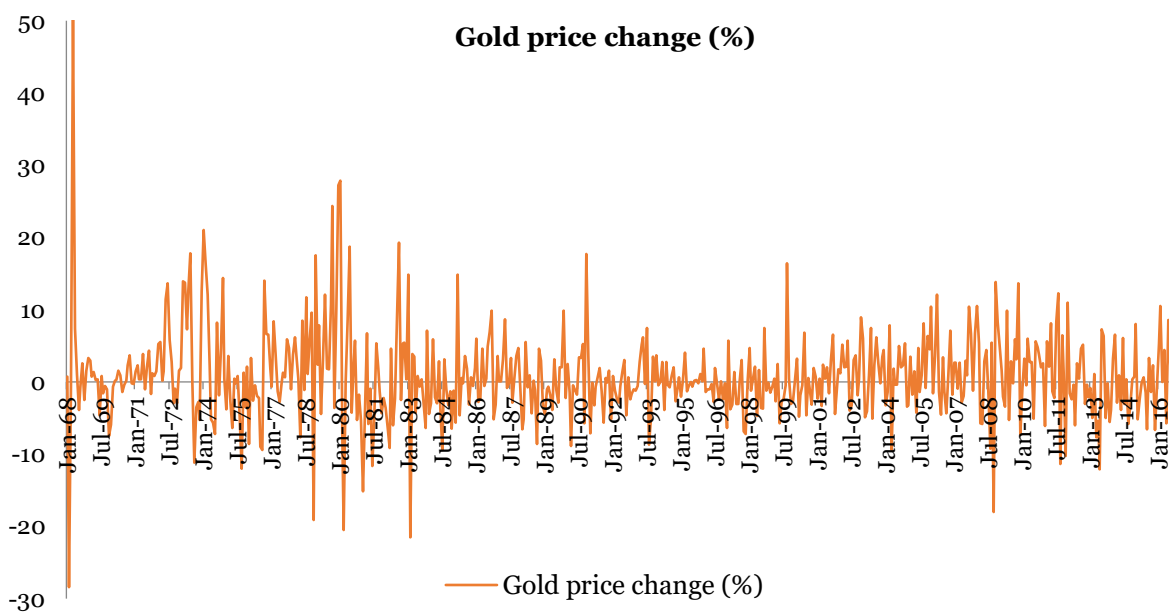


Figure 13 Changes in gold price during 1968-2016

Exchange rate specific variables

Change in GBP exchange rate varies from -12.1% to 13.4% while the change of JPY rate varies from -10.6% to 16.8%, both have average and median values very close to zero.

Stock index specific

Dividend yield has an average value of 3% for the time period and ranges from lowest 1.1% in August 2000 to 6.2% in June 1982.

5.1.3 Correlations

It is important to check for the correlations between the variables before running the models in order to identify important explanatory variables and check for the signs of possible multicollinearity.

It appears that S&P 500 return has the highest correlation with the change in its dividend yield, unemployment rate and market sentiment, with correlations being higher than the 5% critical value of 0.0811 (two-tailed, $n=585$), followed by term spread and change in gold price. The overall correlations with the dependent variable are weak, as they are lower than 15%. Given the fact that correlation matrix captures only linear correlations, there is a possibility that there could be also a non-linear relationship between the variables that is not captured in the matrix but could nevertheless exist, making variables useful in predictions.

Conversely, there are also signs of multi-collinearity, measured by high level of correlation between explanatory variables. Correlations between explanatory variables are the highest between CPI and exchange rates; credit spread and exchange rates; dividend yield and other explanatory variables: exchange rates, CPI, market sentiment, corporate and credit spreads. Judging from the matrix, dividend yield appears to be the most problematic given its high correlations with other variables, followed by exchange rates and CPI. Given this results, few modifications have been made to the variables (taking natural logarithm of CPI series, differencing the dividend yield, and removing similar variables (interest related), keeping only term spread and credit spread (Aaa) and EUR/USD exchange rate). The variables that are used in the model are presented in Table 5.

Refer to the Table 6 for the correlation matrix (cells highlighted in grey identify correlations between explanatory variables that are higher than 40%).

Table 5 **Variables for the first dataset**

N.	Variable	Description
1	S&P500return_n	Lags of the dependent variable, where n=1,2,3...n number of lags
2	Term spread	Difference between 10-year T-bond yield and 3-month T-bill yield.
3	Credit Spread	Difference between 10-year Treasury bond and Moody's Aaa bond yields
4	USD_GBP	Change in the currency exchange rate of U.S. Dollar per British Pounds
5	USD_JPY	Change in the currency exchange rate of U.S. Dollar per Japanese Yen
6	Dividend yield	Difference in S&P500 index dividend yield
7	Market Sentiment	Consumer confidence index measured by University of Michigan
8	CPI	Consumer Price Index, change in prices paid by urban consumers for goods and services
9	Unemployment	Percentage of unemployed people in United States, aged 16 and older at time t-1
10	Gold	Change in the price of gold
11	Crude oil	Change in crude oil price
12	Money Supply	Change in funds that are easily accessible for spending (M1 money Stock in FRED)

Table 6 **Correlation summary**[illegible]

High linear correlation between explanatory variables appears to be between Unemployment and Term Spread variables, CPI and Credit spread, and Market Sentiment and Unemployment. The correlations are significant (50-63%), however, given the theory and previous research, these variables are necessary for the model, therefore, were kept in. Multicollinearity can be further checked during model diagnostics section using Variance Inflation Factor (VIF)-values.

5.2 Dataset 2 description

Time period and frequency

Second dataset that will be used is daily S&P500 index return over the period of 10 years (24th May 2007- 22th May 2017), accounting to 2608 observations over a 10 year time frame.

This is a significantly long period, which allows the forecaster to learn from the long historical performance of the returns and forecast them in the future. It is also of interest to test for a different time period frequency, hence daily frequency has been chosen. It could also be argued that daily frequency time series might have a higher level of persistence than the monthly frequency.

Market

Analogously to the monthly dataset, the choice of picking a US market returns was due to arguable usefulness of this forecast even if one lives in another country (due to spill-over and an overall influence). Furthermore, this paper wanted to test for some new variables, that at this stage are reliably available only for the USA for the specified time period.

5.2.1 Variables

Dependent variable is S&P500 index daily return, which was retrieved from the Factset database.

Independent variables of interest could be grouped in the following manner: interest rate related; exchange rate related; commodities related; other indexes related; main USA companies related; and market sentiment related. The summary of the possible influential factors for S&P5 500 return are described in Table 7.

Table 7 Summary of potential factors useful for predicting the S&P500 return

N.	Variable	Description
1	S&P500return_n	Lags of the dependent variable, where n=1,2,3...n number of lags
2	Term spread	Difference between 10 year US government bond yield and 3 month US Treasury bill yield.
3	Credit Spread	Difference between 10 year Treasury bond yield and Moody's Aaa bond yield
4	USD_GBP	Relative change of currency exchange rate of U.S. Dollar per British Pounds
5	JPY_USD	Relative change of currency exchange rate of Japanese Yen per U.S Dollar
6	USD_EUR	Relative change in currency exchange rate of U.S Dollar per Euro
7	Gold	Relative change in the closing price of gold
8	Crude oil	Relative change in the closing price of crude oil
9	FTSE All_1	Financial Times Stock Exchange (FTSE) All Shares index return at time $t-1$.
10	DAX_1	Deutsche Boerse DAX index Return at time $t-1$.
11	JNJ_1	Johnson & Johnson stock return at time $t-1$
12	MSFT_1	Microsoft Corporation stock return at time $t-1$
13	Ko_1	Coca Cola Corporation stock return at time $t-1$
14	VIX_1	Volatility ("fear") index of market expectations of short term volatility, based on prices of S&P500 options. Measured in relative change at time $t-1$.
15	GT debt	Google Trends data of the proportion of google searches of the term "debt" out of total searches at that location and time
16	GT stocks	Google Trends data of the proportion of google searches of the term "stocks" out of total searches.
17	GT portfolio	Google Trends data of the proportion of google searches of the term "portfolio" out of total searches
18	GT inflation	Google Trends data of the proportion of google searches of the term "inflation" out of the total searches
19	GTS&P500 index	Google Trends data of the proportion of google searches of the term "S&P500 index" out of the total search enquires

Interest rate related independent variables are Term Spread and Credit spread. **Exchange rate related** variables are USD_EUR, USD_GBP and JPY_USD. **Commodities related variables** are Gold and Crude oil. These variables are identical to the ones used in the first dataset (only different frequency and time period). Therefore, the expectations are the same.

Other indexes related variables

Additional variables that are used in this dataset are **other indexes** than S&P 500, as few researchers have found there to be a spillover effect, when a change in one market (index) is happening slightly before the changes in other markers (indexes). For example, Niaki & Hoseinzade (2013) and Sheta et al (2015) in their question of forecasting S&P500 returns include lag returns of FTSE, DAX, CAC 40 index and Hang Seng Index in hopes of explaining current values of S&P500 returns with the lags of returns of other indexes. Therefore, for this study the following indexes have been collected: FTSE 100 (representing 100 largest companies on London Stock Exchange), FTSE All share (all eligible companies listed on London Stock Exchange) and Deutsche Boerse DAX (representing 30 largest and most liquid German companies listed on Frankfurt Stock Exchange) share indexes. The data was retrieved from Eikon database. The expectations are that S&P500 index is correlated with other stock indexes and perhaps some of the changes in other indexes occur earlier than in S&P500, thus giving some prediction power.

Major USA companies related variables

Furthermore, the study also includes share prices and returns of 3 big **companies in USA**: Johnson & Johnson, Microsoft and Coca-Cola corporation. Niaki & Hoseinzade (2013) and Sheta et al. (2015) used Johnson & Johnson, Microsoft, Exxon Mobil and Proctor and Gamble (using indexes' and companies' returns in day $t-1$).

Finally, the variables that have rarely been used in previous studies and therefore provide an interesting edge to this research are **market sentiment related variables**, such as VIX Volatility Index and Google Trends data. VIX is measuring market expectations of the short-term volatility of the S&P500 index, computed from the index's options prices. Google Trends data is also used to measure investors' consensus view of the near future volatility of the stock market, however without the use of complex computations of financial instruments as in VIX, since Google Trends data is solely based on the volume of a certain enquiry relative

to the total searches on all topic in that area and at that time. In their study of predicting Dow Jones Industrial Average prices using Google Trends data, Preis et al (2013) found that certain enquires have a predictive power of the index future prices and that investment strategy that incorporates Google trends data provides a higher return than buy-and-hold strategy.

In their study, they used 98 Google trend terms and found that word “debt” had the highest predictive power among other google searches that they have tested, the search for “stocks” was 3rd most effective, followed by “portfolio” and “inflation”. In addition to the before mentioned Goggle terms, this study is going to test for a predictive power of a new term - “S&P500 index”. Given the results from the study of Preis et al. (2013), the expectation for this research is that Google Trends terms should have a predictive power for stock market returns.

Plotting S&P 500 return and change in VIX index on a demonstrative sample of the observations in 2017 year, shows that there is a negative association between VIX and return. (depicted on Figure 14)

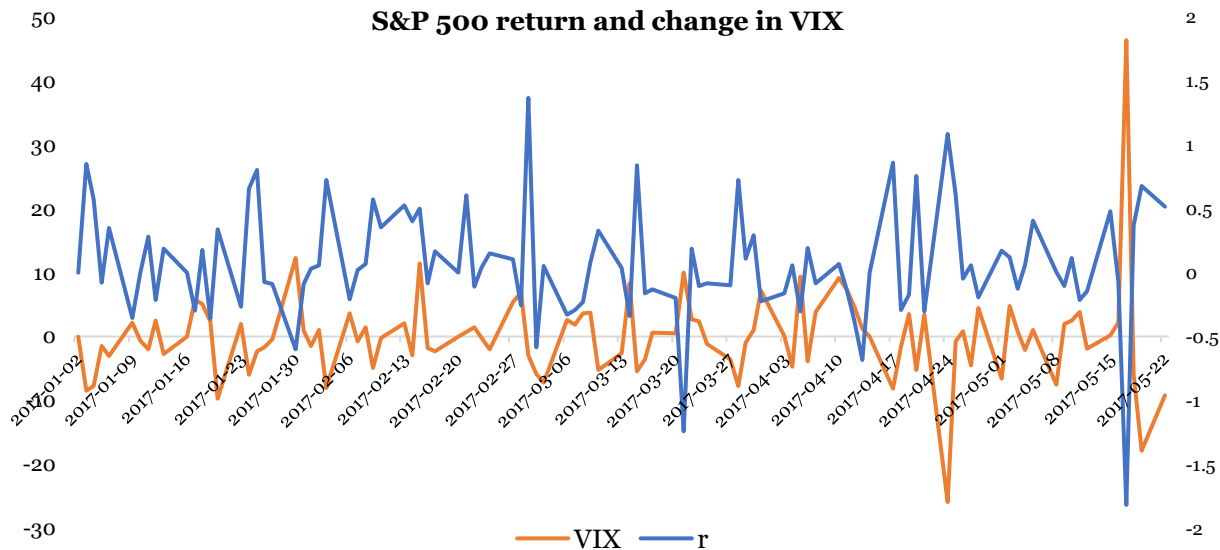


Figure 14 S&P 500 return on change in VIX

In comparison, Figure 15 depicts S&P 500 return and Google searches for the term “debt” series over the same period. Upon examination of this figure one can suspect a negative association to exist between these two variables.

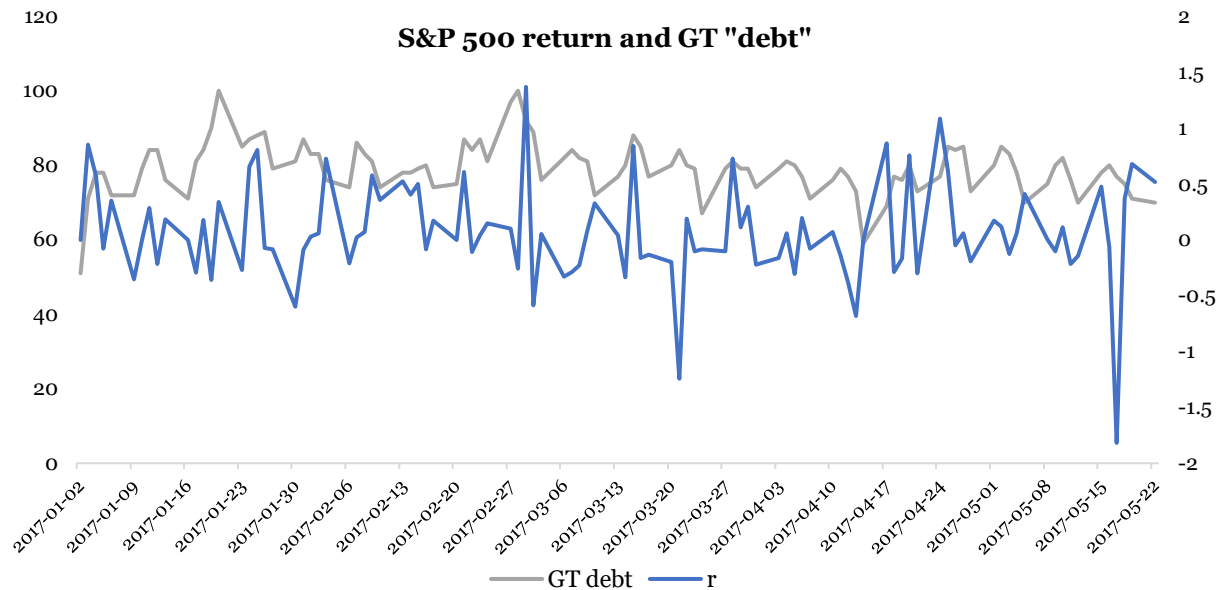


Figure 15 S&P 500 return and Google trends search on term “debt”

5.2.2 Descriptive statistics

This section provides descriptive statistics for the second dataset. The summary of the descriptive statistics is presented in Table 8.

Dependent variable (S&P500return) varies from -9% to 11.6% daily change over the period, with the average and median value of 0.

From the summary of the descriptive statistics one can see that stock indexes returns and individual companies’ returns have mean and median of 0, which means that the series are follow the white noise process (they have no trend) and average of 0. While these variable exhibits little or almost no skewness (meaning negative values as equally likely to appear as positive ones), they have an excess kurtosis, which means that the series have fat tails, i.e. there are many extreme outliers in the series (Brooks, 2014). This is not surprising as it is a typical feature of return series. (Unless the series are winsorized (normalizing the outliers), daily returns of a listed company or a stock index might go up and down quite dramatically, especially during the crises, which this data set includes (2008-2010)).

Table 8 Descriptive statistics summary for the daily dataset

	Mean	Median	Minimum	Maximum	Std. Dev	Ex. Skewness	Ex. Kurtosis
SP500	1550.8	1447.1	676.5	2402.3	423.9	0.2	-1.1
SP500 return	0.0	0.0	-9.0	11.6	1.3	-0.1	10.8
TermSpread	2.2	2.2	-0.2	3.8	0.8	-0.3	0.1
CreditSpread	1.8	1.8	0.6	3.0	0.3	-0.1	3.0
GoldChange	0.0	0.0	-9.2	7.1	1.2	-0.2	5.0
CrudeOilchange	0.0	0.0	-12.0	17.8	2.5	0.4	5.2
USD_EUR	1.3	1.3	1.0	1.6	0.1	-0.2	-0.7
USD_GBP	1.6	1.6	1.2	2.1	0.2	0.7	0.8
USD_JPY	99.3	100.1	75.7	125.6	14.1	0.0	-1.1
EUR change	0.0	0.0	-3.0	4.7	0.6	0.2	6.3
GBP change	-0.0	0.0	-7.8	4.5	0.7	-0.9	16.2
JPY change	0.0	0.0	-5.1	3.4	0.7	-0.2	7.6
FTSE100return	0.0	0.0	-8.8	9.8	1.2	0.1	7.6
FTSEAll_return	0.0	0.0	-8.3	9.2	1.2	0.0	7.1
DAX_return	0.0	0.1	-7.2	11.4	1.5	0.2	6.3
JNJreturn	0.0	0.0	-7.7	12.2	1.0	0.7	14.1
MSFTreturn	0.0	0.0	-11.7	18.6	1.7	0.5	11.2
Koreturn	0.0	0.0	-8.7	13.9	1.2	0.7	14.6
VIX return	0.26	-0.27	-29.6	50.0	7.5	1.2	8.3
GT_debt	58.5	64.0	11.0	100.0	21.5	-0.3	-1.1
GT_stocks	43.5	45.0	7.0	100.0	17.5	0.0	-0.8
GT_portfolio	37.4	41.0	9.0	66.0	13.9	-0.3	-1.0
GT_inflation	22.6	23.0	5.0	49.0	8.9	0.1	-0.8
GT_SP500Index	15.2	15.0	2.0	67.0	6.9	0.8	2.7

Term spread varies from -0.2% to 3.8% during the period with the average and median of 2.2%. Credit spread ranges from low 0.6% to 3%, with the average and median both settling at 1.8% for the time period.

Commodities returns exhibit similar patterns of the stock returns with the average and median value of 0 for both gold and crude oil returns. Series have insignificant skewness, however, similarly to other returns series possess excess kurtosis.

Exchange rates variables appear to be normally distributed. US dollar for Euro rate varies from 1 to 1.6 US dollar for one euro, while US dollar to GB pound varies from 1.2 to 2.1 US dollars for one pound. The change in exchange rate of USD/EUR varies from -3% to 4.7%, change in USD/GBP varies from -7.8% to 4.5% and change in rate of JPY/USD varies from -5.1% to 3.4%. The change in exchange rate series all exhibit excess kurtosis.

Finally, moving on to market sentiment measurement. Google trend data is relatively normally distributed given by its minimal skewness and excess kurtosis. Word “debt” was most commonly used enquiry, followed by “stocks”, “portfolio”, “inflation” and “S&P500 index”, in the order of their popularity. Google searches ranged from low 2 (units of interest as the proportion of total searches) to the high 100 (representing the maximum amount of searches that is possible during that location and that period of time. VIX exhibited slightly higher variation, given by positive skewness of 1.2 and excess kurtosis of 8.3.

5.2.3 Correlation analysis

The highest correlations that S&P500 return has with independent variables are with variables representing other stock market indexes returns (FTSE 100, FTSE All Share, and DAX) and major US companies returns (Johnson & Johnson, Microsoft and Coca Cola corporation). However, the companies’ and indexes’ returns are also highly correlated between themselves (for example, DAX return is highly correlated with FTSE 100 and FTSE All Share returns; Microsoft returns are positively correlated with that of FTSE, DAX, and JNJ), this means that only one variable from each category should be chosen. Exchange rates also show signs of multicollinearity, as US dollar to euro exchange rate is positively correlated with US dollar to pounds rate. Google trends terms are highly collinear. This means that certain variables that share similar characteristics should be removed, in order to minimize the multicollinearity problem. Therefore, the following variables were removed: FTSE 100 and DAX indexes (while keeping FTSE All Share), change in the exchange rate of USD to euro (while keeping exchange rates of USD to GBP and JPY), and Google trends searches on “portfolio”, and “inflation” (while keeping “debt” as it was found to be the most significant in previous studies, and “S&P500index”, as there is an expectation that it will bring explanatory power into the model). The correlation matrix for the variables that were chosen to be left in the model is presented on Table 9. The full correlation matrix is depicted in Appendix 1.

Table 9 Correlation matrix of variables for daily dataset

	R	Term	Credit	Gold	Oil	GBP	JPY	FTSE	MSFT	VIX	GT_debt	GT_index
R	1	0,02	0,00	-0,02	0,34	0,22	0,28	-0,05	-0,06	-0,13	0,01	-0,02
Term		1	0,34	-0,01	0,01	0,01	0,02	0,01	0,01	-0,02	-0,03	-0,21
Credit			1	0,00	0,01	0,02	0,02	0,00	0,00	-0,02	-0,10	-0,15
Gold				1	0,15	0,25	-0,24	0,02	0,03	0,01	0,01	0,02
Oil					1	0,27	0,14	-0,01	0,02	-0,26	0,00	-0,03
GBP						1	-0,01	0,04	0,12	-0,17	0,02	-0,04
JPY							1	0,00	0,08	-0,26	0,00	-0,04
FTSE								1	0,42	0,02	0,00	-0,05
MSFT									1	0,05	0,00	-0,04
VIX										1	-0,01	0,03
GT_debt											1	0,67
GT_index												1

The most significant correlation S&P500 return has with the following independent variables: change in oil price, change in JPY exchange rate and change in GBP exchange rate, measured by 34%, 28% and 22%, respectively, indicating moderate relationship between them. High correlation between explanatory variables themselves could be seen between return of FTSE index and return on MSFT stock (42%), and between Google trend variables (GT debt and GT S&P500index) with 67% correlation, meaning that only one GT term should be used in the model. This study proceeds with the GT term of “S&P500 index”.

6 RESEARCH METHOD

6.1 Research Question and Research Hypothesis:

Based on the discussed theories and literature review, the **research question** of this study is: Do artificial neural networks produce a more accurate forecast than traditional forecasting methods?

Based on previous research, the author of this paper develops the following **research hypothesis**: ANN has a better forecasting accuracy than the traditional linear forecasting methods.

6.2 Traditional long-term forecasting methods

Many studies have compared the performance of ANN with Box–Jenkins’ ARIMA (autoregressive integrated moving average), random walks and exponential smoothing (Hwarng, 2001; Maia & Carvalho, 2011; Panda & Narasimhan; 2007). Therefore, this study will test ARIMA models for univariate prediction of the return with linear models. Furthermore, it will use Vector Autoregressive models (VAR) as a proxy for traditional multivariate forecasting methods (linear).

6.2.1 ARIMA

At first, it is good to start from building the simplest univariate model that would predict future values of the series given only its own historical values and past values of its error term. ARMA (Autoregressive moving average) models combine Autoregressive (of order p) and Moving Average (of order q) components of time series.

Following Box and Jenkins methodology (1976) for modeling time series with ARIMA, the first step is the model identification (ACF, PACF, and Information criteria (AIC, SBIC and HQIC) will be used in identifying appropriate model). Second step will be to perform parameter estimation (using least squares or maximum likelihood). Third step would be to check the model, using residual diagnostics and overfitting. (Brooks, 2008). Additionally, diagnostic checks will be used to measure model adequacy.

6.2.1.1 General model description for ARMA

ARMA (p, q) model:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

or using lag polynomial, ARMA process can be written as:

$$\varphi(L)x_t = \theta(L)\varepsilon_t$$

6.2.1.2 Stationarity assumptions

One of the model assumption is that the series are stationary, which is not often the case in financial time series. Stationarity of the time series means that that series have a constant mean, constant variance and constant auto-covariance structure. (Brooks, 2008). It is important to distinguish between stationary and non-stationary series as this feature affects the behaviour and properties of the series which needs to be accounted for in econometric modelling.

In order to create stationary time-series, it is necessary to difference the original series. Often 1st differencing is enough to make non-stationary series into stationary, however some series require further differencing of order d , where $d=0,1,2,\dots$ (if $d=0$, the original series were stationary; if $d=1$, one order differencing was required to make a series stationary, etc.). For example, S&P 500 price series are not stationary as evident on the Figure 5, its differenced series (S&P500 return) appears to be stationary. Apart from the graphical analysis, it is necessary to run tests to ensure the return series are stationary.

One such test is Augmented Dickey-Fuller (ADF) test which examines whether the series contain a unit root: if the series contain a unit root, it is a sign of non-stationarity, hence, one needs to difference the series to make it stationary. If a null hypothesis of the existence of a unit root is rejected, there is no need to perform the differencing.

According to the ADF test on the monthly dataset, all variables were stationary except for CPI and Dividend yield (original series). Therefore, in order to make each respective series stationary, natural logarithm was taken out of CPI, while Dividend yield was differenced (showing the change in dividend yield). Variable representing exchange rates (original series) were also differenced in order to get stationary variables (relative change in exchange rates). Similar changes have been conducted to the daily dataset, where ADF test indicated that original series of exchange rates were not-stationary series, therefore their difference was taken.

6.2.2 VAR

Vector autoregressive models (VARs) is a system (or vector) regression model that allows for more than 1 dependent variable, permitting for the time series to depend not only on its own lags and disturbance terms, but also on the other explanatory variables and their lags. VAR is often considered as a hybrid between univariate autoregressive models and simultaneous equations models. (Brooks, 2008). It is possible to create VARMA - extended multivariate regression model of ARMA.

The benefit of VAR model is that it does not impose the restriction of identifying certain variables as exogenous, leaving the researcher with more flexibility as all variables are considered endogenous. Furthermore, several researches found the forecasts of VAR models to be superior to that of the traditional structured models. (Brooks, 2008).

According to Brooks (2014), one of the main weaknesses of the VAR modelling is the difficulty of interpreting the coefficients, given its large structure and diverse relationship with different lags (one lag could have a positive association with the variable of interest whereas next its lag might have a negative association). However, as the Brooks points out, this VAR feature becomes a problem only when the sole purpose of the researcher is to interpret the coefficients. There is no problem using VAR when the purpose is to forecast, which is the case for this study.

While VAR represents a system of linear equations each of them is being very similar to a simple OLS regression, it offers additional benefits. First benefit is its autoregressive component, that allows to incorporate not only past lags of the dependent variable but also the lags of the independent variables. Moreover, having an autoregressive component allows for a more profound time series forecasting. Furthermore, VAR allows much more deeper analysis of the relationship between the variables in time-series setting, using granger causality, impulse response and variance decomposition tests. This is particularly useful in this study since one of the purposes of this study was to analyze the relationship between market sentiment and stock index returns.

6.2.2.1 General model description for VAR

As VAR can be sought of as a system of linear equations, it is possible to write down VAR as n -amount of equations, where n is the number of variables.

For example, a VAR model with two variables would be written in the following way:

$$y_{1t} = \beta_{10} + \beta_{11} y_{1t-1} + \alpha_{11} y_{2t-1} + \mu_{1t}$$

$$y_{2t} = \beta_{20} + \beta_{21} y_{2t-1} + \alpha_{21} y_{1t-1} + \mu_{2t}$$

or when there are k amount of lags it could be written compactly as below (where each symbol is a vector of values):

$$\mathbf{y}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{y}_{t-1} + \boldsymbol{\beta}_2 \mathbf{y}_{t-2} + \dots + \boldsymbol{\beta}_k \mathbf{y}_{t-k} + \boldsymbol{\mu}_{1t}$$

Using lag operator, the model can be written as following

$$\boldsymbol{\Pi}(L)\mathbf{Y}_t = \mathbf{c} + \varepsilon_t, \text{ where } \boldsymbol{\Pi}(L) = \mathbf{I}_n - \boldsymbol{\Pi}_1(L) - \dots - \boldsymbol{\Pi}_p(L)^p$$

6.2.3 In sample forecasting versus out-of-sample

In order to perform out of sample forecasting, the datasets are split into two sub-samples, one of which will be used for the model estimation and second one will be used as out of sample forecasting. For the monthly data series, the estimation sample is from January 1968 to December 2013, thus leaving observations from January 2014 to September 2016 for out-of-sample forecasting (N=153 observations). The estimation sub-sample for the daily data set is from 24th of May 2007 to 28st of March 2016, thus leaving observation from 29st of March 2016 to 22 May 2017 (N=300 observations) for out-of-sample forecasting.

6.3 Research method for ANN

The chapter discusses the process of building, training, evaluating and using the neural network.

6.3.1 Process of constructing and utilizing ANN

6.3.1.1 Building

Each type of neural network differs by in its architecture. The process of building an ANN is explained below of the example of multi-layer perceptron (MLP).

MLP has one input (matrix of independent variables) and one output layer (the dependent variable in question). In addition, in order to allow a more complex interaction between variables, hidden layers are used. Typically, one or two hidden layers are chosen. Rarely does

a network include more than three hidden layers as it complicates the training and computation and can result in over-fitting. Moreover, the researcher can determine specific number of neurons in the hidden layers. Theoretically, more complicated system can produce much more accurate result (for example deep-learning), however, the training of such system is usually very long as well as the accuracy of the results can be corrupted by problem of over specification. (Kriesel, 2007)

Another aspect of the neural network is the transfer function, which refers to the way neurons react to the input values. There are few common types of the transfer function, namely, 1) binary threshold function (if the input surpasses the defined threshold, the function changes from one value to another, but otherwise remains the same); 2) logistic function which allows for the range of values between 0 and 1; and 3) hyperbolic tangent which can take values from -1 to 1.

6.3.1.2 Learning processes

There are different learning processes: supervised, unsupervised, semi-supervised and reinforcement learning. Unsupervised learning means no human intervention in the process of how the system learns the algorithm from the data provided, i.e. there is only the input given and no output (no right answers), and the network attempts to recognize patterns and to classify them into categories. (Kriesel, 2007).

Supervised learning – the output is provided for the training process. This is one of the first learning methods created, and it is also proven to be effective in forecasting, hence often practiced by the professionals (Kriesel, 2007).

Reinforcement learning, the system tries through trial and error to find a solution that maximizes the reward (right answers).

6.3.1.3 Training methods

Different training methods are available for training the neural network, some of them are listed below.

- **Backpropagation learning (BP)** supervised learning algorithm, the workhouse for the neural networks training.
- **Levenberg-Marquardt method (LM).** This training algorithm requires more memory but less time. Training automatically stops when generalization stops

improving (which occurs when mean square error of the validation samples starts to increase). It has received its popularity given its favorable quality of approximating the second-order derivative with no need to compute the Hessian matrix, which consequently reduces the learning time significantly.

- **Bayesian Regularization (BR)** training algorithm which often results in good forecasting for difficult, small or noisy datasets, although requiring more training time. Training process stops according to adaptive weight minimization (regularization).
- **Scaled Conjugate Gradient (SCG)** algorithm is also very fast because it requires less memory. By a similar design to BR, SCG's training process stops when generalization stops improving as indicated by a growth in the mean square error of the validation samples.
- **Extreme learning machine (ELM) learning algorithm** is used for training of single hidden layer feedforward neural network (SLFN) for training MLP network. Huang et al (2006) and Ding et al.(2015) find that ELM is a simpler neural network, it avoids the problem of overfitting and specification of the parameters (n,p,q), its learning speed is much faster compared with other methods, hence the training time only take seconds, and it also avoid the problem with correlation of the variables. (Grigorievskiy et al., 2013).

6.3.1.4 In- sample versus out-of-sample forecasting

Network allocates 15% of the data to testing. Those observations in the testing sub-sample are not used for the network modelling, however it is still considered as the in-sample forecasting (as the time period of the testing data is included in the total dataset). On the other hand, out-of-sample forecasts independent values during the period that lies out of the specified sample. The researchers suggest that neural network's performance varies significantly between in-sample and out-of-sample forecasting, therefore both measures will be used. In order to perform out-of-sample forecasting the in-sample observations are reduced by 153 in the monthly data set and 300 in the daily dataset (identical to the ARIMA and VAR out of sample definition), allowing to use them for the out-of -sample forecasting. Thus, the estimation sample for the monthly dataset is from January 1968 to December

2003, leaving observations from January 2014 to September 2016 for out-of-sample forecasting (N=153 observations). The estimation sub-sample for the daily data set is from 24th of May 2007 to 28st of March 2016, thus leaving observation from 29st of March 2016 to 22 May 2017 (N=300 observations) for out-of-sample forecasting.

6.3.2 NAR network

For the time-series problem, it is advisable to use dynamic networks, where the output depends not only on the current input to the network, but also on the current or previous inputs, outputs or states of the network (MATLAB, 2016). Dynamic networks are considered to be more powerful than static, feed-forward networks, as dynamic networks can be trained to learn time-varying or sequential patterns, however they might be more difficult to train. One of such networks belonging to the recurrent-dynamic networks group is NAR network which is non-linear auto-associative time-series network. NAR makes the future prediction for time-series using that series past values.

6.3.2.1 NAR network structure

NAR network structure could be written in a formula format as below:

$$y(t) = f(y(t-1), y(t-2), \dots, y(t-n)) + e(t)$$

where $y(t)$ is a time series which values network is trying to forecast using series own n number of lags, $e(t)$ is the error term that occurs as a result of a difference between forecasted and actual values, while $f(\cdot)$ is the transfer function of network which is often log-sigmoid (as predetermined by MATLAB), however it could be re-specified if desired.

$y(t-1)$, $y(t-2)$, \dots , $y(t-n)$, are called feedback delays, and can be thought of as an input layers in the system, while $y(t)$ is the output of the network.

The graphical illustration of one NAR network is depicted on Figure 16, where $y(t)$ is the series that we want to predict (whose lags we use as an input). “1:2” notation means that this network was designed for 2 delays. w stands for weights vector and b for biases. The transfer function of the network’s hidden layer is tan-sigmoid, while in the output layer it is linear transfer function. The transfer function of the output layer is linear due to question of forecasting of the actual return level, not the likelihood of up or down movement, for which tan-sigmoid function should be used.

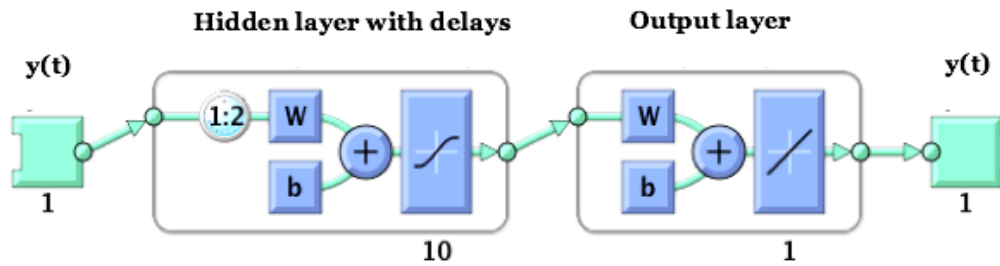


Figure 16 NAR architecture

6.3.2.2 Hidden layers and hidden neurons specification

Number of hidden layers varies given the data structure, however, it is often advised not to overfit the network by introducing many hidden layers into the network. Therefore, the network with one hidden layer will be used. Number of neurons in the hidden layer is also flexible, and by default MATLAB software suggests 10 neurons. While the robustness of the forecasting performance of the network given different number of neurons in the hidden layer can be checked, some researches warn of not increasing the number too much and making the system too complex and inefficient, and not decreasing too much and restricting the network from attaining best the generalization capabilities and computing power. (MATLAB, 2017; Ruiz et al., 2016).

6.3.2.3 Data spilt specification

The data is split in the following manner:

- 70% will be used for training of the network,
- 15% for validation (measure network generalization and to stop training when generalization stops improving) and
- 15% for testing (provide and independent measure of network accuracy as these testing data observations have no effect on the training).

This data sample split was found to be optimal by MATLAB and is used as a default that is not recommended to change (MATLAB, 2017). This means that for monthly data set, 409 observations are used for training, and 88 for validation and testing, while for daily data set there will be 1826 observations used for training, 391 and 391 used for validation and testing, respectively.

6.3.2.4 Training algorithm

For this research Levenberg-Marquardt (LM) learning algorithm will be used and the robustness of the results will be checked if the network is trained Bayesian Regularization (BR) and Scaled Conjugate Gradient (SCG).

6.3.2.5 Number of lags

The default suggestion of MATLAB software is to use 2 lags in the network (it is often found to be superior, as it catches some dynamics of the data, yet doesn't not complicate the network too much). Therefore, the base network will be using 2 lags and the robustness of the results will be checked using 6 or 12 lags.

6.3.3 NARX network

While NAR is the basic network with explanatory variables but its own lags, it is often advisable to use a more sophisticated network that also includes other explanatory variables and their lags into the network, as those networks usually can capture more of the dynamics of the time series. One such recurrent dynamic neural network is NARX - nonlinear autoregressive network with exogenous inputs (NARX). It has feedback connections surrounding several layers of the network. NARX is often found to be superior in forecasting accuracy than NAR networks (Ruiz et al., 2016).

6.3.3.1 NARX structure

NARX can be presented as:

$$y(t) = f(x(t-1), x(t-2), \dots, x(t-n_x), y(t-1), y(t-2), \dots, y(t-n_y)) + e(t)$$

where $y(t)$ is a dependent output, that is regressed over the its own previous values as well as the previous values of the independent input (x). The input can could be one variable (for example, GDP or unemployment rate), or could represent a matrix of variables values over the time-series.

Transfer function $f(\cdot)$ will be log-sigmoid, as it is often mentioned in the previous works as well as it is the pre-determined function in MATLAB. The function however could be re-specified, if desired.

Graphical illustration is presented below on Figure 17.

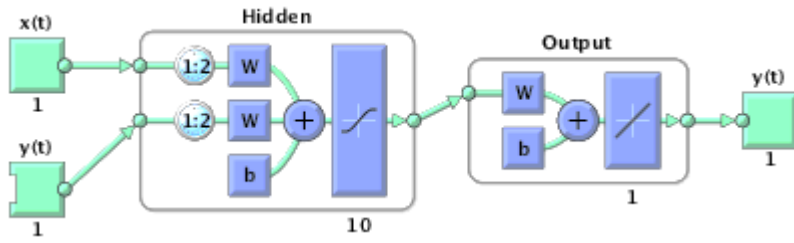


Figure 17 NARX network structure (Source: MATLAB, 2016)

6.3.3.2 Hidden layers and hidden neurons

Typically, there is one hidden layer, thus it will be used in this paper as well.

6.3.3.3 Training algorithm

The training methods will be the same as for NAR network: Levenberg-Marquardt back propagation, which is recommended by the MATLAB guide (2016) as a method that quickly and accurately trains the network, and to test it against Bayesian Regularization, and Scaled Conjugate Gradient learning algorithms.

6.3.3.4 Number of lags

Number of lags depended on the dynamics of the data. First, the network with 2 lags will be built after which the results can be compared to the ones from the networks that use more lags. This study uses networks with 2, 6 and 12 lags.

6.4 Method for comparing the model's forecast accuracy

There is no clear superior measurement for the model accuracy, as the choice of the measurement depends on the reasons for performing a forecast, does the person wants to find the absolute value, or can he/she sacrifice some accuracy in order to get better results of the overall movement of the variable in question. Hence, the measurements for the forecast accuracy can be divided in the following categories: statistical loss functions and financial/economic loss functions.

6.4.1 Statistical loss functions

Given the fact that in-sample and out-of-sample forecasting differs, the measures of statistical loss functions are subsequently divided into measures for in-sample and out-of-sample forecasting measures.

6.4.1.1 In-sample measures

In order to check which model produces higher accuracy for in-sample forecasts, the following evaluation criteria can be used: mean squared error (MSE), mean absolute error (MAE), square root of mean squared error (RMSE), mean absolute percentage error (MAPE), Theil's U- statistics and others (Brooks, 2008). Most common measures in the previous research were MAE and MSE (Khashei & Bijari, 2010; Panda & Narasimhan, 2007).

MSE provides a quadratic loss function that penalizes large forecast errors much heavily and can be written as follows, where e_t is an error that occurs as a result of the difference between actual value y_t and its forecasted value (in-sample), n is the number of observations:

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

MAE measures average absolute forecast error. In comparison to MSE it penalizes the large errors equally proportionately as small errors (Brooks, 2014) The formula is given below:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

Root mean square error is simply MSE squared:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

Mean Absolute Percentage error formula is given below:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right|$$

6.4.1.2 Out-of-sample measures

Out-of-sample measures are marginally different to the statistical measures of the in-sample forecasting. One of the most known measure of out-of-sample statistical loss function is Mean Square Prediction Error (MSPE), which is similar to the MSE, but for the out-of-sample forecasting, where the error is the difference between forecasted value for out-of-

sample and the actual value. The formula is given below, where m is the number of forecasting periods, and j is the number of steps ahead in forecasting.

$$MSPE = \frac{1}{m} \sum_{j=1}^m (\hat{y}_{n+j|n+j-1} - y_{n+j})^2$$

Another one is Mean Absolute Percentage Error (MAPE). It is a relative measure of forecast accuracy and hence is widely used in forecast comparison. It is a very close measure to the MAE for the in-sample forecasting.

$$MAPE = \frac{1}{m} \sum_{j=1}^m |\hat{y}_{n+j|n+j-1} - y_{n+j}|$$

Franses and van Dijk (2003) note that in some cases returns display erratic behavior and might have large outliers, for which reason sometimes it is better to use median values, hence the formula would use instead Median counterparts, for example, MedSPE (median square prediction error) instead of MSPE (mean error prediction error)

6.4.2 Financial/economic loss function

However, some researchers suggest that statistical loss functions may be not so relevant in the real world, hence giving some motivation for the creating of financial/economic loss functions.

The researchers argue that the exact value of the predicted stock return is not that much of interest to most people, whereas general indication of the market movement for the future time could be very useful as the trade can be executed if the forecasted price is higher than the current one and vice versa. (Pesaran & Timmerman, 1992).

6.4.2.1 Percentage of correct direction change predictions

One financial measure to measure the accuracy of the forecast prediction of the stock index is to calculate the percentage of the correct market movement predictions. (Brooks, 2014).

The formula for it is given below:

$$\% \text{ correct direction change prediction} = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^T z_{t+s}$$

$$\begin{aligned} \text{where } z_{t+s} &= 1 & \text{if } (y_{t+s} - y_t) \cdot (f_{t,s} - y_t) > 0 \\ z_{t+s} &= 0 & \text{otherwise} \end{aligned}$$

6.4.2.2 Percentage of correct sign prediction

Formula for correct sign of the forecast prediction is the following:

$$\% \text{ correct sign prediction} = \frac{1}{T - (T_1 - 1)} \sum_{t=T_1}^T z_{t+s}$$

$$\begin{aligned} \text{where } z_{t+s} &= 1 & \text{if } (y_{t+s} \cdot f_{t,s}) > 0 \\ z_{t+s} &= 0 & \text{otherwise} \end{aligned}$$

It is also called Success ratio (SR) (in Franses and van Dijk (2003)) or Hit ratio in Kumar (2009). Success ratio formula is given below (where I is an indicator variable that takes value of 1 if the condition described in brackets is satisfied, and 0 if not:

$$SR = \frac{1}{m} \sum_{j=1}^m I_j [y_{n+j} \cdot \hat{y}_{n+j|n+j-1} > 0]$$

When evaluating the metrics of forecast accuracy such as percentage of correct direction or sign predictions of the total predictions, one can use simple heuristic rule of thumb, that any model offering a ratio of higher than 50% is bringing a competitive edge and has some usefulness. (50% due to the fact that the result of the forecast takes only two meaningful values, either it predicts that the stock goes up or down; or whether the stock return is positive or negative).

A more sophisticated user can Pesaran and Timmermann **Direction Accuracy (DA)** test (1992) that accounts for success rate in case of independence (SRI) and therefore can be used as a measure of significance of the Success Ratio (i.e. testing whether produced Success Ratio is statistically different from a Success Ratio that could have occurred by a random prediction).

For this test, the following formula is used:

$$DA = \frac{(SR - SRI)}{\sqrt{\text{var}(SR) - \text{var}(SRI)}} \sim N(0, 1)$$

The necessary definitions for it are:

$$P = \frac{1}{m} \sum_{j=1}^m I_j[y_{n+j} > 0]$$

$$\hat{P} = \frac{1}{m} \sum_{j=1}^m I_j[\hat{y}_{n+j|n+j-1} > 0]$$

Success rate in case of independence (SRI) is

$$SRI = P \hat{P} + (1 - P)(1 - \hat{P})$$

and variances are given by:

$$\text{var}(SRI) = \frac{1}{m} \left[(2\hat{P} - 1)^2 P(1 - P) + (2P - 1)^2 \hat{P}(1 - \hat{P}) + \frac{4}{m} P\hat{P}(1 - P)(1 - \hat{P}) \right]$$

$$\text{var}(SR) = \frac{1}{m} SRI(1 - SRI)$$

6.4.2.3 Sign test by Diebold and Mariano for comparing forecasts

Comparing the accuracy of forecasts and concluding which forecast has the best accuracy can be done using the Sign test developed by Diebold and Mariano (1995). The test aims at comparing “loss differential” of the forecasts, to determine whether the difference is significant or not. (Franses & van Dijk, 2003).

The test will be conducted following Diebold and Marino (1995) methodology. First step of the test is to take the forecast errors e_j of two forecasts that will be compared. (Forecast errors are defined as a difference between actual value of time-series y_j and forecasted value of that series at time j (that was made at $(j-1)$ time period).

$$e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t}$$

Then apply appropriate loss function $g(\cdot)$ to the errors. Most common examples of loss function are squared-error loss function or absolute error loss. Then the loss differential d_j will be defined as a difference between errors of forecast 1 and 2 (with applied functions):

$$d_j = g(e_{1j}) - g(e_{2j})$$

Loss differential is then used in **Sign test (S test)**, as below

$$S = \frac{2}{\sqrt{m}} \sum_{j=1}^m \left(I[d_j > 0] - \frac{1}{2} \right) \sim N(0,1)$$

where $I(d_j)$ is an indicator variable that takes value of 1 if $d_j > 0$ and 0 otherwise.

The test has a null hypothesis of no difference between the forecast errors of two comparing forecasts and of loss differential being not significantly different from zero. If the test gives a value that is statistically different from zero (outside the confidence interval of ± 1.96), it implies that the forecasts differ in the forecasting accuracy and one method is superior to another in solving that particular time-series prediction problem. (Diebold & Mariano, 1995)

6.5 Statistical hypotheses

From the research methods and data section description, this thesis is aiming to answer the following statistical hypotheses. There are hypotheses that are related to forecasting measures and performance of neural networks as proxy for AI compared to traditional forecasting methods, and there are also statistical hypothesis that are indirectly related to the AI field and arise from the selected dataset: namely, the application of Big Data in forecasting (having Google Trend term as a proxy) and other market sentiment variables (that would support the behavioral economists and give a further support to the study of AI, that could surpass human prejudices, fears and other biases). All statistical hypotheses are summarized in Table 10.

Table 10 **Table of statistical hypotheses**

<i>Econometric assumptions (statistical hypotheses)</i>
<p>Ho: Lags of S&P500 returns do not have a predictive power over its future values</p> <p>H1: Lags of S&P500 returns have a predictive power over its future values</p>
<p>Ho: Explanatory variables such as Term spread, Gold return, Credit spread etc. and their lags do not have a predictive power over S&P500 returns</p> <p>H1: Explanatory variables such as Term spread, Gold return, Credit spread etc. and their lags have a predictive power over S&P500 returns</p>
<p>Ho: Lags of explanatory variables that represent market sentiment (e.g. Google Trends data) do not have a predictive power over S&P500 returns</p> <p>H1: Lags of explanatory variables that represent market sentiment (e.g. Google Trends data) have a predictive power over S&P500 returns</p>
<p>Ho: Performance of NAR forecast is no different to the performance of ARIMA forecast</p> <p>H1: Performance of NAR forecast is different to the performance of ARIMA forecast</p>
<p>Ho: Performance of NARX forecast is no different to the performance of VAR forecast</p> <p>H1: Performance of NARX forecast is different to the performance of VAR forecast</p>
<p>Ho: Performance of NARX forecast is no different to the performance of NAR forecast</p> <p>H1: Performance of NARX forecast is different to the performance of NAR forecast</p>

7 EMPIRICAL RESEARCH

This chapter presents the models that have been chosen for each dataset from each model category (univariate vs. multivariate, linear model vs. non-linear neural networks), explaining the rationale for their choice and presenting the model output and goodness of fit, where applicable.

7.1 ARIMA modeling for monthly dataset

First step in building ARMA model is to determine the order of the model to use that would capture best the time varying features of the series. It is performed by graphical analysis of ACF (autocorrelation function) and PACF (partial autocorrelation function), as well as Information criteria (AIC, SBIC and HQIC).

ACF and PACF analysis

























ACF measures linear dependence of a current value and its historical value at time k . PACF –measures the correlation between current value and its value k period ago, after controlling for the correlation of the intermediate lags. At lag 1 ACF and PACF are the same (given no intermediate lags), while for other lags, the values differ. Having ACF significant at further lags (of order p) than the first one is a sign of Autoregressive component of the data, while having PACF significant at further lags (of order q) than the first one is a sign of Moving Average component of the data. If both ACF and PACF are significant at further lags that means ARMA model (of p, q order) combining both characteristics is more appropriate.

Autocorrelation are considered significant if it is outside $\pm 1.96 \times \frac{1}{\sqrt{T}}$, where T is the number of observations (Brooks, 2008). Given $T=585$, the confidence interval is -0.081 to 0.081. This would mean that autocorrelation coefficient is significant for lag 1 and 5, while partial autocorrelation coefficient is significant t lag 1, 5, 6 and 11, which implies that a mixed ARMA model might be appropriate.

Graphical illustration of ACF and PACF for the monthly dataset is depicted in Table 11.

Table 11 ACF and PACF illustration for monthly dataset

This table presents autocorrelation and partial autocorrelation functions for ARIMA model, fitted for the monthly data-set for the period of 1968-2016. Functions highlighted in a bold font represent significant autocorrelations that are beyond the confidence interval.

Autocorrelation	Partial Correlation	Lag	AC	PAC
		1	0,25	0,25
		2	-0,01	-0,07
		3	0,03	0,05
		4	0,05	0,03
		5	0,10	0,09
		6	-0,06	-0,12
		7	-0,06	-0,01
		8	0,03	0,04
		9	-0,02	-0,05
		10	-0,01	0,01
		11	0,06	0,09
		12	0,01	-0,03

Information criteria analysis

The hypothesis of a mixed ARMA model could be further tested using the following information criteria: Akaike's information criterion (AIC), Schwartz's Bayesian information criterion (SBIC) and Hannan-Quinn criterion (HQIC). Detailed information on these criteria and their formulas can be found in Brooks (2014, p.275). The results of the information criteria test on model selection are illustrated in Table 12, where the cells highlighted in a bold font show the models that are the best fit according to the respective information criteria.

AIC Information criterion suggested to use ARMA (5,2) model, while SBIC Information criterion proposed MA(1) model.

Table 12 Summary of information criteria coefficients, monthly dataset

The tables below depict the result of Akaike's information criterion (AIC), Schwartz's Bayesian information criterion (SBIC) used for identifying optimal ARIMA structure for the monthly dataset for the period of 1968-2016. Horizontal line indicates q number of error lags, whereas vertical axis indicates p number of lags of the dependent variable. The cells highlighted in a bold font represent the smallest information criteria, indicating the optimal model.

AIC								
p/q	0	1	2	3	4	5	6	MA
0	5,417	5,355	5,358	5,360	5,363	5,348	5,350	
1	5,362	5,358	5,358	5,360	5,363	5,351	5,351	
2	5,360	5,357	5,361	5,363	5,367	5,345	5,349	
3	5,361	5,363	5,364	5,367	5,357	5,349	5,349	
4	5,363	5,365	5,366	5,352	5,354	5,348	5,351	
5	5,358	5,354	5,345	5,348	5,351	5,353	5,355	
6	5,347	5,350	5,348	5,351	5,352	5,354	5,356	
AR								

SBIC								
p/q	0	1	2	3	4	5	6	MA
0	5,425	5,377	5,388	5,398	5,408	5,400	5,410	
1	5,384	5,388	5,395	5,405	5,415	5,411	5,418	
2	5,390	5,395	5,405	5,416	5,426	5,422	5,424	
3	5,398	5,407	5,416	5,426	5,424	5,434	5,431	
4	5,408	5,418	5,426	5,419	5,429	5,430	5,441	
5	5,410	5,414	5,412	5,423	5,433	5,442	5,452	
6	5,407	5,418	5,423	5,433	5,442	5,452	5,460	
AR								

7.1.1 ARMA (5,2) model

Therefore, the first model this study is going to use for the monthly dataset is ARMA (5,2).

Model description for ARMA

ARMA (5,2) model:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_5 x_{t-5} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

Model output

Table 13 ARMA (5,2) model output, monthly dataset

This table depicts ARMA (5,2) model output for the monthly dataset for the period of 1968-2016.

Variable	Coefficient	Std. error	t-statistic	Prob.
C	0,60	0,22	2,69	0,007
AR(1)	-0,49	0,16	-3,06	0,002
AR(2)	-0,47	0,13	-3,59	0,000
AR(3)	0,15	0,06	2,57	0,010
AR(4)	-0,01	0,05	-0,27	0,786
AR(5)	0,16	0,04	3,90	0,000
MA(1)	0,77	0,15	4,99	0,000
MA(2)	0,60	0,15	3,95	0,000
Adjusted R-squared	0,083	R-squared		0,095
Prob(F-statistic)	0,000	Schwarz criterion		5,41
Durbin-Watson stat	1,998			

Model's goodness of fit can be measured by R-squared. R-squared is the ratio of explained sum squared variations to total sum squared variation. However, the better technique for its measure is to use Adjusted R-square, which is R-squared that is adjusted for the loss in degrees in freedom associated with the introduction of each new explanatory variable. (Brooks, 2014). The higher Adjusted R-square, the better model explains the variation in the variable of interest. ARMA (5,2) has Adjusted R^2 of 0.083, which means that 8.3% of the variation in the S&P500 index return can be explained by the variation its own lags and lags of the error terms.

F-statistics measures joint significance of all variables in the model. It is highly significant, meaning that all explanatory variables should be kept in the model. According to the individual statistical probabilities of each variable, all variables are significant except for the 4th lag of S&P500 return.

From the coefficients, one can observe that first two lags of S&P500 return have a negative association with the current S&P500 return, whereas lag 3 and 5 have a positive association. However, according to Brooks (2014), the coefficients of ARMA models should not be interpreted in the traditional way; instead, the model plausibility should be measured in terms of how well does the model fit the data and model's forecasting accuracy.

7.2 ARIMA modelling for daily dataset

























Similar procedure is repeated for ARIMA modelling for the daily dataset.

ACF and PACF analysis

Starting from analysis ACF and PACF of S&P500 return and using the same formula from Brooks (2014) in determining significant correlations: $\pm 1.96 \times \frac{1}{\sqrt{T}}$, where T is the number of observations, one can determine the confidence interval. Given T=2608 in this data sample, the confidence interval is -0.038 to 0.038. This means that there are significant autocorrelations for the first 2 lags as well as lag 5 and 8, while partial autocorrelation coefficient is significant at first 2 lags and at lag 5, 7 and 8, as depicted in Table 14. This implies that a mixed ARMA model might be appropriate, for which it could be further tested using Information criteria.

Table 14 ACF and PACF, daily dataset

This table presents autocorrelation and partial autocorrelation functions for ARIMA model, fitted for the daily data-set for the period of 2007-2017. Functions highlighted in a bold font represent significant autocorrelations that are beyond the confidence interval.

Autocorrelation	Partial Correlation	Lags	AC	PAC
		1	-0,11	-0,11
		2	-0,05	-0,06
		3	0,02	0,01
		4	-0,01	-0,01
		5	-0,04	-0,04
		6	0,00	-0,01
		7	-0,04	-0,04
		8	0,05	0,04
		9	-0,04	-0,03
		10	0,02	0,02
		11	0,02	0,02
		12	-0,03	-0,02

Information criteria analysis

According to the results from information criteria, AIC was the smallest for ARMA (4,5), while the smallest SIC was for MA (1), suggesting to use the respective structures.

7.2.1 ARMA (4,5) model

Therefore, given ACF and PACF graphical examination and Akaike's information criterion, ARMA (4,5) model has been chosen for the daily dataset.

Model description for ARMA (4,5)

ARMA (p,q) model:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_4 x_{t-4} + \theta_1 \varepsilon_{t-1} + \dots + \theta_5 \varepsilon_{t-5}$$

Model output

ARMA (4,5) estimation output presented below, Table 15.

Table 15 ARMA (4,5) model output, daily dataset

This table depicts ARMA (4,5) model output for the daily dataset for the period of 2007-2017.

Variable	Coefficient	Std. error	t-statistic	Prob.
c	0,02	0,03	0,72	0,469
AR(1)	-0,02	0,20	-0,09	0,930
AR(2)	0,75	0,15	5,06	0,000
AR(3)	-0,40	0,14	-2,85	0,004
AR(4)	-0,70	0,20	-3,53	0,000
MA(1)	-0,10	0,20	-0,48	0,634
MA(2)	-0,80	0,14	-5,57	0,000
MA(3)	0,50	0,14	3,50	0,001
MA(4)	0,69	0,20	3,50	0,001
MA(5)	-0,15	0,04	-3,82	0,000
R-square	0,028			
Adjusted R-squared	0,023	Durbin-Watson	2,00	Prob (F-test) 0,0

Adjusted R-square is 0.023, which means that only 2.3 % of the variation in the S&P500 return can be explained by the variation its own lags and lags of the error terms. That is very low explanatory power and perhaps other explanatory variables should be introduced into the model in order to create a better fitted model for this data.

F-statistics, measuring joint significance of all variables in the model, is highly significant, meaning that all explanatory variables should be kept in the model. Given individual probabilities of each lag, all lags are significant except for the 1st lag of S&P500 return and 1st lag of the residual component.

7.3 NAR network

Moving onto the proxy for the artificial intelligence in univariate forecasting, this section describes which non-linear auto-associative time-series network (NAR) have been chosen for both dataset.

7.3.1 NAR network for monthly data set

Given that network performance varies from each training and initiation, the network had to be retrained and re-run numerous times in order to get statically significant results. One such network is presented below and will be explained in detail.

7.3.1.1 NAR network example 1

The main NAR network type for this data set was NAR with 2 lags, 1 hidden layers with 10 hidden neurons in it. The network was trained with Levenberg-Marquardt (LM) learning algorithm.

One such network had been trained with 11 epoch iterations and 6 validation checks. The performance of this network was measured in MSE and was 12.4. (RMSE is 3.5). The graphical illustration of this stage of network training is presented in Figure 18.

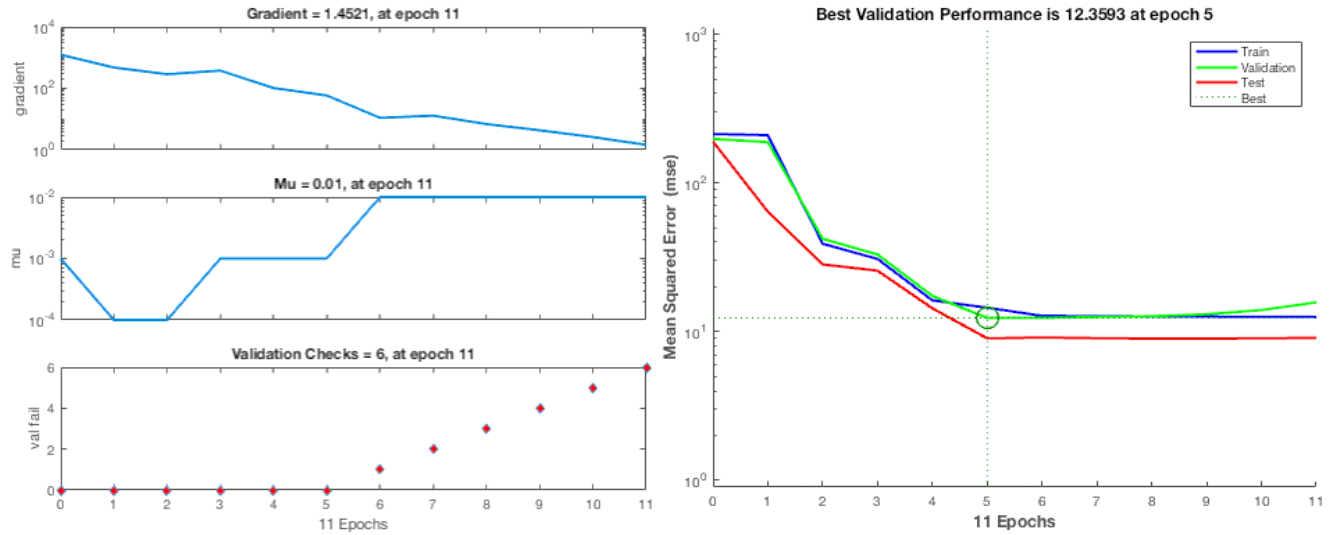


Figure 18 NAR training and validation graph

The explanatory power of the network, measured by the correlation coefficient (R) between the output and input is 20% for test observations. This is moderately high coefficient, thus giving a reason to suspect that network was able to capture some non-linear characteristics of the data. R coefficient is presented below in Figure 19.

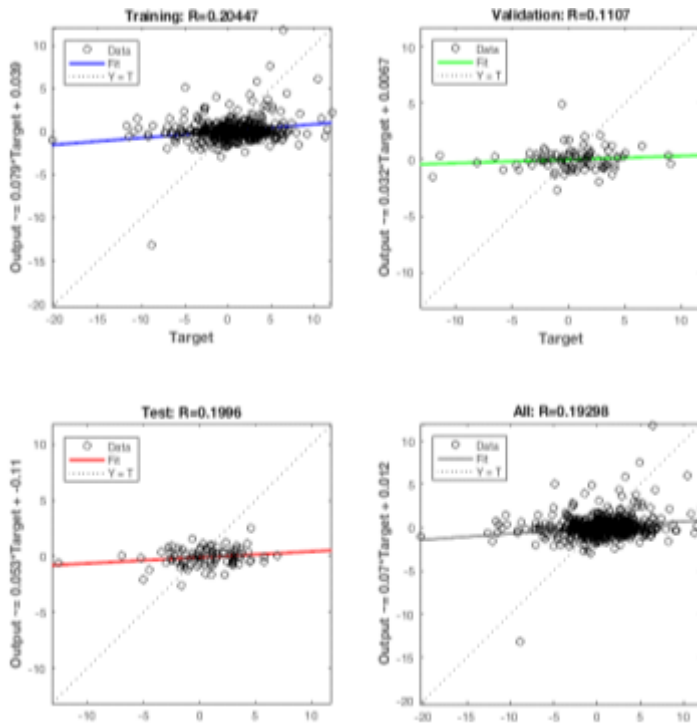


Figure 19 Correlation coefficient of neural network

7.3.1.2 NAR network example 2

Another network that will be demonstrated differs in some parameters from the previous network. Keeping 1 hidden layer with 10 hidden neurons in it, the number of lags have been increased to 12 lags. Furthermore, Bayesian Regularization learning algorithm was used for its training.

The training has been achieved after 886 iterations, which is much larger than before, and took much longer time to process. (that is due to the different learning algorithm, as it has been stated before, BR learning algorithm takes much longer time to train). Training and validation performance can be observed on Figure 20.

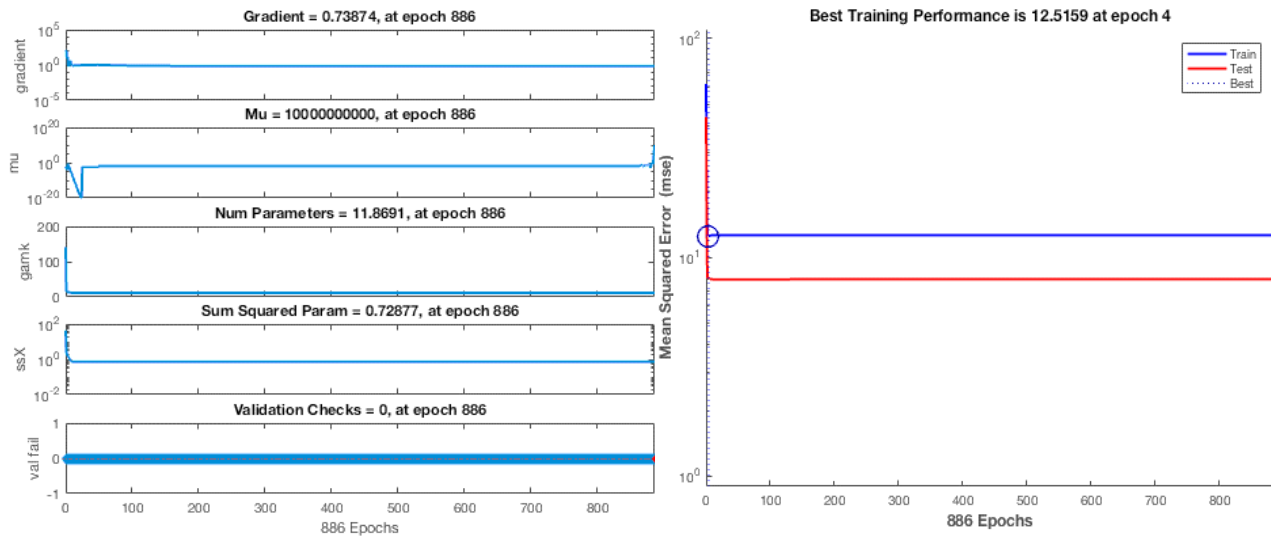


Figure 20 Training and validation summary NAR 12 lags monthly

The explanatory power of this network, measured by R is higher than that of the previous network and can be examined on Figure 21. This is due to the increased number of lags and different learning algorithm that fitted the data better.

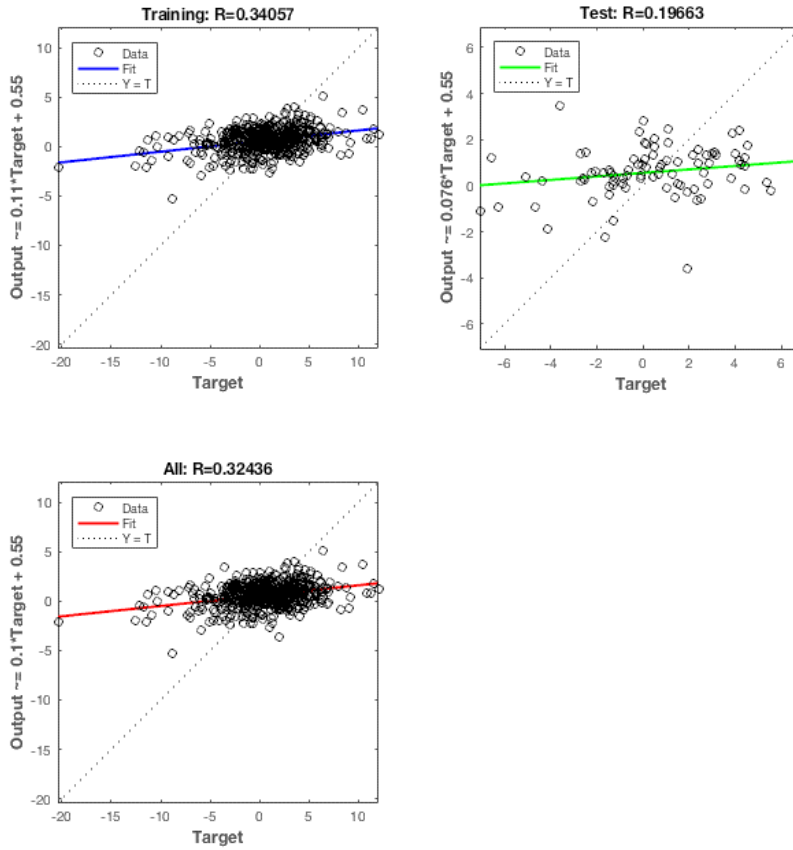


Figure 21 Correlation coefficient NAR 12 lags monthly dataset

7.3.2 NAR network for daily data set

The process is repeated for the networks for daily dataset.

7.3.2.1 NAR network example 1

The first network that is presented is a neural net that has two lags of the S&P500 daily return as the input variables. The training is with Levenberg-Marquardt learning algorithm has been accomplished after 11 iterations and 6 validation checks as depicted on Figure 22. The performance for the validation set, measured by MSE, was 1.67, while for the test observations it was reduced to 1.52.

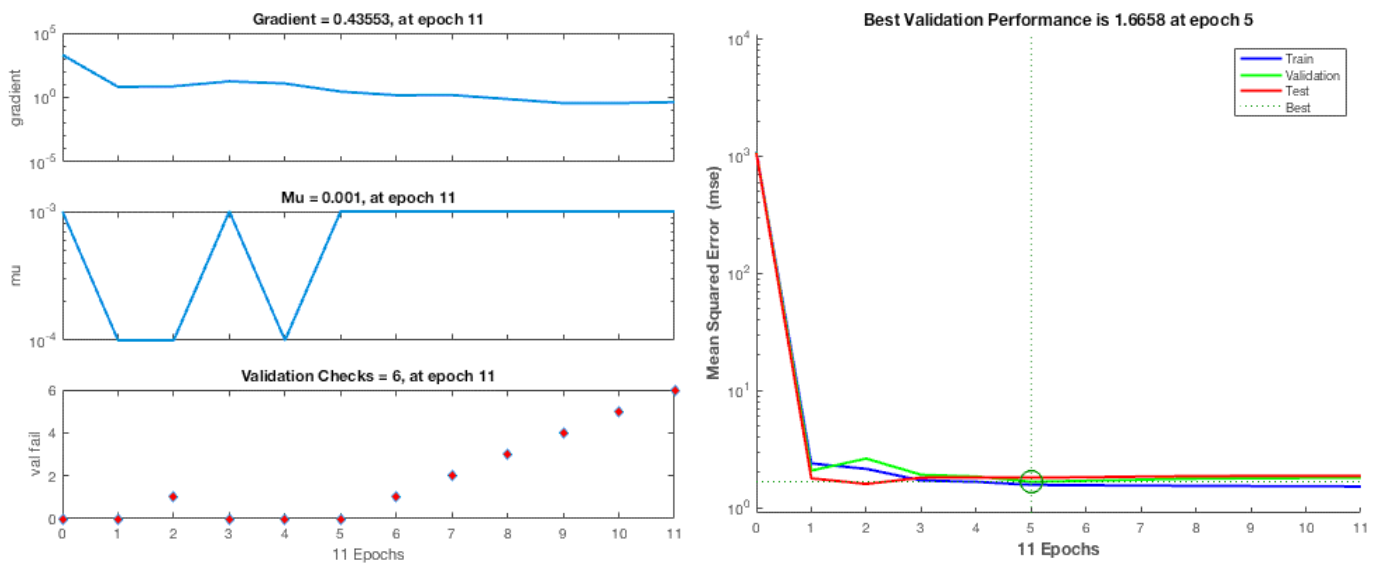


Figure 22 Training and validation NAR 2 lags daily dataset

The correlation coefficient (R) between output and input of this network was 19.4% (depicted on Figure 23).

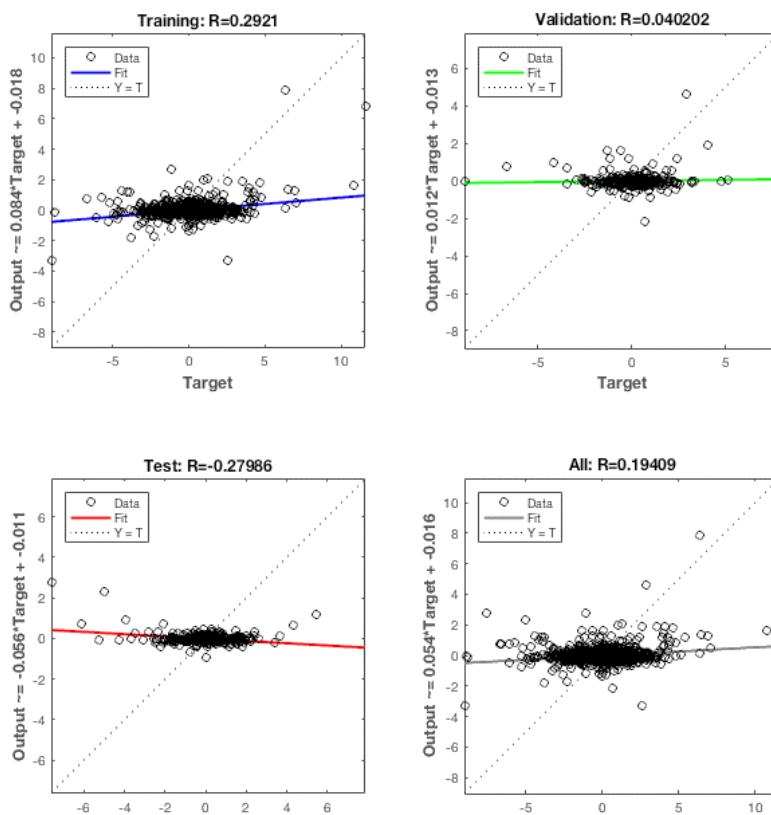


Figure 23 Correlation coefficient of NAR 2 lags daily dataset

7.4 VAR model

This section describes linear multivariate model for modelling S&P500 returns – VAR. It will first describe VAR models for monthly dataset, followed by the description of VAR models for the daily dataset.

7.4.1 VAR modelling for monthly data set

7.4.1.1 Lag selection

Lag selection is performed using cross-equation restrictions and information criteria. Information criteria do not require errors to be normally distributed, and they function as a trade-off between an increase in RSS with each lag and a decrease in degrees of freedom. Formulas for the multivariate versions of information criteria are the following (where p is the number of lags in the model) :

$$AIC_{(p)} = \ln|\tilde{\Sigma}(p)| + \frac{2}{T}pn^2$$

$$BIC_{(p)} = \ln|\tilde{\Sigma}(p)| + \frac{\ln T}{T}pn^2$$

$$HQ_{(p)} = \ln|\tilde{\Sigma}(p)| + \frac{2\ln \ln T}{T}pn^2$$

Table 16 Lag length selection using information criteria, monthly dataset

This table depicts the result of Akaike's information criterion (AIC), Schwartz's Bayesian information criterion (SBIC) and Hannan-Quinn criterion (HQIC) used for identifying optimal lag structure for VAR model for the monthly dataset for the period of 1968-2016. The cells highlighted in a bold font represent the smallest information criteria, indicating the optimal model.

Lag	AIC	SBIC	HQ
0	51,1	51,2	51,1
1	29,1	30,3*	29,5
2	28,6*	30,9	29,5*
3	28,7	32,0	30,0
4	28,8	33,2	30,5
5	29,0	34,5	31,1
6	29,0	35,7	31,6
7	29,2	36,9	32,2
8	29,3	38,2	32,8
9	29,4	39,4	33,3
10	29,4	40,5	33,7
11	29,4	41,5	34,1
12	28,9	42,1	34,0

Given the fact that both AIC and HQ suggested VAR (2) model, it will be therefore used in this study.

7.4.1.2 Lag exclusion test

Another test that could be performed in order to check the fitness of a chosen VAR model is a test on lags exclusion (presented in Table 17). The statistics of the test suggest that both lags should be kept in the model.

Table 17 Lag exclusion test, VAR (2), monthly dataset

This table portrays the results of the lag exclusion test on the VAR (2) model for the monthly dataset over the period of 1968-2016. Statistical probabilities are shown in brackets.

	S&P500 return	Joint
Lag 1	71,0 [0,000]	5811,5 [0,000]
Lag 2	32,0 [0,004]	645,1 [0,000]
df	14	196

7.4.1.3 VAR (2) monthly dataset

Therefore, an estimated VAR model has 2 lags and 12 variables: S&P500 return, term spread, credit spread, oil price change, gold price change, money supply change, CPI (nat. logarithm), unemployment rate, market sentiment, dividend yield change and relative changes in exchange rates of GBP and JPY to US dollar.

The full model output is too large to be presented here, hence, only the regression for the variable of interest- S&P 500 return is presented in Table 18.

Adjusted R-square, measuring how much of the variation in the dependent variable is explained by the variation in independent variables, is 0.087. This means that the fitted model explains 8.7% of the variation in S&P500 index returns. Explanatory power of VAR (2) model is higher than that of ARMA model, meaning that addition of other explanatory variables in the model help explaining the movements in the S&P500 return.

F-statistics, measuring the joint significance of the coefficients is statistically significant, therefore, suggesting that VAR (2) is an appropriate model for the data.

Table 18 **VAR (2) model output, monthly dataset**

This table depicts VAR (2) model output for the monthly dataset for the period of 1968-2016, using 12 variables: S&P500 return (R), term spread, credit spread, gold, oil, money supply, unemployment, market sentiment, dividend yield, GBP, and JPY exchange rates, and CPI (all with two lags), totalling to the 24 parameter estimations.

Variable	Coefficient	Standard errors	T-statistics
R(-1)	0,09	0,32	0,27
R(-2)	0,04	0,32	0,11
Term(-1)	0,79	0,41	1,91
Term(-2)	-0,88	0,41	-2,15
Credit(-1)	2,64	1,18	2,25
Credit(-2)	-3,10	1,18	-2,63
Gold(-1)	-0,02	0,03	-0,82
Gold(-2)	0,01	0,03	0,52
Oil(-1)	-0,01	0,02	-0,61
Oil(-2)	0,04	0,02	2,02
Money S.(-1)	0,02	0,10	0,20
Money S.(-2)	0,05	0,10	0,54
Unempl.(-1)	2,36	0,96	2,47
Unempl.(-2)	-1,99	0,93	-2,13
Market Sent.(-1)	0,02	0,04	0,53
Market Sent.(-2)	0,00	0,04	0,11
DY(-1)	-0,18	0,31	-0,56
DY(-2)	0,09	0,31	0,29
GBP(-1)	0,03	0,06	0,44
GBP(-2)	-0,02	0,06	-0,29
JPY(-1)	-0,01	0,05	-0,19
JPY(-2)	-0,04	0,05	-0,83
CPI(-1)	-119,76	56,99	-2,10
CPI(-2)	119,77	56,94	2,10
C	-3,22	2,46	-1,31
R-squared	0,127	Prob.F test	0,000
Adj. R-squared	0,087		

From the output presented in Table 18, one can see that both lags of Term spread, Credit Spread and Unemployment variables are significant at explaining S&P500 return. However, it is not possible to go further into interpretation of the relationship with just the model

output, given the fact that the coefficients are given in vectors and for multiple lags, thus making it very hard to infer (Brooks, 2014). If one wants to test for the relationship between the variables, he/she should utilize impulse response tests, which measure the effect of all specified lags of one variable on another, thus allowing for a more detailed interpretation.

7.4.2 VAR modelling for the daily dataset

Similar approach is applied in building VAR model for the daily data set.

7.4.2.1 Lag selection

In order to identify the optimal number of lags necessary for the model, information criteria (AIC, SBIC and HQIC) are used. Different information criteria choose different number of lags: AIC chose 6, SBIC 1 and HQ 2. (presented in Table 19 below). It is important to note that information criteria choose different models given their different approach in evaluation of an increased explanatory power at the cost of the loss in degrees of freedom. AIC often chooses one among the largest models, while Schwartz's Bayesian information criteria normally chooses the smallest one.

Table 19 Lag length selection using information criteria, daily dataset

This table shows the result of Akaike's information criterion (AIC), Schwartz's Bayesian information criterion (SBIC) and Hannan-Quinn criterion (HQIC) used for identifying an optimal lag structure for VAR model for the daily dataset for the period of 2007-2017. The cells highlighted in a bold font represent the smallest information criteria, indicating the optimal model.

Lag	AIC	SC	HQ
0	71,8	71,9	71,8
1	54,3	55,2*	54,6
2	53,9	55,5	54,47*
3	53,7	56,2	54,6
4	53,6	56,9	54,8
5	53,5	57,7	55,0
6	53,3*	58,3	55,2
7	53,5	59,2	55,6
8	53,5	60,1	55,9
9	53,6	61,0	56,3
10	53,6	61,9	56,6

7.4.2.2 Lag exclusion test

Lag exclusion test assesses whether the lags in the model are statistically significant. The test performed on this VAR (2) model suggests that both lags are statistically significant, separately and jointly, and do not need to be excluded. Evident in Table 20.

Table 20 Lag exclusion test, VAR (2), daily

This table depicts the result from the lag exclusion test on the VAR (2) model for the lags of S&P500 return and jointly, for the daily dataset over the period of 2007-2017. Statistical probabilities are shown in brackets.

	S&P500 return	Joint
Lag 1	71.63953 [0.000]	22832.39 [0.000]
Lag 2	35.31938 [0.0127]	1994.036 [0.000]
df	19	361

Due to the results provided by HQIC and lag exclusion test, VAR (2) model have been chosen. Therefore, VAR model was estimated with 2 lags and 11 variables: S&P500 return, term spread, credit spread, oil price change, gold price change, changes in exchange rate of GBP and JPY to US dollar, FTSE All share index return at time t-1, Microsoft return at time t-1, change in VIX index at t-1, and Google trends term “S&P500 index”.

7.4.2.3 VAR (2) daily dataset

The full model output is too large to be presented here hence only the regression for the variable of interest- S&P 500 return is depicted in Table 21.

Adjusted R-square for the S&P500 index return equation in VAR model is 0.023, which means that the fitted model explains 2.3% of the variation in S&P500 index returns. This is significantly lower explanatory power than that in the monthly dataset. However, when comparing the explanatory power of this VAR model with its univariate counterparts for this dataset, it becomes apparent that VAR model has slightly higher Adjusted R-square, meaning that addition of the other variables helps explaining the movements in the S&P500 return.

F-test on joint significance of the coefficients is highly significant, meaning that the model is adequate and it is better to keep the variables in the model.

Table 21 VAR (2) model output, daily dataset

This table presents VAR (2) model output for the daily dataset for the period of 2007-2017, using 10 explanatory variables, such as term spread, credit spread, change in gold price, change in oil price, GBP, and JPY exchange rates, FTSE All-Share Index return, Microsoft stock return, VIX index return, and Google Trends term “S&P500 Index”, all of which with two lags, in addition to the 2 lags of the dependent variable (S&P500 return), totalling to the 22 parameter estimations.

	R	Standard errors	T-statistics	
R(-1)	-0,11	-0,04	-4,70	***
R(-2)	-0,02	-0,05	-0,31	
Term(-1)	0,98	-0,43	2,60	***
Term(-2)	-0,95	-0,43	-2,55	***
Credit(-1)	1,30	-1,05	1,47	
Credit(-2)	-1,23	-1,04	-1,40	
Gold(-1)	0,03	-0,03	1,08	
Gold(-2)	0,07	-0,03	1,67	
Oil(-1)	-0,03	-0,02	-1,22	
Oil(-2)	0,00	-0,02	0,26	
GBP(-1)	-0,10	-0,06	-1,71	
GBP(-2)	0,01	-0,06	0,14	
JPY(-1)	0,03	-0,05	0,63	
JPY(-2)	-0,05	-0,06	-1,00	
FTSEALL_1(-1)	-0,09	-0,04	-2,29	***
FTSEALL_1(-2)	-0,03	-0,03	-0,93	
MSFT_1(-1)	0,00	-0,03	-0,10	
MSFT_1(-2)	0,05	-0,02	2,27	***
VIX(-1)	0,01	-0,01	0,82	
VIX(-2)	-0,01	-0,01	-0,84	
GT_index(-1)	-0,01	-0,02	-0,38	
GT_index(-2)	0,02	-0,02	1,87	*
C	-0,18	-0,21	-0,86	
R-squared	0,032	Prob. F-test	0,000	
Adj. R-squared	0,023			

It becomes evident that its own first lag (R-1), Term spread (both lags), Gold (2nd lag), GBP (1st lag), FTSE index, MSFT index and Google trend search for “S&P500 index” all have been found to be significant at explaining S&P500 return. Further examination of the relationship between variables (especially the effect of Google Trends variable on S&P500 return is of interest) is performed using the impulse response test, which measures the effect of all specified lags on the variables, thus allowing for a more detailed interpretation and is presented in the results section.

7.4.2.4 VAR (6) model

Additionally, due to the indication of AIC criterion to use 6 lags in VAR model, it will be tested for as well. An estimated VAR model has 6 lags and contains 11 variables: S&P500 return (treated as an endogenous variable), term spread, credit spread, oil price change, gold price change, changes in exchange rate of GBP and JPY to US dollar, FTSE All share index return at time $t-1$, Microsoft return at time $t-1$, change in VIX index at time $t-1$, and Google trends term “S&P500 index”.

Adjusted R-square for the S&P500 index return equation in VAR model is 0.054, which means that the fitted model explains 5.4% of the variation in S&P500 index returns. When comparing the explanatory power of this VAR model with its univariate counterparts for this dataset, it becomes apparent that VAR model has higher Adjusted R-square, meaning that an addition of the other variables helps explaining the movements in the S&P500 return. It is also possible that having 6 lags in the model allows for a better capture of the dynamic component of the data. For example, GT_index variable was also highly significant at lag 6, implying that there is some time series dynamics beyond first two lags that are stored in the data.

7.5 NARX network

As a proxy for artificial intelligence multivariate forecasting methods, this study uses NARX (non-linear autoregressive network with exogenous inputs).

7.5.1 NARX network for monthly dataset

7.5.1.1 NARX network example 1

First NARX network demonstrated has 2 lags, and has been trained with LM training.

Model description

$$y(t) = f(\mathbf{x}(t-1), \mathbf{x}(t-2), y(t-1), y(t-2))$$

where $y(t)$ is S&P500 return, that is regressed over the previous values of itself as well as the previous values of independent variables, represented in a matrix of \mathbf{X} , which consists of 11 variables: term and credit spreads, changes in oil and gold prices, change in money supply and dividend yield, CPI, unemployment rate, market sentiment, and changes in exchange rates of GBP and JPY to US dollar over the time period of 01/1968-09/2016.

Testing and validation

The training process stopped when the network reached 6 validations. It completed by then 9 iterations and achieved 9.3 MSE. Its training performance is depicted on Figure 24.

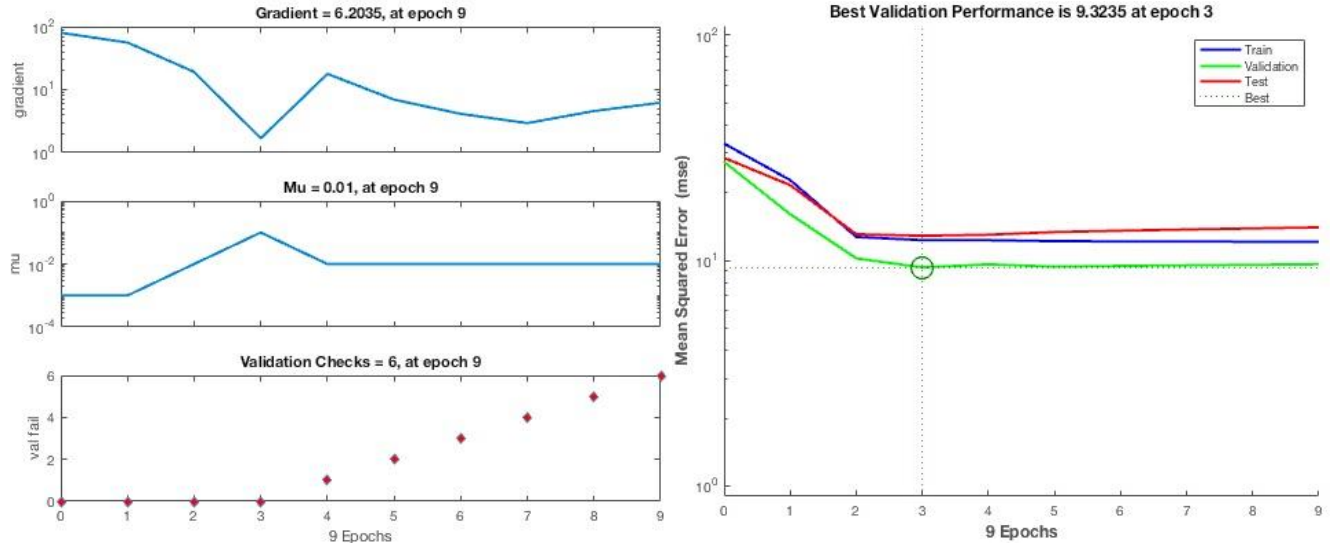


Figure 24 Training and validation image of NARX 2 lags monthly

Coefficient correlation of this network for the test subsample is 15.5% and 30.2% for the whole sample, as depicted on Figure 25.

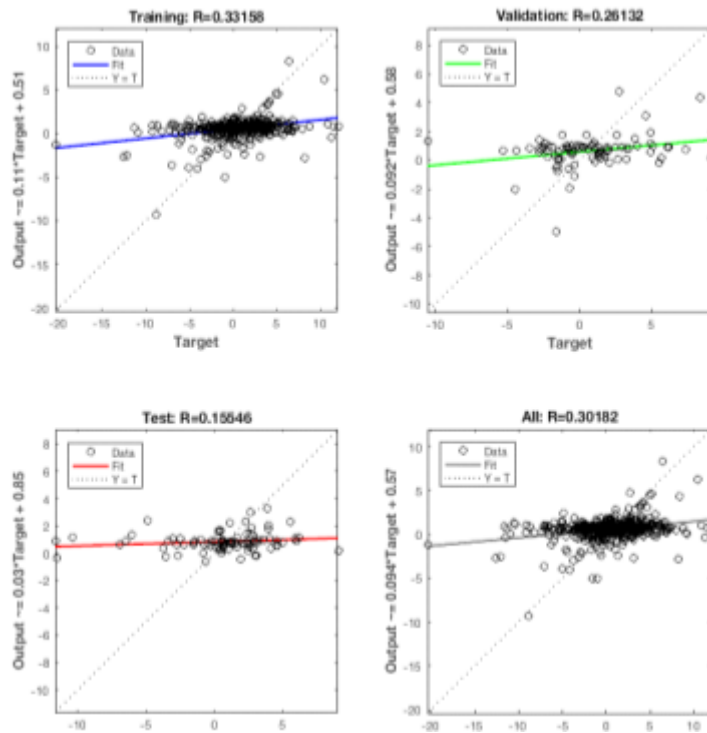


Figure 25 Explanatory power NARX 2 lags monthly

7.5.2 NARX network modelling for the daily dataset

NARX network for the daily dataset is presented below. The network configurations were two lags and LM training methods.

Model description

$$y(t) = f(\mathbf{x}(t-1), \mathbf{x}(t-2), y(t-1), y(t-2))$$

where $y(t)$ is S&P500 return, that is regressed over the previous values of itself $y(t-n)$ as well as the previous values of independent variables, represented in a matrix of \mathbf{X} .

Matrix \mathbf{X} for the daily dataset consists of term spread, credit spread, oil price change, gold price change, changes in exchange rate of GBP and JPY to US dollar, FTSE All share index return at time $t-1$, Microsoft return at time $t-1$, change in VIX index at time $t-1$, and Google trends term “S&P500 index” over the time period of 2004-2017.

Testing and validation

Training and validation process of the network is depicted on Figure 26. There were 6 validation checks, with best performance for the validation sample being 2.38 MSE at epoch iteration 11.

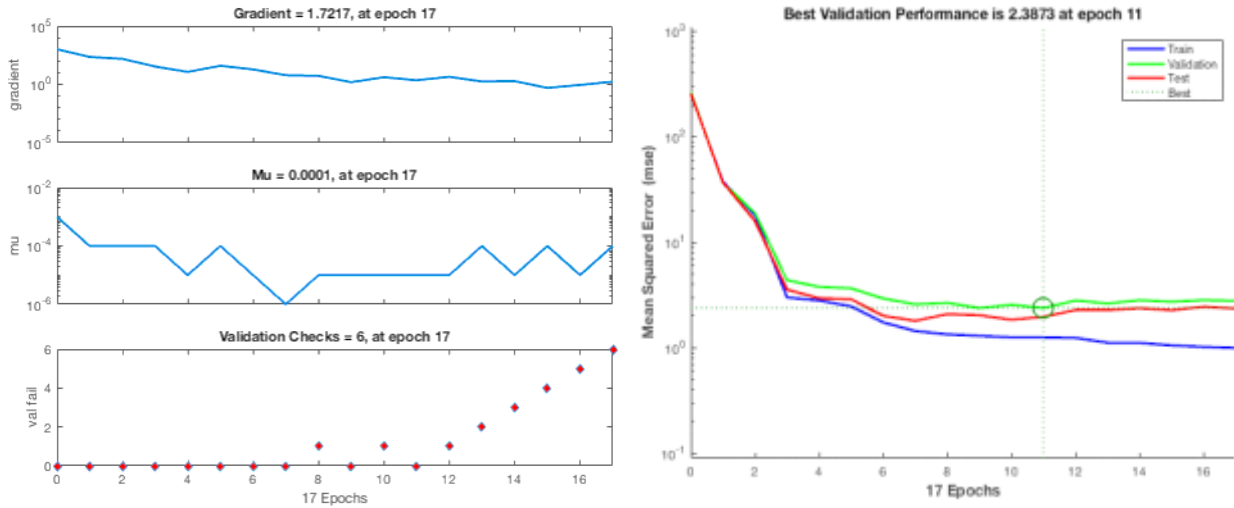


Figure 26 Training and validation illustration NARX 2 lags daily

Correlation coefficient

The correlation between output and input in the testing subsample was low (8.1%), whereas for the training data it was 44%. According to MATLAB (2016) that occurs when there is an overfitting problem, when the network learned very well during the training set and fitted the model very close, however, it is lacking the capability of adjusting to the new observations. It is observed on Figure 27 below.

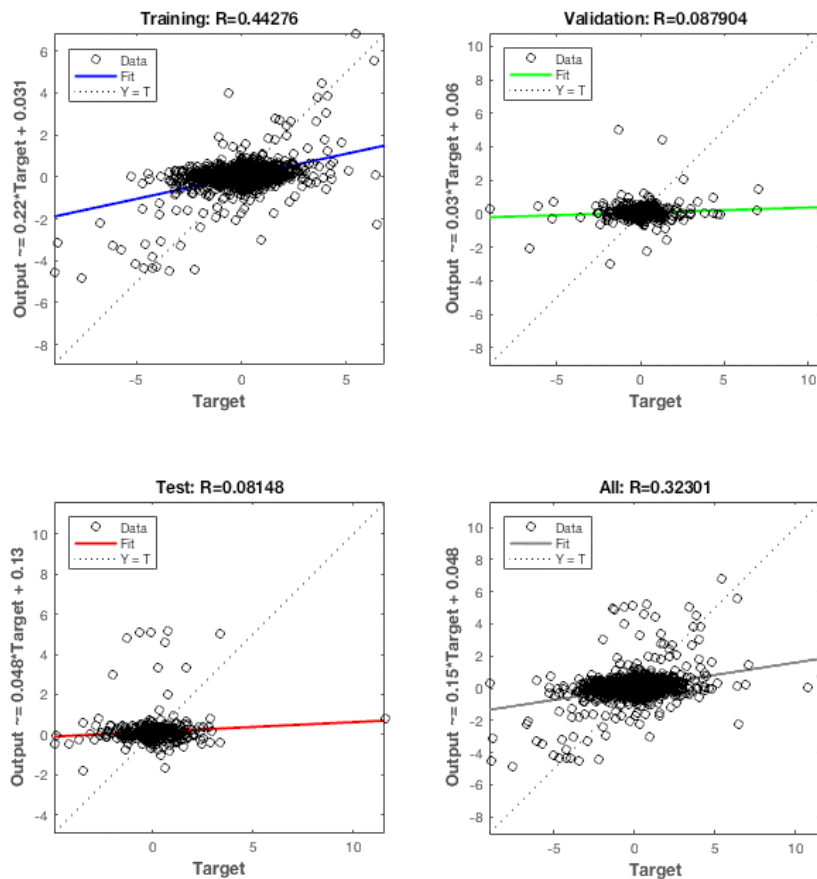


Figure 27 Explanatory power NARX 2 daily dataset

8 MODEL DIAGNOSTICS

This chapter depicts model diagnostics for the models and networks described in the empirical research part. This chapter is structured in a way that it, firstly, summarizes main model diagnostics results in the first subchapter, and then presents model diagnostics details in the following sections. This allows an impatient reader upon finishing the summary subchapter to move onto the results chapter, and the scrupulous reader to continue reading the model diagnostics description.

8.1 Model diagnostics summary

This subchapter presents the summary of the model diagnostics for the models and networks described in the empirical research part.

Stationary of ARMA and VAR

The assumptions of the ARMA and VAR models imply that in order for the model to be stationary, the inverse roots of AR/MA polynomials should be less than one in absolute value and, hence, to lie inside the unit circle.

The test on ARMA and VAR stationarity has concluded that all roots lie inside the unit circle; hence the models are stationary and invertible.

Autocorrelation

Model diagnostics for ARMA model consists of examining the residuals. If the fitted model is appropriate for the data, the residuals should be distributed similarly as white noises processes (0,1), or iid (0,1). This could be observed from the plot of the residuals series and residuals' ACF (which should not contain autocorrelations that exceed the boundaries). There were no significant autocorrelations between the residuals in ARMA models for both daily and monthly dataset hence, there is no need to reject the fitted models.

For VAR models Breusch–Godfrey LM serial correlation and Portmanteau test for autocorrelation were used. Null hypothesis of no serial correlation VAR (2) monthly model was rejected at lags 1, 2, 6 and 7, while the null of the Portmanteau autocorrelation test was rejected for all lags. Also, VAR (2) and VAR (6) model for daily dataset rejected the null of no autocorrelation (or serial correlation) in the residuals. This means that some model re-specification should be made to account for time-series dynamics left in the data. Tests for autocorrelations for residuals in VAR models become insignificant when the lag order is

increased to 30, however, increasing lag order by that much introduces new problems into the models, therefore, the lag order of 2 and 6 has been kept in accordance with the results suggested by information criteria. In order to correctly evaluate the significance of the variables, robust standard errors were used.

Normality test

From Jarque-Bera tests performed on the residuals from the models, one can conclude that the residuals are not normally distributed. Typically, the skewness was minimal, however, there was always an excess kurtosis, indicating that the series have fat tails (outliers). This is an expected result for financial time series (especially for the time-series that include the time period of a financial crisis).

Nevertheless, given the fact that both datasets contain large enough number of observations (585 observations for the monthly dataset and 2608 observations for the daily dataset), central limit theorem could be applied, stating that t-statistics will asymptotically follow the normal distribution, hence should not intervene with the interpretation of the coefficients.

Heteroscedasticity

In order to check for the existence of possible heteroscedasticity problem, White heteroscedasticity test has been used.

It concluded that no heteroscedasticity was detected in ARMA models for monthly and daily datasets, since all p-values were significantly higher than 0.05. Therefore, the models seem appropriate and valid for the data set.

Using White test for VAR models did not reject the null hypothesis of homoscedasticity for VAR monthly model (except lag 1,2, and 7) and rejected the null hypothesis of homoscedasticity for VAR (2) and VAR (6) for daily dataset, denoting that there is still some heteroscedasticity in the data left unaccounted for. In order to correctly evaluate the significance of the variables, robust standard errors were used.

Neural Network diagnostics

As have been previously described, one of the short-comings of the neural networks is the lack of network diagnostics tests available, thus limiting a traditional structural evaluation of model validity. However, one of the reasons for its absences is that there are no

assumptions required for the network, regarding the distribution of the variables or stationarity, consequently, making network diagnostics redundant. Nevertheless, the few network diagnostics that are available for the networks are autocorrelation of the residuals, which occasionally found to be an autocorrelation in the residuals when only two lags are used, and autocorrelation is not more detected when the lag order is increased to 6 or 12. Normality test also depicted not normally distributed residuals, but similarly to the traditional models, one can apply central limit theory and continue with the research. Test for correlation between the input and the respective error did not flag any diagnostics problem in the network, therefore, all networks are continued to be used in the research.

Overall:

- ARIMA models for monthly and daily dataset are adequate and can be used for forecasting
- VAR models for monthly dataset are stable and adequate according to in-sample measurements of model adequacy and can be used for forecasting
- VAR models for daily dataset exhibits model diagnostics problems, such as autocorrelation in the residuals, and heteroscedasticity. In forecasting, the fit of the model and its validity is often judged by out-of-sample measures, such as MSPE and MAPE, thus poor in-sample model can still provide a good out-of-sample results. Therefore, these models will continue to be in the research and their adequacy will be checked against its forecasting accuracy.
- Artificial Neural Networks (given no necessary assumptions for network, its stationarity or distribution) are continued to be used in the research
- To conclude, all described models will be tested for forecasting accuracy

This is the end of the summary section. The following subchapters will present the results of the model diagnostics test for each model/network used.

8.2 Model diagnostics ARMA (5,2)

This subchapter presents the model diagnostics results for ARMA (5,2) model for the monthly dataset. Model diagnostics are presented for the full model (using all of the sample observations), as the tests results did not differ from the ones run on the estimation sample (without including the observations used for forecasting out-of-sample).

8.2.1 Model stationarity

The assumptions of the ARMA model imply that in order for the model to be stationary, the inverse roots of AR/MA polynomials should be less than one in absolute value and, hence, to lie inside the unit circle. The test on ARMA stationarity has concluded that there all roots lie inside the unit circle; hence the model is stationary and invertible.

8.2.2 Residual diagnostics

Model diagnostics for ARMA model consists of examining the residuals. If the fitted model is appropriate for the data, the residuals should be distributed similarly as white noises processes (0,1), or iid (0,1). This could be observed by analyzing autocorrelation and partial autocorrelations functions of the residuals, (which should not contain any significant autocorrelations, as all autocorrelations should have been accounted for by a model).

Table 22 ACF and PACF of the residuals of ARMA (5,2), monthly dataset

This table depicts autocorrelation and partial autocorrelation functions of the residuals series from ARMA (5,2) model, fitted for the monthly data-set for the 1968-2016 period.

Autocorrelation	Partial Correlation	Lag	AC	PAC
		1	0,00	0,00
		2	0,00	0,00
		3	0,00	0,00
		4	0,00	0,00
		5	-0,01	-0,01
		6	-0,02	-0,02
		7	-0,02	-0,02
		8	-0,03	-0,03
		9	0,01	0,01
		10	-0,01	-0,01
		11	0,05	0,05
		12	0,02	0,02

From the observation of ACF and PACF depicted in Table 22, it is evident that there are no significant autocorrelations left in residuals, hence there is no need to reject the fitted model.

Additionally, one can examine the model fit by plotting the residuals (Figure 28), that should resemble white noises process if the model is appropriate. The graph of plotted residuals resembles the WN (0,1), which further supports that ARMA (5,2) model is appropriate.

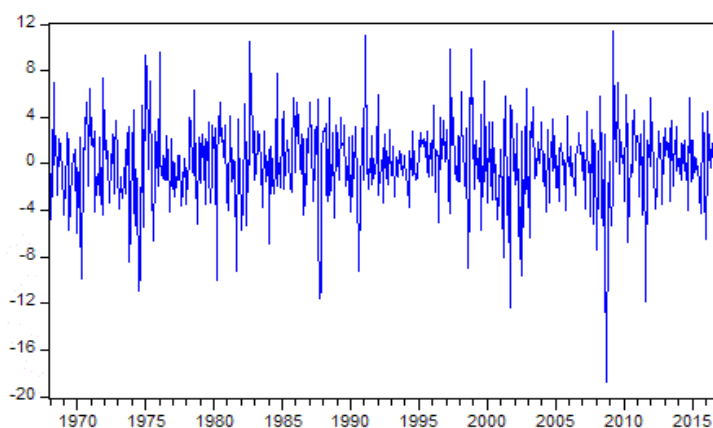


Figure 28 Model diagnostics ARMA (5,2)- plot of residuals

8.2.3 Normality test

From Jarque-Bera test performed on the residuals from the model, one can conclude that the residuals are not normally distributed. While skewness is minimal (-0,5), the kurtosis is 5.6, indicating that the series have fat tails. Indeed, one can see few fat tails on residuals' histogram, depicted on Figure 29. Given the fact that the dataset contains 585 observations, central limit theorem could be applied, stating that t-statistics will asymptotically follow the normal distribution, hence should not intervene with the interpretation of the coefficients. (Brooks,2014).

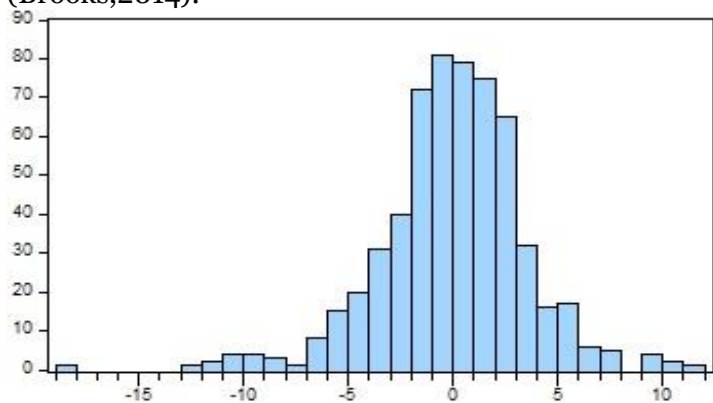


Figure 29 Distribution of the residuals

Mean	0,0
Maximum	11,4
Minimum	-18,8
Std. Dev.	3,5
Skewness	-0,5
Kurtosis	5,6
Jarque-Bera	190
Probability	0,00

8.2.4 Heteroscedasticity test

White heteroscedasticity test did not detect any heteroscedasticity, since all p-values were significantly higher than 0.05.

8.3 Model diagnostics ARMA (4,5)

Next, this subchapter will present model diagnostics for the ARMA (4,5) model that was used for the daily dataset.

8.3.1 Model stationarity

























Test on stationarity of VAR model determined that all inverse roots of AR/MA polynomials lie inside the unit circle, hence, ARMA (4,5) model is stationary and invertible.

8.3.2 Residual diagnostics

ACF and PACF graph of the residuals depict that there is no correlation left in the residuals, as all coefficients are below the critical value of significance. Therefore, the model seems appropriate.

Table 23 ACF and PACF of residuals of ARMA (4,5), daily dataset

This table presents autocorrelation and partial autocorrelation functions of the residual series from ARMA (4,5) model, fitted for the daily data-set for the period of 2007-2017.

Autocorrelation	Partial Correlation	Lag	ACF	PACF
		1	0,00	0,00
		2	0,00	0,00
		3	0,00	0,00
		4	-0,01	-0,01
		5	-0,01	-0,01
		6	-0,02	-0,02
		7	0,00	0,00
		8	0,01	0,01
		9	-0,02	-0,03
		10	-0,01	-0,01
		11	0,01	0,01
		12	-0,02	-0,02

8.3.3 Normality test

Jarque-Bera test of normality on the residuals rejected the null hypothesis of normal distribution of the residual. Residuals are not normally distributed, as the test is highly significant. It appears that while there is only minimal skewness of -0.25, there is an excess kurtosis. The histogram of the distribution is depicted on Figure 30.

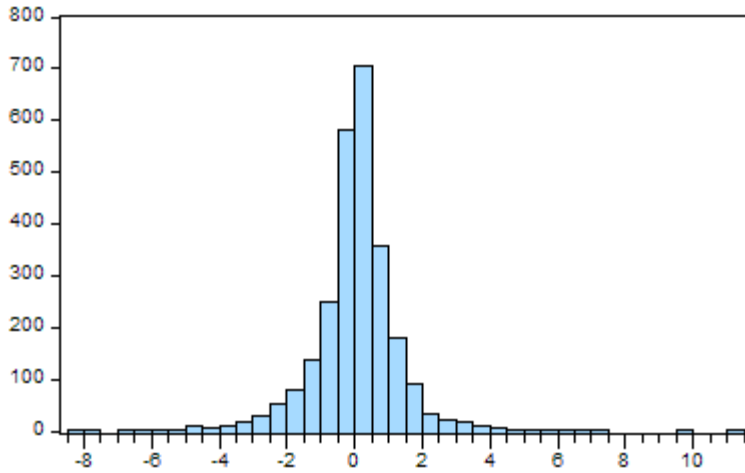


Figure 30 The distribution of residuals from ARMA (4,5) daily dataset

The dataset consists of 2608 observations, which is a large enough number to apply central limit theorem and conclude that even though the residuals are not distributed normally, t-statistics will asymptotically follow the normal distribution, hence should not intervene with the interpretation of the coefficients. (Brooks,2014).

8.3.4 Heteroscedasticity test

In order to check for the existence of possible heteroscedasticity problem, the White heteroscedasticity test has been used. It concluded that no heteroscedasticity was detected, since all p-values were significantly higher than 0.05. Therefore, the model seems appropriate and valid for the data set.

8.4 Network diagnostics NAR (2), LM training, monthly dataset

Moving to univariate network for the monthly dataset, the first network that is discussed NAR net with 2 lags and LM training.

As have been described before, one of the short-comings of the neural networks is the lack of network diagnostics tests available, thus limiting a traditional structural evaluation of model validity. However, one of the reasons for its absences is that there are no assumptions

required for the network, regarding the distribution of the variables or stationarity, consequently, making network diagnostics redundant. Nevertheless, this section will present few network diagnostics that are available.

8.4.1 Test for autocorrelation of the residuals

The first check that could be performed on the NAR network with 2 lags and LM training algorithm is the test for the autocorrelation of the residuals. Similarly, to the liner models, the NAR network is considered a good fit if there are not autocorrelations left in the residuals. On the Figure 31 one can observe that there still some autocorrelations left up to lag 5. This suggests that the higher order lag should be chosen to capture more dynamics of the data and therefore improve the network fit.

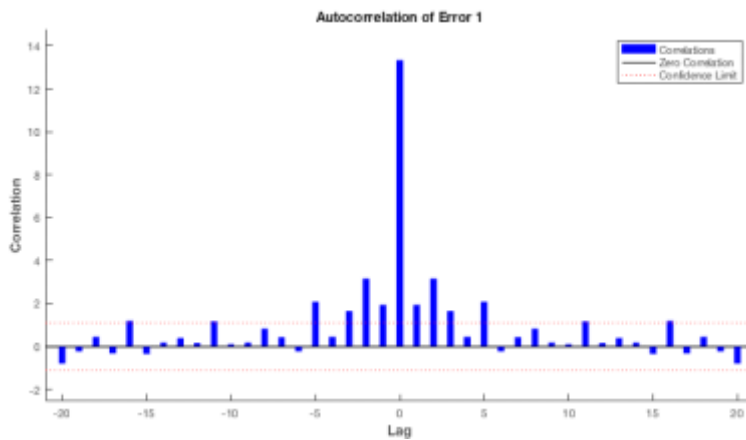


Figure 31 Autocorrelation of the residuals

8.4.2 Normality test

Normality test of the residuals from the network modelling can also be performed. Error histogram is depicted on Figure 32. It is evident that there is a slight skewness, thus indicating that the errors are not non-normal distribution. However as have been discussed previously, given large number of the observations, not-normal distribution of the residuals does not represent a problem.

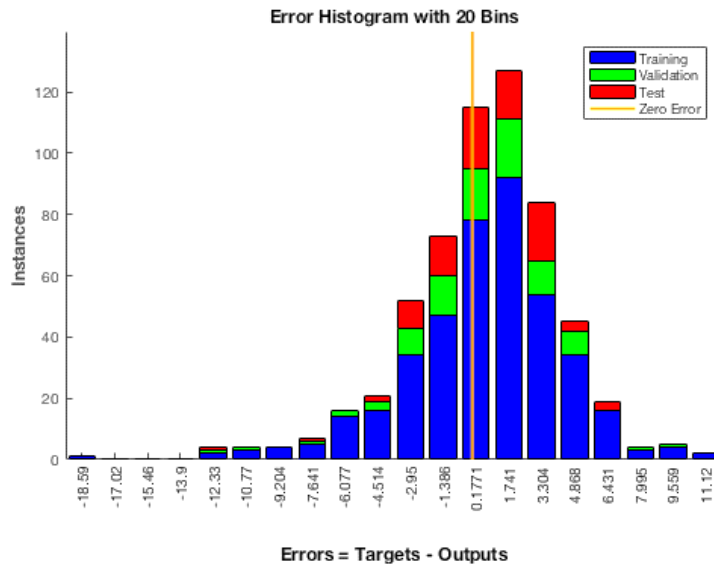


Figure 32 Error distribution

8.5 Network diagnostics of NAR (12), BR training, monthly dataset

Given the results from the autocorrelation test in the previous network (NAR, 2 lags and LM training for the monthly dataset) that have showed that there is still some autocorrelation left in the residuals, the network with higher order lag is proposed next. One such network is described below: it has 12 lags and was trained with Bayesian Regularization.

8.5.1 Test for autocorrelation of the residuals

Performing an autocorrelation test on the new NAR model with 12 lags, it becomes apparent that there is no more significant autocorrelation left in the residuals, indicating that the network has accounted for the dynamics of the data (presented on Figure 33).

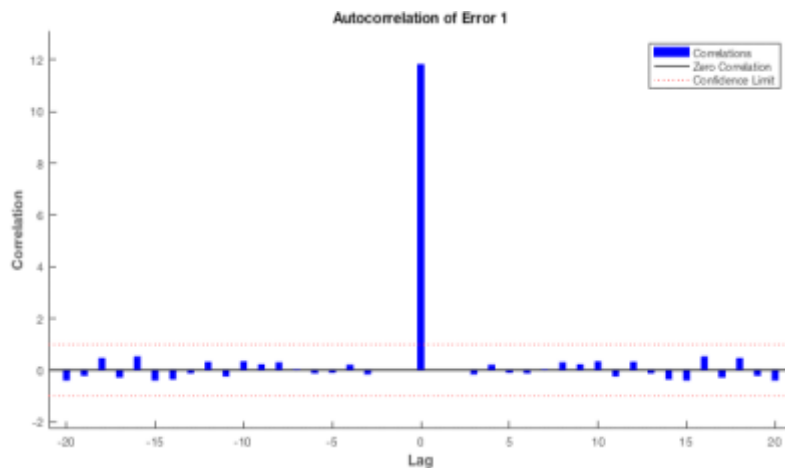


Figure 33 Error autocorrelation NAR 12 lags

8.5.2 Normality test

Despite the network improvement in the autocorrelation test, the test for normality of the residuals continued to be significant, as the residuals from this network also appear to be non-normally distributed, as depicted on error histogram on Figure 34. As it was mentioned before, given large enough number of observations, it does not represent any complications for the network.

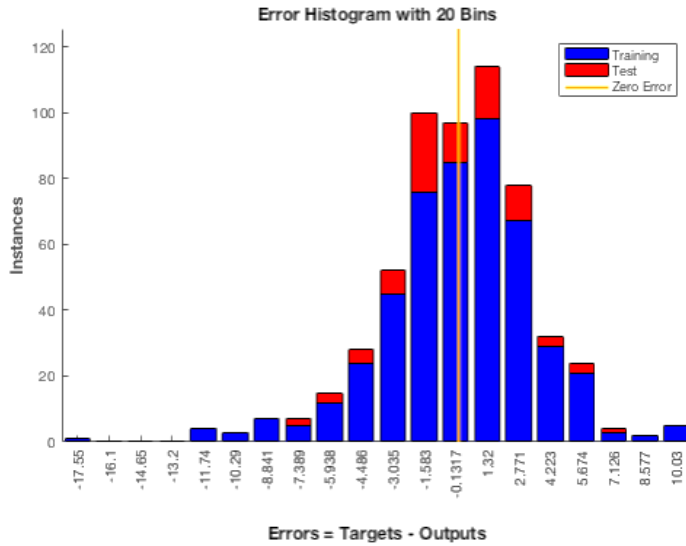


Figure 34 Residuals distribution NAR 12 lags

8.6 Network diagnostics NAR (2), LM training, daily dataset

Moving onto NAR networks for the daily dataset, the first network that is discussed is a simple NAR net with 2 lags, that was trained with LM training method.

8.6.1 Test for autocorrelation of the residuals

From the graphical examination of the autocorrelation of the residuals on Figure 35, one can notice that there are few autocorrelations at lag 8,10,15,17 and 20, that are slightly over the confidence limit specified. This suggests that there is still some dynamics left in the series that have been captured. One suggestion would be to increase the number of lags, as it was helpful in NAR example in monthly dataset. However, increasing number of lags to 20 might significantly complicate the network, create overfitting and deliver even worse results.

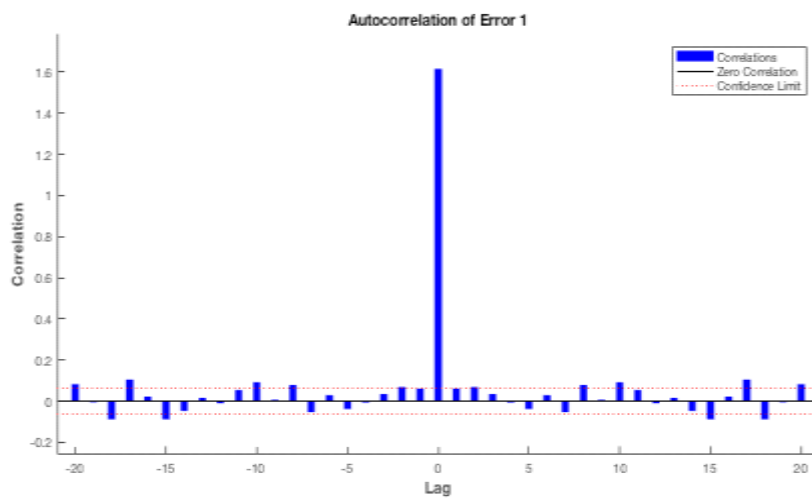


Figure 35 Test of autocorrelations in the residuals

8.6.2 Normality test

According to error histogram (depicted on Figure 36), the errors are clearly not-normally distributed. There is a significant excess kurtosis present (peakedness of the series) and moderate positive skewness. However, given large number of observations, there is no need to be concerned with it and one can continue with forecasting using this network.

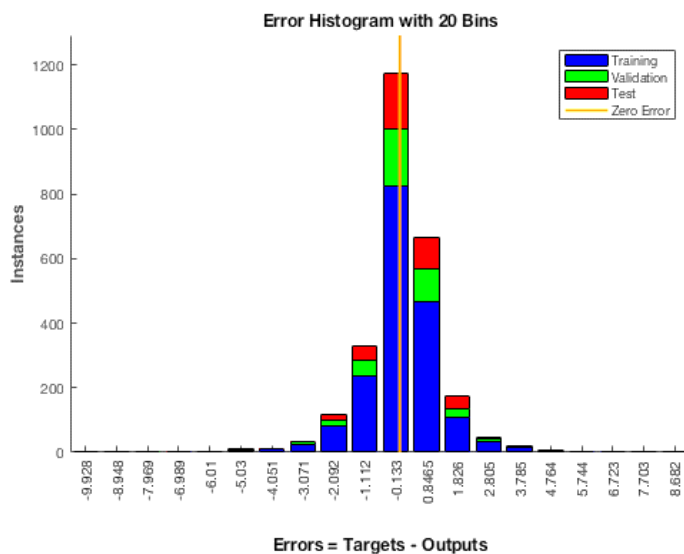


Figure 36 Residuals distribution NAR 2 lags daily dataset

8.6.3 Correlation between series at time t and error at time t .

Additionally, the network can produce graphical examination of the correlation between input and the error (Figure 37). Correlation between input and error does not represent a problem for this network as they are all lower than the significance level (except for the minimal excess of lag 16 and 17).

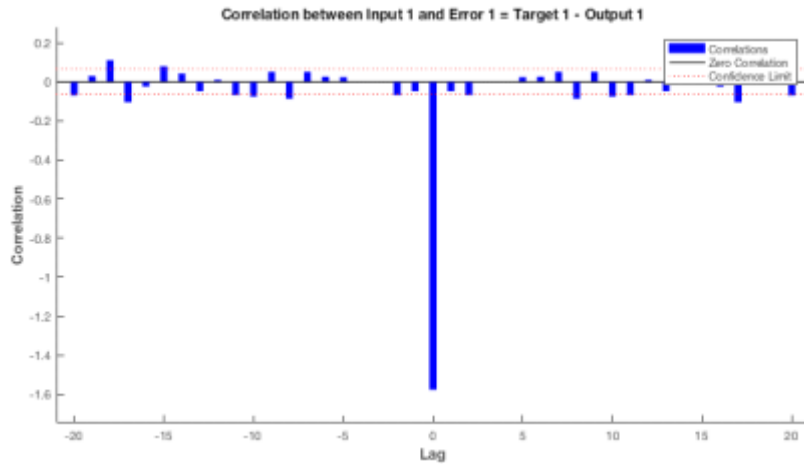


Figure 37 Network diagnostics: Correlation between Input and Error

8.7 Model diagnostics VAR (2) monthly dataset

VAR models contain more assumptions than networks, therefore, it is necessary to check the model adequacy.

8.7.1 Test for VAR stability/Inverse roots of AR characteristic polynomial

In order to ensure a stable VAR (p), model needs to be stationary and ergodic with means, variances and auto-covariances. The applicable test for VAR stationarity is the examination of the eigenvalues of the companion matrix. The VAR (p) is stable if the eigenvalues of the companion matrix are less than 1 in absolute value (i.e. lie inside the unit circle). As evident on Figure 38, all values are inside the unit circle, meaning that VAR satisfies the stability condition.

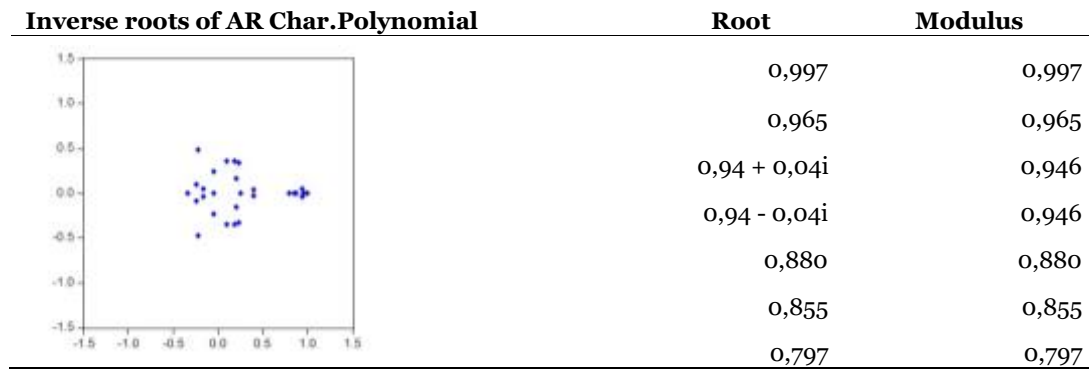


Figure 38 VAR stability test

8.7.2 LM test for serial correlation

Breusch-Godfrey LM test analyses whether there is a serial correlation in the residuals, with the null hypothesis of no serial correlation. Test statistics suggest that there is no serial correlation after lag two, which means that the model is appropriate for the dynamic of the data.

8.7.3 Normality test

The test for normality, similarly to ones on other models, rejected the null of no normality, however, as have been mentioned before, application of central limit theory allows to continue with the model.

Summary:

Overall, based on the results from all the model diagnostics tests, VAR (2) model is adequate for the monthly data. Hence, it could be safely used for forecasting.

8.8 Model diagnostics VAR (2) daily dataset

The following section discusses model diagnostics of VAR (2) for the daily data set.

8.8.1 Test for stability of VAR model

In order to test for the model stability, the test of characteristic polynomial is performed. The purpose of the test to check for stationarity of the model, by measuring the inverse roots of the characteristic AR polynomial. If the model is appropriate and stationary all roots should be less than 1 in absolute value and lies inside the unit circle. Refer to the Figure 39 for graphical examination.

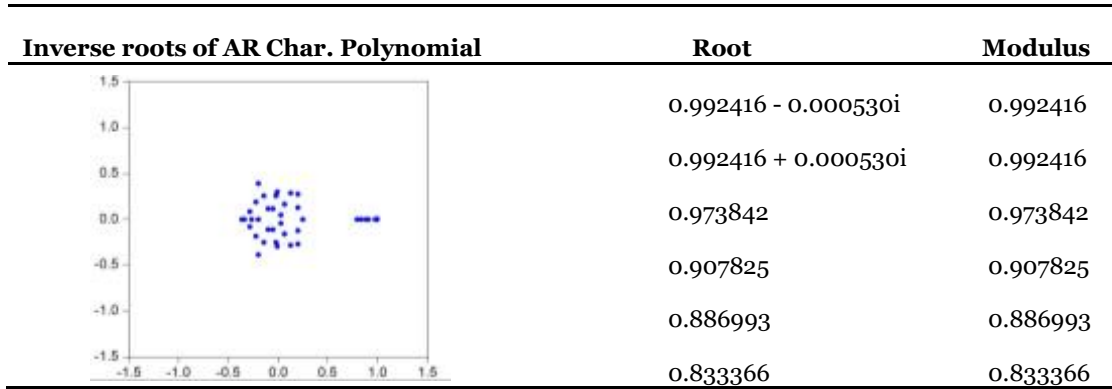


Figure 39 Illustration of VAR stability test for daily data set

According to the figure, all roots lie inside the circle. There are few roots that are very close to 1, however their absolute value is less than 1: the largest roots have values of 0.992 and 0.97. This means that the model satisfies the stationarity condition.

8.8.2 Normality test

There are few normality tests for the residuals that are available. For this study, Cholesky and Urzua tests have been chosen. The null hypothesis of the tests is that the residuals are multivariate normal (the test compares skewness and kurtosis to that of the normally distributed data). The p-values are significant for almost all components, hence suggesting the rejection of the null hypothesis, thus concluding that the errors are not-normally distributed. Having large enough sample allows to apply central limit theorem, and continue with the chosen model.

8.8.3 Portmanteau test for autocorrelation

This test aims at measuring whether there is an autocorrelation in residuals, with null hypothesis stating that there is no residual autocorrelation. Test is statistically significant at 1% level, thus rejecting the null hypothesis, meaning that there is an autocorrelation between the residuals. In order to correctly evaluate the significance of the variables, robust standard errors were used.

8.8.4 LM test for serial correlation

Breusch–Godfrey LM test aims at measuring whether there is a serial correlation in the residuals, with the null hypothesis of no serial correlation. Test statistics are statistically significant at 1% level, hence rejecting the null hypothesis, implying that there is a serial correlation in the residuals.

8.8.5 Heteroscedasticity test

White test for heteroscedasticity was significant and rejected the null of homoscedasticity. This means that there is heteroscedasticity in the data. In order to correctly evaluate the significance of the variables, robust standard errors were used.

Summary

While model diagnostics tests indicate that VAR (2) model is not the best fit for the daily dataset according to the measures of in sample fit, however, one alternative measure of a model fit would be to test its forecasting accuracy, as the model that is good at prediction out-of-sample does not necessarily need to be perfectly fit for the in-sample.

8.9 Model diagnostics VAR (6) daily data set

This subchapter presents results for the model diagnostics tests for VAR (6) model for the daily dataset.

8.9.1 Test for stability of VAR model

First test for VAR (6) model for the daily dataset is the test of characteristic polynomial, measuring the stability of the VAR model. The largest value is 0.994 (which is still lies inside the unit circle). This means that the model satisfies stability condition. (Figure 40).

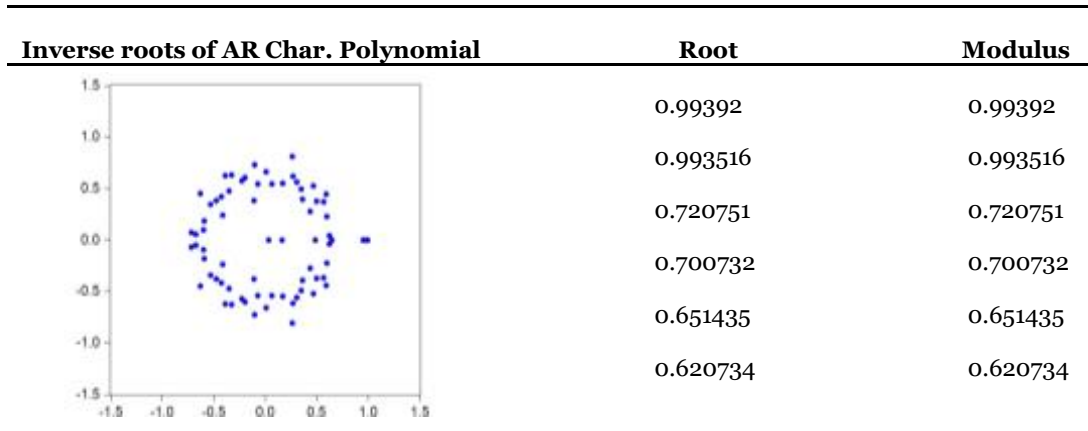


Figure 40 Illustration of VAR stability test for daily data set

8.9.2 Normality test

Jarque-Bera test for normality of the residuals indicated that the residuals are not normally distributed as the null of zero skewness and no excess kurtosis was rejected at 1% level of significance (as evident from the Table 24).

Table 24 Normality test of VAR (6), daily dataset

Test type		Test statistics	df	Prob.
Joint test	for skewness	3347.387	12	0.0000
Joint test	for kurtosis	186345.5	12	0.0000

8.9.3 Portmanteau test for autocorrelation

This test aims at measuring whether there is an autocorrelation in residuals, with a null hypothesis stating that there is no residual autocorrelation. Test is statistically significant at 1% level, thus rejecting the null hypothesis, meaning that there is an autocorrelation between the residuals. In order to correctly evaluate the significance of the variables, robust standard errors were used.

8.9.4 LM test for serial correlation

Breusch–Godfrey LM test aims at measuring whether there is a serial correlation in the residuals, with the null hypothesis of no serial correlation. Test statistics were statistically significant at 1% level for all lags, except lag 7, hence rejecting the null hypothesis, implying that there is a serial correlation in the residuals.

8.9.5 Heteroscedasticity test

White test for heteroscedasticity was significant and rejected the null of homoscedasticity. This means that there is heteroscedasticity in the data. In order to correctly evaluate the significance of the variables, robust standard errors were used.

Summary

While model diagnostics tests indicate that VAR (6) model is not the best fit for the daily dataset according to the measures of in sample fit, one alternative measure of model fit would be to test its forecasting accuracy, as the model that is good at prediction for out-of-sample does not necessarily need to be perfectly fit for the in-sample.

8.10 Network diagnostics NARX (2), LM training, monthly dataset

This subchapter presents results of model diagnostics tests performed on NARX network with 2 lags, and LM training algorithm, that was used for the monthly dataset.

8.10.1 Test for autocorrelation of the residuals

From the graphical examination of autocorrelations in the residual on Figure 41, it becomes evident that there is no significant autocorrelation present except for the lag 5. This, therefore, suggests that incorporating higher lag order into the network might capture some additional dynamic.

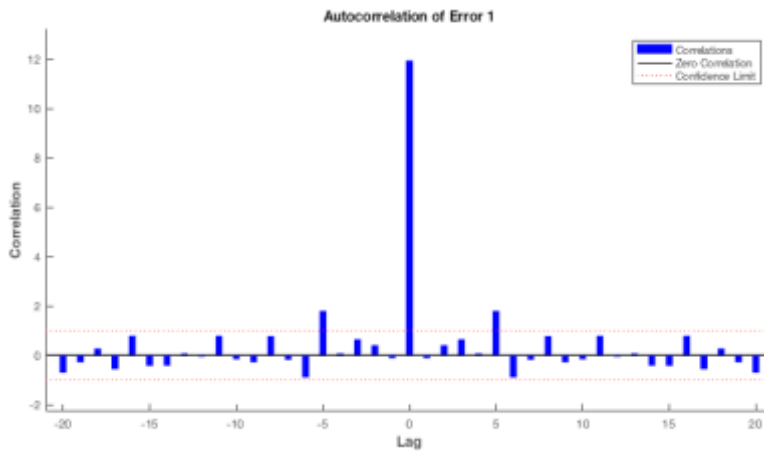


Figure 41 Autocorrelation of the residuals NARX 2 lags monthly

8.10.2 Correlation between input and error

There was no significant correlation detected between the input and error, except for lag 5. (Figure 42). The network seems overall appropriate for the data, although the tests suggest that a network with 5 or 6 lags might capture some of the correlation that is left unaccounted.

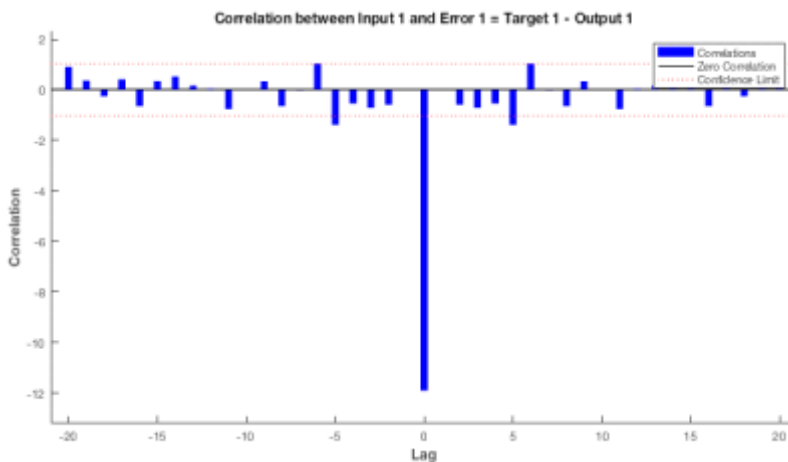


Figure 42 Correlation between input and error NARX monthly

8.11 Network diagnostics NARX (2), LM training, daily dataset

This subchapter portrays network diagnostics for NARX with 2 lags and LM training algorithm that is used for the daily dataset.

8.11.1 Test for autocorrelation of the residuals

There is no autocorrelation detected in the residuals, as depicted on Figure 43, thus suggesting that the network has captured the dynamic of the data well.

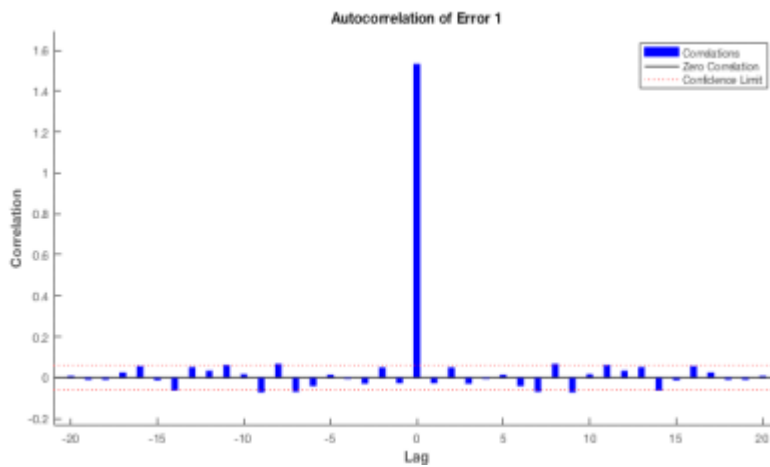


Figure 43 Autocorrelation of the residuals NARX 2 daily

8.11.2 Normality test

Errors of this network appear to have minimal skewness; however, the series has an excess kurtosis, thus indicating the presence of non-normality (Figure 44). This means that none of the models and networks produced normally distributed residuals.

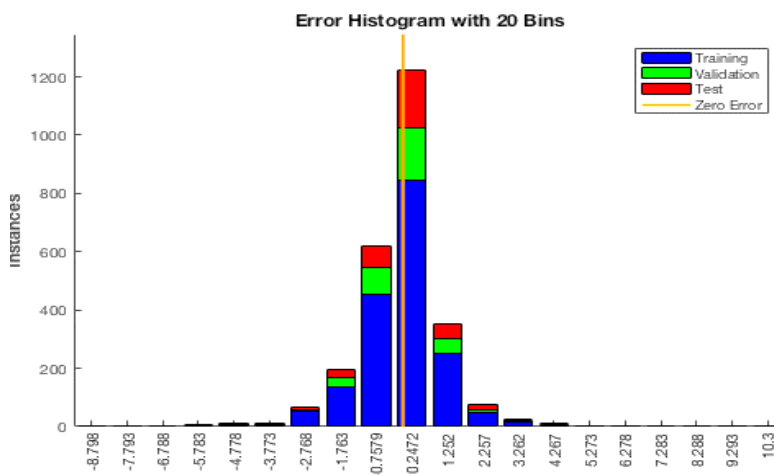


Figure 44 Residuals distribution NARX 2 daily

Summary of the model and network diagnostic sections:

- All ARIMA models are adequate and can be used for forecasting
- The null hypothesis of normal distribution of the error term was rejected in all models and networks. However, the samples contain large enough number of observations in order to apply central limit theorem and continue using the models, as it should not have an impact on forecasting.
- VAR models for monthly dataset is stable and adequate according to in-sample measurements of model adequacy.
- Most model did not have any autocorrelations left in the residuals. All the other models that exhibited autocorrelations in the residuals were fixed by introducing higher lag order, and thus capturing the dynamics. The autocorrelations after the increase in lags become insignificant. The only model that could not get remove autocorrelation problem is VAR for daily data set (discussed below). In order to correctly evaluate the significance of the variables, robust standard errors were used.
- VAR models for daily data set exhibit model diagnostics problems, such as autocorrelation in the residuals and heteroscedasticity. In forecasting, the fit of the model and its validity is often judged by out-of-sample measures, such as MSPE and MAPE, thus poor in-sample model can still provide a good out-of-sample results. Therefore, these models will continue to be in the research and their adequacy will be checked against their forecasting accuracy.
- To conclude, all described models will be tested for forecasting accuracy

9 RESULTS

This chapter will present the results from the empirical research analysis. In addition to the main models described in detail in this study, few further models were introduced, such as MA(1) and VAR(6) for the monthly dataset, and ARMA (3,3), MA(1) for the daily dataset.

9.1 In sample fit

First part of the results consists of comparing statistical and economic measures of in-sample forecasting fit. The results from traditional methods such as ARMA and VAR models presented as given, whereas the results of neural networks represent an average of 35 iterations in order to provide statistically meaningful result. The result is presented below in Table 25, where MSE is a mean square error, SR is a success ratio (percentage of correct signs predictions) and Direction is a percentage of correct directions prediction.

Table 25 Comparison of in-sample forecasting accuracy

This table presents the results from the in-sample forecasting accuracy analysis. The values highlighted in a bold font represent best result for a respective category. The measurements of forecasting accuracy were MSE, Success Ratio and a percentage of correct predictions of a stock index movement (Direction). The models include ARMA (5,2), MA(1), VAR(2) for the monthly dataset; ARMA (4,5), ARMA(3,3), MA(1), VAR(2) and VAR(6) for the monthly dataset; and NAR and NARX neural networks representing the average of 35 iterations of each.

	<i>Monthly dataset</i>					<i>Daily dataset</i>						
	ARMA (5,2)	MA (1)	VAR (2)	NAR	NARX	ARMA (4,5)	ARMA (3,3)	MA (1)	VAR (2)	VAR (6)	NAR	NARX
MSE	11.4	12.0	11.8	12.2	11.4	0.637	0.65	0.64	0.638	0.65	1.8	1.7
SR %	60.5	58.3	64.6	62.4	61.2	50.2	49.5	50.3	50.3	51.3	54.9	55.7
Direction %	73.3	72.6	73.2	73.4	73.8	75.7	75.0	75.7	75.0	73.7	75.8	75.5

From the evaluations using statistical measures of forecast accuracy, ARMA models have surpassed the average of neural networks performance. In a monthly dataset, ARMA (5,2) has the lowest MSE, although followed very closely by a NARX network, while ARMA(4,5) has the smallest MSE among the models for the daily dataset and neural networks had the largest MSE.

However, judging from an economic evaluation of forecast accuracy, neural networks demonstrated a competitive performance. In monthly dataset, both **NAR and NARX**

produced better prediction of the direction of the stock return, than the traditional models (both ARMA and VAR). According to the Success Ratio measurement, networks had second best performance and were surpassed only by VAR model. As for the daily dataset, **artificial neural networks had better prediction of the sign of the forecast**, measured by a Success Ratio, than traditional methods (ARMA and VAR). The difference between the superior accuracy of artificial neural networks and traditional model for sign prediction is 5% and could make a significant difference for the trader. **The forecast accuracy for direction prediction was the highest for NAR networks**, followed very closely by ARMA (4,5) and MA (1). The difference between accuracy of NAR and ARMA models in this case was insignificant.

As it was discussed there is a significant difference between in-sample and out- of-sample forecasting. For example, from Table 26, one can observe the difference in statistical measures of forecast accuracy between in-sample and out-of-sample prediction. Clearly, the forecasting error becomes larger when the model is forecasting out-of-sample. Therefore, it is important to compare models based on their out-of-sample forecasting accuracy.

Table 26 In-sample fit vs. out-of- sample forecasting

This table presents the results of the in-sample forecasting accuracy results, and out-of-sample forecasting accuracy results for both monthly and daily dataset. The measurements of the forecasting accuracy are RMSE, MAE, MAPE and Theil U.

		<i>In-sample forecasting</i>				<i>Out-of-sample forecasting</i>			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
Monthly data	ARMA (5,2)	3,369	2,369	152,747	0,681	3,439	2,403	153,510	0,701
	MA (1)	3,470	2,413	162,499	0,728	3,472	2,412	162,341	0,731
	VAR (2)	3,432	2,381	504,964	0,702	3,842	2,547	723,610	0,729
Daily data	ARMA(4,5)	0,798	0,561	165,226	0,868	0,800	0,561	162,080	0,855
	ARMA(3,3)	0,803	0,561	160,654	0,872	0,803	0,562	166,606	0,875
	MA(1)	0,800	0,557	169,494	0,897	0,801	0,557	166,526	0,882
	VAR(2)	0,799	0,572	1565,954	0,822	0,801	0,572	1565,954	0,822
	VAR(6)	0,807	0,589	1162,301	0,758	0,903	0,656	639,417	0,726

9.2 Out of sample prediction accuracy

As have been discussed before, the forecasting out-of-sample needs to be compared using different statistical and economic measures than the in-sample fit. Statistical measures for out-of-sample measures are Mean Square Prediction Error (MSPE) and Mean Absolute Prediction Error. Furthermore, economic measures for out-of-sample forecasting are Success ratio (proportion of correctly predicted signs of the forecast) and Direction ratio (proportion of correctly predicted directions forecasts).

The summary of the out-of-sample forecasting is presented in Table 27. The following sections will explain the results in detail.

Table 27 Forecasting out-of-sample accuracy results

This table shows the results for the out-of-sample forecasting accuracy analysis for both daily and monthly datasets, predicting S&P500 stock index return one step-ahead. The models include ARMA (5,2), MA (1), VAR (2) for the monthly dataset; ARMA (4,5), ARMA(3,3), MA(1), VAR(2) and VAR(6) for the daily dataset; and NAR and NARX neural networks representing the average of 35 iterations of each and for both dataset. The measurements of the forecasting accuracy are MSPE (mean square prediction error), MAPE (mean absolute prediction error), Success Ratio (and corresponding Direction Analysis ratio measuring its statistical significance), and a percentage of correct predictions of a stock index movement (Direction). The values highlighted in a bold font represent best results for a respective category.

Monthly dataset							Daily dataset						
	ARMA (5,2)	MA (1)	VAR (2)	VAR (6)	NARX (2)	NARX (6)	ARMA (4,5)	ARMA (3,3)	MA (1)	VAR (2)	VAR (6)	NARX (2)	NARX (6)
MSPE	11,8	12,0	14,8	15,8	14,9	13,8	0,36	0,37	0,36	0,40	0,44	0,40	0,37
MAPE	2,40	2,41	2,55	2,73	2,63	2,53	0,41	0,42	0,41	0,46	0,48	0,45	0,42
SR %	60,8	60,8	62,7	66,7	64,1	64,1	56,3	51,0	56,0	48,3	48,3	50,3	55,0
DA	1,1	0,9	1,7	2,9	1,1	1,3	1,57	0,2	1,55	-0,4	-0,4	0,1	1,3
Direction %	75,8	75,8	74,5	72,5	74,6	74,5	78,9	80,3	80,9	76,9	72,9	76,9	78,6

9.2.1 Artificial neural networks versus traditional models

When comparing the forecasts based on the statistical measures such as MSPE and MAPE, ARMA models have the smallest forecasting error out of the tested models. However, when evaluating from the economic significance perspective, neural networks become more competitive. For the monthly dataset the **NARX networks had better accuracy in**

predicting the sign of the forecast, than VAR (2) model, ARMA (5,2) and MA(1); being surpassed only by VAR model with 6 lags. Networks also had **higher prediction rate of the direction of the future returns than respective VAR models**. However, the statistical significance of this performance is low. Based on the S-test that measures whether the one forecast is superior to another, the difference between NARX with 2 lags forecast and VAR (2) forecast was not significant (S test=0.81 and 0.89 for daily and monthly forecast, respectively, as evident in Table 28).

For the daily dataset, the forecast produced by NARX network with 2 lags had slightly better accuracy than the forecast produced by VAR (2) based on both MAPE and Success ratio, whereas **NARX network with 6 lags produced higher accuracy forecast than VAR (2) and VAR (6) models** based on MSPE, MAPE, SR and Direction forecast for both daily and monthly datasets. S-test, comparing the performance of these two models, indicated that forecast produced by artificial neural networks is statistically more accurate than the forecast by VAR (6), depicted in Table 28.

Success ratio of a NARX network with 6 lags was 55% in a daily dataset and 64,1% in a monthly dataset, which indicates that the model was able to correctly predict the sign of the stock index return 55% and 64,1% of the time, respectively; which is better than the statistical 50% chance. However, at 10% level of significance, Directional Accuracy (DA) test did not reject the null of no difference between the success ratio of NARX and success ratio in case of independence.

When neural networks are compared to ARMA models, neural network's forecast tends to be less accurate than the one produced by univariate models. However, the S-test was 1,15 and 1,54 for daily and monthly dataset, respectively, which is not significant at 10% level. Therefore, S-test's null hypothesis of no difference between the forecasts cannot be rejected, meaning that ARMA models did not outperform neural networks.

Overall, this implies that while neural networks have better performance than some traditional models (for example NARX (6) vs. VAR (6)), the null of statistical hypotheses stated in this thesis that the accuracy of neural networks is not different to the one of traditional methods cannot be rejected at 10% confidence level.

This means that there was no model that was superior in forecasting than others, in all economical and statistical measures.

Table 28 S-test pair-wise comparison of forecasts' performance significance

The following tables present the results from the S-tests and pairwise metrics of forecasting accuracy produced by competing forecasting models. The models that are compared are NARX (6) and VAR (6), NARX (2) and VAR (2), and NARX (2) and ARMA (4,5) models for daily dataset; NARX (6) and VAR (6), NARX (2) and VAR (2), and NARX (2), and NARX (2) and ARMA (5,2) models for the monthly dataset. The measurements of the forecasting accuracy are MSPE (mean square prediction error), MAPE (mean absolute prediction error), Success ratio (shown in percentages) and respective Direction Accuracy statistics, and Direction ratio (percentage of correct direction prediction, expressed in percentages). The S-test measures the significance of the difference between the two forecasts, and implies that the difference is significant when S-test statistics is higher in absolute value than the confidence interval.

<i>S-test: NARX vs. VAR daily dataset</i>				
	NARX (6)	VAR (6)	NARX (2)	VAR (2)
MSPE	0,37	0,44	0,40	0,40
MAPE	0,42	0,48	0,45	0,46
SR %	55	48,3	50,3	48,3
DA	1,27	-0,42	0,05	-0,42
Direction %	78,6	72,9	76,9	76,9
S-Test	1,96**		0,81	

<i>S-test: NARX vs. VAR monthly dataset</i>				
	NARX (6)	VAR (6)	NARX (2)	VAR (2)
MSPE	13,84	15,75	14,65	14,76
MAPE	2,53	2,73	2,93	2,55
SR %	64,1	66,7	53,6	62,7
DA	1,32	2,88	1,51	1,68
Direction %	74,5	72,5	70,6	74,5
S-Test	0.89		0,89	

<i>S-test: NARX vs. ARMA</i>				
<i>Monthly dataset</i>			<i>Daily dataset</i>	
	NARX (2)	ARMA (5,2)	NARX (2)	ARMA (4,5)
MSPE	14,97	11,82	0,40	0,36
MAPE	2,63	2,40	0,45	0,41
SR %	64,1	60,8	50,3	56,3
DA	1,11	1,11	0,05	1,57
Direction %	74,5	75,8	76,9	78,9
S-Test	1,54		1,15	

9.2.2 Picking the best model

This research has utilized different measures of accuracy for out-of-sample forecasting, and identified different models that got the best performance in each of the categories.

9.2.2.1 Monthly dataset

Starting from the monthly dataset: ARMA (5,2) had the lowest Mean Square Prediction Error and Mean Absolute Prediction Error than other models for the dataset. It implies that ARMA (5,2) had smaller forecasting error, measured as the difference between a forecasted and actual value.

However, VAR models (especially VAR (6)) were significantly better at predicting the sign of the forecast. The Directional Accuracy (DA) test rejected the null that the Success ratio produced by VAR was no different to the success ratio in case of independence. This implies that VAR model was able to model the variation of S&P500 return and forecast it for out-of-sample, and that predictive power is economically significant. Given the result that 67% of the time the forecast produced by VAR correctly predicts the sign of the return (will there be a negative return or positive), one can perform a competing trading strategy based on this prediction.

Prediction of the stock movement (Direction ratio) appeared to be very successful for all models, however, ARMA models (ARMA (5,2) and MA (1)) had the highest percentage of correct direction forecast (75.8%).

9.2.2.2 Daily dataset

Superior forecasting models for this dataset were ARMA models. Based on MSPE, MAPE and SR measures, ARMA (4,5) had the best forecasting accuracy, as it had the lowest forecasting error and had higher success ratio than other models. MA (1) was the best in Directions ratio and had equally high Success Ratio as ARMA (4,5).

9.3 Sensitivity analysis

This subchapter presents the results from the sensitivity analysis of this study. It consists of sensitivity analysis of the artificial neural networks, given different number of lags and different learning algorithm, as well as the robustness of the results given a number of iterations.

9.3.1 Sensitivity of results given a number of iterations

While researching and analysing the forecasting accuracy of artificial neural networks it is important to acknowledge that they often produce different result every time they are run, thereby making it very challenging to compare with the traditional models. While the average values for 35 network iterations were presented in the results, it is also important to consider the range of values that occurred. The table for the range of value for the in-sample fit is presented below (Table 29).

Table 29 Range of values for ANN iterations for in-sample fit

This table reports the range of Success Ratios and Direction predictions for the forecasts produced by artificial neural networks, given number of iterations (35). The table shows minimum, maximum and average values for the metrics for daily and monthly datasets.

		NARX			NAR		
		Min	Max	Average	Min	Max	Average
Monthly data	SR	55,1 %	64,2 %	61,2 %	54,4 %	64,6 %	62,4 %
	Direction	67,0 %	75,1 %	73,8 %	68,7 %	75,7 %	73,4 %
Daily data	SR	52,2 %	60,9 %	55,7 %	51,3 %	61,8 %	54,9 %
	Direction	69,4 %	76,1 %	75,5 %	73,6 %	76,8 %	75,8 %

9.3.2 Sensitivity to number of lags

Out-of-sample forecasting accuracy was very sensitive to the number of lags. Networks with two lags significantly underperformed in comparison with higher order lag networks. For both monthly and daily dataset NARX networks with 6 lags produced better forecasts, measured by MSPE and MAPE at 5% level of significance.

However, the increase from 6 to 12 lags deteriorated the performance of the network for the monthly dataset, causing it to underperform in comparison to the one with 6 lags. On contrast, the performance of NARX 12 was slightly better than that of NARX 6 for the daily dataset (the results are presented in Table 30).

9.3.3 Sensitivity to training algorithms

As it was mentioned in this paper one of the important network configuration is the training algorithm with which it learns how to fit the data. The comparison between ANN fit for different datasets (monthly and daily), different number of lags (2,6,12) and different training algorithm (LM, SCG and BR) is presented below (Table 30). The results represent the average of 35 iterations of each configuration.

Table 30 Robustness of ANN given different network configuration

The following tables report the forecasting accuracy of neural networks given various network modifications. Measures of the forecasting accuracy are MSE and RMSE. Network modifications include the robustness of the results given a number of lags (2, 6 and 12), training algorithm (Levenberg-Marquardt, Scaled Conjugate Gradient and Bayesian Regularization), used on different datasets with different time periods and time-series frequencies (monthly dataset for 1968-2016 period and daily dataset for 2007-2017 period), and using univariate or multivariate networks (NAR vs. NARX networks).

<i>Robustness of NAR, monthly dataset</i>								
Training algorithm	Levenberg-Marquardt			Scaled Conjugate Gradient			Bayesian Regularization	
N.of lags	2	6	12	2	6	12	2	6
RMSE	3,50	3,52	3,49	3,96	3,47	3,56	3,84	3,71
MSE	12,23	12,40	12,19	15,66	12,07	12,65	14,74	13,75

<i>Robustness of NARX, monthly dataset</i>								
Training algorithm	Levenberg-Marquardt			Scaled Conjugate Gradient			Bayesian Regularization	
N.of lags	2	6	12	2	6	12	2	6
RMSE	3.81	3.52	3.38	3.96	3.63	3.50	3.90	3.72
MSE	14.53	12.39	11.43	15.70	13.15	12.23	15.23	13.87

<i>Robustness of NAR, daily dataset</i>								
Training algorithm	Levenberg-Marquardt			Scaled Conjugate Gradient			Bayesian Regularization	
N.of lags	2	6	12	2	6	12	2	6
RMSE	1.35	1.37	1.44	1.46	1.35	1.51	1.39	1.28
MSE	1.83	1.89	2.08	2.14	1.83	2.29	1.94	1.65

<i>Robustness of NARX, daily dataset</i>								
Training algorithm	Levenberg-Marquardt			Scaled Conjugate Gradient			Bayesian Regularization	
N.of lags	2	6	6	2	6	6	2	6
RMSE	1.40	1.32	1.32	1.48	1.31	1.32	1.45	1.32
MSE	1.95	1.75	1.75	2.18	1.71	1.71	2.10	1.74

Firstly, explaining the results from the monthly dataset, it is evident that almost in all instances MSE was lower for the network with 6 lags when compared to the one with 2 lags. However, the network with 12 lags with LM training outperformed the one with 6 lags. Networks with two lags and Scaled Conjugate Gradient (SCG) learning algorithm performed worse than the networks that were trained with different learning algorithms, keeping everything else unchanged. SCG was also a worse learning algorithm for the network with 12 lags as the network's performance was inferior than that of a network with LM training and 12 lags. However, NAR network using SCG training and 6 lags had lower forecasting error than LM network with 6 lags. Overall, the **best model for the monthly dataset was NARX network with 12 lags trained with Levenberg-Marquardt** algorithm, followed by a NAR network with 6 lags and Scaled Conjugate Gradient algorithm.

From the results of the networks in a daily dataset, the networks with 6 lags continued to outperform the networks with 2 lags, which is consistent with the results in a monthly dataset. The network that had the **lowest MSE** for the daily dataset was a **NAR network with 6 lags and Bayesian Regularization**, followed by a NARX network with 6 lags and SCG training algorithm, and a NARX network with 6 lags and LM training algorithm. Networks that had the worst fit were NARX net with 2 lags and SCG algorithm, followed by a NAR net with 2 lags and SCG training algorithm.

The forecasting accuracy (both in-sample and out-sample; and measured by statistical and economic measures) of networks with 6 lags is significantly higher than that of the networks with 2 lags, thus rejecting the statistical null hypothesis that stated that the lags of S&P500 return have no effect on the current values of S&P500, implying that past values of S&P500 return are useful in forecasting future returns.

9.4 Effect of market sentiment variables on stock return

Artificial Intelligence in finance comes not only in the forecasting methods, but also from its ability to use large amount of data from around the world and incorporate it in the model. However, despite the ease of access to the data and its inclusion in the models, it has not been yet widely accepted. Using Google trends data as a proxy for a Big data of market sentiment in a daily dataset, and Michigan's Consumer Sentiment Index as a proxy for a market sentiment in a monthly dataset, this study was testing whether simple data on market sentiment can explain the market returns.

9.4.1 Market sentiment, monthly dataset

In order to test for the importance of market sentiment in predicting stock market return, this paper have applied granger causality tests, and impulse responses tests.

Granger causality VAR (2), monthly dataset

Block significance tests (Granger causality tests) examines the correlation between past values of one variable and current value of another variable.

Granger causality tests did not find a granger causality between market sentiment, measured by a Consumer Sentiment Index from the university of Michigan, and monthly S&P500 index returns, as evident on Table 31. This means that Market Sentiment variable in a monthly dataset has not contributed significantly to the forecasting of S&P500 returns.

Table 31 Granger causality test, monthly dataset

This table depicts the results of the Granger causality test performed on VAR (2) model for the monthly dataset over the period of 1968-2016, where term, credit, oil, gold, money supply, CPI, dividend yield, unemployment, market sentiment and exchange rate variables are tested for granger causality of S&P500 return, by examining whether the changes in before-mentioned variables appear before the changes in S&P500 return. Significant probabilities are highlighted in a bold font, indicating that a particular variable g-causes the changes in S&P500 return, meaning that that variable is useful in forecasting stock index return.

Variable	Prob.
Term	0,13
Credit	0,02 **
Oil	0,18
Gold	0,64
Money Supply	0,88
CPI	0,11
Unemployment	0,01 ***
Market Sentiment (Michigan proxy)	0,26
DY	0,92
GBP	0,81
JPY	0,78
All	0,04

9.4.2 Market sentiment, daily dataset

Granger causality VAR (2) daily dataset

The study of the effect of a market sentiment on S&P500 return in a daily dataset has provided more promising results. The proxy for a market sentiment in daily dataset was a GT_index variable, which corresponds to Google trends term “S&P500 index”. It is a new

variable that appeared with Big data, and its significance on S&P500 returns has not been studied before due to the skepticism from the traditional finance theory holders.

Granger causality test indicated that GT_Index statistically significantly granger-causes S&P500 returns, as depicted in Table 32. This means that changes in the concentration of google searches for “s&p500 index” happen before the changes in actual S&P 500 index returns, hence the variable should be included in the forecasting model as it should help in the prediction of the return.

Table 32 Granger causality test, VAR (2), daily dataset

This table depicts the results from the Granger causality test performed on VAR (2) model for the daily dataset over the period of 2007-2017, where term, credit, oil, gold, FTSE All share index return, Microsoft stock return, VIX index return and Google trends term “S&P500 index” are tested for granger causality of S&P500 return, by assessing whether the changes of the before-mentioned variables appear before the changes in S&P500 return. Significant probabilities are highlighted in a bold font, indicating that a particular variable g-causes the changes in S&P500 return, meaning that that variable is useful in forecasting stock index return.

Variable	Chi-sq	Prob.
Term	7,593	0,022 **
Credit	2,635	0,268
Gold	4,922	0,085 *
Oil	3,094	0,213
FTSEAll_1	5,073	0,079 *
MSFT_1	5,245	0,073 *
VIX	2,178	0,337
GT_index	5,678	0,070 *
All	33,012	0,007

Impulse responses VAR (2) daily dataset

In order to better understand the nature of the relationship between the variable, one could run an Impulse response test, to measure the unit change of one variable given the shocks in lags of another variable. Given the interest in Google trend term and its significant granger causality of S&P500 return, the impulse response test was performed and the output can be viewed on Figure 45, where vertical axis is expressed in units of dependent variable, and the blue line is a point estimate for the expected change in S&P500 return following a unit impulse after the number of periods of the GT_index variable (horizontal axis).

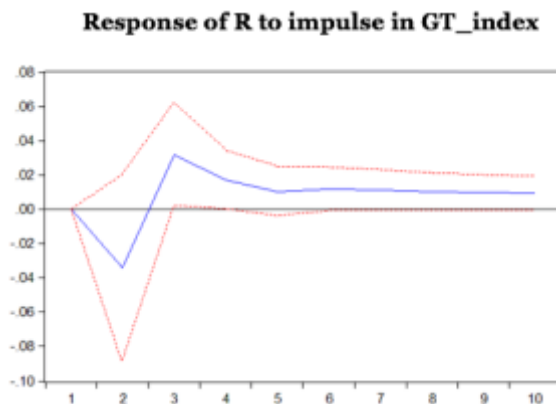


Figure 45 Impulse response of R to Google search “S&P500 index”

From the graph, it is evident that S&P500 return has a varying correlation with GT_index variable. It firstly has a negative association, which changes to the positive one after lag 3.

Granger Causality VAR (6) daily

It is of a further interest to check whether the granger causality is persistent with more lags. Upon examining the results of Granger causality test on VAR (6) model (depicted in Table 33), it becomes evident that at 5% significance level, the lags of GT “S&P500 index” variable g-cause the current values S&P500 return.

Table 33 Granger causality test, VAR (6), daily dataset

This table depicts the results of the Granger causality test performed on VAR (6) model for the daily dataset over the period of 2007-2017, where term, credit, oil, gold, FTSE All share index return, Microsoft stock return, VIX index return and Google trends term “S&P500 index” are tested for a granger causality of S&P500 return, by assessing whether the changes of before-mentioned variables appear before the changes in S&P500 return. Significant probabilities are highlighted in a bold font, indicating that a particular variable g-causes S&P500 return, meaning that that variable is useful in forecasting stock index return.

Variable	Chi-sq	Prob.
Term Spread	24,30	0,004 ***
Credit Spread	17,75	0,012 **
Gold	12,17	0,046 **
Oil	10,58	0,112
GBP	4,24	0,645
JPY	10,69	0,098 *
FTSE All_1	10,07	0,122
MSFT_1	12,04	0,058 *
VIX	4,43	0,618
GT_S&P500Index	14,12	0,028 **
All	135,09	0,000

Impulse responses VAR (6), daily

The persistence of the significance of GT_index variable for S&P500 return continue from VAR (2) model into VAR (6), therefore, it is interesting to investigate further the nature of the relationship. The impulse response of stock index return to Google trend search for 10 lags is depicted on Figure 46.

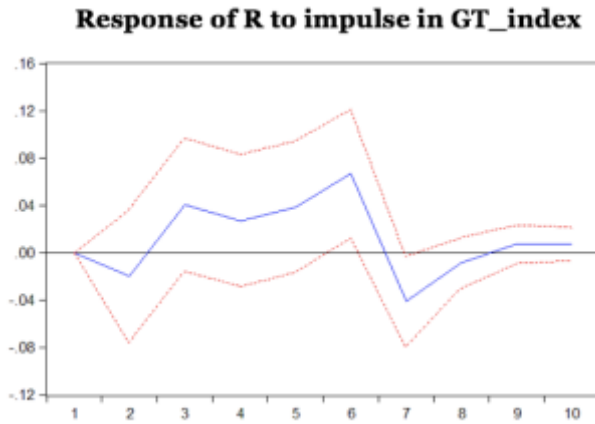


Figure 46 Impulse response of S&P500 return to Google search of “S&P500 index”

It is evident that S&P500 index return and Google trends data have a time-varying association, with the first 2 lags and lag 7 having a negative association with stock index returns and a positive one at other lags. While it is difficult to interpret the exact numerical value of the relationship coefficient at each lag, it is nevertheless evident from the tests performed and that Google trends search term “S&P500 index” has some explanatory power in predicting the stock index and hence is useful for building a forecasting model.

9.5 Effect of explanatory variables on S&P500 return

In the majority of the iterations NARX networks (including a matrix of independent variables) outperformed NAR network (networks with no other input but its own lags). This suggests that explanatory variables have predictive power of S&P500 return.

Furthermore, from the VAR regression output for a monthly dataset, Term spread, Credit Spread and Unemployment variables were found to be significant at explaining S&P500 return (Table 18).

Additionally, Granger causality tests conclude that at 2% confidence level, credit spread granger-causes S&P500 index return; at 1% confidence level unemployment rate granger

causes S&P500 Index return. Furthermore, at 13% and 11% significance level, lags of term spread and CPI g-cause the current values S&P500 index return, as evident in Table 31.

From the VAR model output in a daily dataset, return's own first lag (R-1), Term spread (both lags), Gold (2nd lag), GBP(1st lag), FTSE index (1st lag), and 2 lag of MSFT index and Google search for "S&P500 index" have been found significant in explaining S&P500 return (Table 21). It appears that at 5% significance level lags of Term spread granger-causes the current values of the S&P 500 returns, while at 10% significance lags of gold, FTSE all share index returns, lags of MSFT stock returns, and lags of returns of GT_index granger cause the current returns of S&P500, as evident from Table 32. These results are supported by the granger causality tests performed on VAR (6) model on daily dataset, where at 1% significance level lags of Term spread variable granger cause the current values of S&P500 index returns, and at 5% significance level the lags of Credit Spread, Gold and Google trend term "S&P500 index" g-cause the current values S&P500 Index return, while at 10% significance level lags of Microsoft returns and change in JPY to USD currency exchange rate g-cause the current values of S&P500 return.

This means that the statistical null hypothesis of explanatory variables having no predictive power over the S&P500 return is rejected, meaning that these explanatory variables are useful in modeling S&P500 return.

10 DISCUSSION

10.1 Answering the Research question

This study explored various artificial intelligence applications in a finance field. It identified and discussed the main areas for AI in finance: portfolio management, bankruptcy prediction, credit rating, exchange rate prediction and trading. This paper provided numerous examples of the companies that have invested in AI and type of tasks they use it for. The main type is efficiency maximization by using algorithms that scan substantial amount of data in a shorter time period than any human could do. For instance, data mining, market sentiment and profile scanning are examples of the most applicable tasks for AI.

Furthermore, this research investigated one of the proxies of Artificial Intelligence, - Artificial Neural Networks - for the prediction accuracy of stock market returns. Few previous researches, using different neural network types, have found that ANN outperforms its traditional counterparts, such as ARIMA models, exponential smoothing or logit models. This paper has used recurrent dynamic networks for forecasting S&P500 return and compared it to the forecasting accuracy of ARIMA and VAR models, using statistical measures (Mean Square Prediction Error and Mean Absolute Prediction Error) and economic measures (Sign prediction and Direction prediction) of forecast accuracy. Even though some networks have outperformed certain linear models, the overall result is mixed. ARMA models were the best in minimizing the forecast errors, while networks often had better accuracy in sign or direction prediction. Artificial neural networks outperformed respective VAR models in many parameters, but the difference of this outperformance was not significant, measured by S-tests. Therefore, on 10% level of confidence, there was no model/network that was superior than others.

While the result suggests inconclusive results, this does not imply insignificance of Artificial Intelligence in the finance field. AI is developing rapidly with many new algorithms appearing in the world yearly, which allows for testing of more sophisticated networks that haven't been used in this study. Moreover, this study has utilized ANNs that use supervised learning, and given recent breakthrough into the area of unsupervised learning, further investigation of ANN and testing of the networks with unsupervised learning could provide different results.

However, this study found other interesting results: Google trends search terms have been found useful in modeling of S&P 500 returns. This implies further study of Big data and the use of algorithms and other AI applications to incorporate vast amount of statistics available online, in order to capture more accurate dynamics of the financial time-series. These and other findings will be discussed in the following sections.

10.2 What to learn from companies that implement AI

One of the main focuses of this thesis was the investigation of the various AI application in a finance field. The following forms of AI application in finance were identified: anomaly detection, portfolio management, algorithms trading, text mining, market sentiment analysis, credit evaluation and loan underwriting. The key message that companies that invest in AI portray is improved efficiency, use of big data and sophisticated data analytics, automated systems reducing the cost, biases and human errors.

10.2.1 Credit card, banks and major financial institutions

One of the applications of AI for credit card companies, banks and other major financial institutions could come in the form of anomaly detection. These organizations have a lot of papers work, transactions and money transfers, making it impossible to physically check every single transaction by hand. Use of algorithms can maximize the efficiency of the organizations by taking over the routine and manual tasks, such as checking the patterns in transactions, and leaving the human professional to be occupied with subjects that deserve immediate attention and humane intervention (for example, to deal with only a handful of transactions that received a red flag by an algorithm).

Furthermore, banks and loan issuing companies can benefit from AI application in profile scanning of the applicant. Algorithms could identify the factors that are strongly correlated with the default probability. As it has been discussed it is often not the most commonly believed factors that affect the probability of default, and it could be due completely different reasons why a person defaulted on the loan. This means that AI algorithm will minimize the human prejudice (employee's prejudice towards young people, unemployed, Hispanic, gender-based or others), thus allowing for a more efficient search for good applicants, more loans to be issued while minimizing the default rate.

10.2.2 Portfolio management firms

For portfolio management firms AI represents a challenge and an opportunity, depending whether the company manages to quickly adjust to the changing environment. Robo-advisory brings new advantages for the consumers in terms of more time- and cost- efficient advice, and brings challenges to the current portfolio managers that have not adapted to it yet. Moving onto the digital age, less and less people require human face-to-face interaction. This is surely driven by the trends of modern lifestyle: online education, social networking (where people communicate without face-to-face interactions), online shopping etc. Therefore, the job of a human portfolio manager could become obsolete as most of his/her tasks could be done more efficiently by an algorithm, whereas the clientele's desire to see the actual human professional (with whom they need to schedule an appointment time and for whom they need to often wait and travel to, as well as to pay a higher fee) will decrease. The challenge for current portfolio managers is to generate a better performance than algorithms, and while at its current stage of AI development it is still possible, when the unsupervised learning will advance to its next level, it will not be achievable. The best strategy is to learn about AI and to incorporate it in the current working model, thus maximizing already current performance and stepping towards the forthcoming changes.

10.2.3 Investment and trading companies

Traditional research in investments has identified few factors that correspond to the company's future performance, whether those are quality measures (from companies' financial statements), ownership or trading valuations. However, each company has much more features that could affect its return and which have not been studied yet, given laborious work. AI in the form of data-mining algorithmic system could scan large amount of performance-related information of the companies to identify true causes of the performance. Hence, AI could improve the screening of the companies that are searching for a profitable investment. The algorithms could perform this search in a matter of seconds and identify a handful of exceptional companies in emerging markets or where the information is hardly available for the human user (different language, lack of knowledge, personal prejudices). This would maximize the efficiency of the human traders or investment specialists that would spend less time searching, as they would need to consider only some top picks that have been provided to them by the algorithm.

Furthermore, trading companies can take the use of data-mining function of AI. Algorithms could scan social media activity (Twitter, Facebook etc.) or online behavior (Google or Wikipedia searches) in the region and understand how does a market feel about a certain event (that happened or is about to happen). As this research found, market sentiment could be very useful in modeling stock returns, therefore, including it in the analysis should provide a competitive advantage to trading firms and possibly improve the results.

10.3 Choosing the best model

Forecasting performance varies between the models, for example VAR models had the best performance in predicting the sign of the future return during the monthly dataset, but in the daily dataset their prediction was the lowest. This means that there was no model that was superior in forecasting that the other model.

It is therefore hard to determine which model was the best, and whether the explanatory variables helped the predictions or not (because MSPE increased (bad result), but so did the Success ratio (good result)). Therefore, the choice of the model should depend on the preferences of the person who is forecasting. What is the ultimate goal of the forecast? To minimize the forecasting error? Predict the sign of the return? Or the stock movement?

10.4 Importance of forecasting

This thesis highlights the importance of the forecasting model and its economic significance. VAR (6) in this research was able to predict correct sign of the future stock return (will there be a negative return or positive), with 67% accuracy, which was statistically significant result. This has important implications for the professional traders, as one can create a trading strategy based on this prediction and possibly realize an excess return.

Furthermore, all models produced significant results of accurate direction prediction for the S&P500 return. 75-80% of the time the models predicted correct change in S&P500 return. This result is even more noteworthy for the professionals in finance as one may not be interested in the exact value of the return, but whether it is forecasted to increase from today's one or decrease, as traders can perform an investment strategy that invests in the stock index when the forecast predicts an increase in return and sell the shares when the forecasts indicates a decrease in returns. Moreover, this result is even more meaningful given the use of AI and algorithms, as one can create this strategy that performs trades automatically (such as algorithmic trading firms described in this paper) or to program the

algorithm that would send a notification to the trader of what change is predicted for the stock index. This will consequently reduce the amount of time that a human trader needs for the investigation of the matter and the search of investment opportunities.

10.5 Usefulness of Big Data in predicting stock returns

While Big Data is still considered as a black box in a financial area, given its debatable base to the traditional finance theories, it proves to be useful in forecasting stock market returns in this study.

This paper found Google term “S&P500 index” to be a significant factor in modelling S&P500 return, measured by the statistical significance of its coefficients during VAR in-sample model estimation, Granger Causality tests and Impulse response. Moreover, lags of this Google trend variable are also significant, thus, have a predictive power of the S&P500 return. The result is in accordance with the expectation that Google trend data, being a proxy for a market sentiment, will have a predictive power on stock index return. (This finding is also in accordance with Preis et al. (2013) that have performed a study using different google trend searches and found some significant terms for predicting Dow Jones Industrial index).

The argument against the use of Big data in traditional finance research has been that the method is based on a data-mining and thus is not related to any of the traditional finance theories. However, one can argue that the use of “S&P500 index” in predicting S&P 500 index is a proxy for a market sentiment, i.e. how much anxiety there is in the market, which is in our digital age is given by how many people google the exact same term over a certain period of time in a certain location. Being representative of a market sentiment, the use of this variable is justified by the behavioral finance.

The explanation for the exact nature of the association between Google searches and stock return is not clear, one plausible theory could be that market is nervous about the future, searches online for the information about a stock market index, and the holders of the stock sell it, causing further herding among the market participant and a further price deterioration (hence negative returns). Due to these behavioral biases, the stock return varies disproportionately, hence allowing for a time-varying abnormal return to exist. (In accordance with Andrew Lo’s hypothesis of a dynamic market ecosystem with crashes, bubbles, hysteria and time-varying abnormal returns).

This finding is statistically significant and it is, to the author's best knowledge, the first study that uses "S&P500 index" Google Trend search for predicting S&P500 return for this time period. Additional benefit of this study is that this result is very current (using observations till May 2017). For the professional traders, this means that the study of Big data deserves more attention than it gets currently. Furthermore, it is important to identify what particular aspects of large amount of data is especially relevant and useful for the particular task. The search for significant factors could be more efficiently performed by an algorithm, thus giving a finance professional another reason for its investigation.

10.6 Non-linear vs linear

This thesis discussed the difference between linear and non-linear models for forecasting stock returns. Clements et al. (2014), in evaluating the models, suggest that there is no ultimately a model that would be superior in the forecasting accuracy in all problems and datasets. He states that for some tasks and data types, non-linear models are more efficient as they are able to capture non-linear relationship in the data, whereas for other problem types, where the nature of the association between the variables is linear, non-linear models would produce worse forecasts than its linear counterparts. The results of this research are in accordance with his findings, as for certain problems (prediction of sign or direction) neural networks, being non-linear models, performed better than its linear competitors, whereas in other parameters such as minimizing the absolute forecasting error, networks underperformed.

For researchers and professional traders this implies that non-linear forecasting methods should receive more attention and exploration. Moreover, the use of non-linear models should be combined with the use of linear models, which would provide two different perspectives to the forecast.

10.7 Caution in network configuration

This paper has extensively tested various neural network configurations, such as number of lags, training algorithms and number of input variables. The range of the results was quite large, meaning that there were networks that performed exceptionally well, and there were others that underperformed. This brings the risk if a professional only uses one iteration of the neural network and does not test for the sensitivity.

Therefore, one of the suggestions would be not to use only one network's forecast, but to combine it with another forecast (produced either by linear or non-linear methods) and to check for the robustness of the result. Another suggestion is to test for the network robustness, given different number of lags and training algorithm. Furthermore, there are many other networks types, each of them would have an extensive specification and configuration options, therefore, allowing for the numerous number of individual networks. This is even more acute given the current research into the unsupervised learning, which should be able to remove any human bias from the algorithm, thus unleashing new opportunities for the AI application in finance.

10.8 Statistical versus Economic measures

This research have used both statistic measures for the evaluation of the forecast accuracy, such as MSPE and MAPE, and economic measures, such as Success ratio and Direction ratio. The objective of these measures is very different. The former measures the forecasting error, namely the difference between predicted and actual forecast, and suggests picking the model with the smallest error (similar to Mean Square Error and Mean Absolute Error), however, the latter is focused not on the actual level of the forecasted series, but on the correct movement prediction (suggesting that the model with the highest percentage of correct prediction is the best).

Given this fundamental difference in the objective of measuring a forecast accuracy, the measures choose different models. This is evident in the evaluation of the results, where the model that has the lowest forecasting error was not necessarily the one that predicted the movement the best.

Therefore, it is suggested to use both measures for the evaluation of the forecast and to allow the user to decide which measure of a forecast accuracy is more applicable for his/her question.

11 CONCLUSION

This thesis aimed at researching various artificial intelligence application in finance and found that most common areas for AI application is portfolio management, bankruptcy prediction, credit rating, exchange rate prediction and trading. Using artificial neural networks as a proxy for Artificial Intelligence, this thesis conducted a study comparing the performance of ANN (as data driven, non-linear models, that use learning algorithms to learn from the data) to the performance of models that have existed for a long time in the area of forecasting, such as ARIMA and VAR models. Even though some models were found to be superior than other (for example, artificial neural network with 6 lags had a higher accuracy forecast than VAR (6) model) the overall result is inconclusive. While some methods outperformed in certain measures, they have underperformed in others. Moreover, S-tests, measuring the significance of the difference in the forecast accuracy, indicated that there was no model that would be consistently and significantly better in forecasting. Nevertheless, all models appear to have good forecasting accuracy in the prediction of the stock return directions, as they predicted it with 75%-80% accuracy. This is arguably a more important result for the investor as the exact price level of the stock index might not be of that much interest, as an accurate prediction of the stock market movement.

11.1 Implications

Theoretical

This study has tested a new proxy for a market sentiment and investor's behavior, - Google trends data. These variables are easy to extract and they might be very useful in understanding particular concerns or interests of the market. This could be applied into many research questions, for example if one wants to investigate how much 1st day premium will the company receive in an IPO, he/she could check for the market interest using the Google trends search terms. It is expected that the more hype there is in the market about the forthcoming IPO, the higher will be the premium. Moreover, the use of Google trends data can be argued to have a base in the behavioral finance and is in accordance with Andrew Lo's Adaptive Market Hypothesis.

Managerial

This research finds that many companies employ Artificial Intelligence in their work, and the applications of AI in finance are very diverse, varying from trading to loan issuance, and

keep on expanding. For the professionals, it implies to invest resources into the explorative studies of artificial intelligence applications that are useful for their work. Furthermore, for the professionals it also means to re-identify the task that the person does, his or her contribution and whether his/her skills can be easily replicated by an algorithm. If, after this evaluation, the professional finds himself at a risk of being obsolete in the near future, one should think of how he/she can embrace the forthcoming change and try to adapt to it, thus making themselves more useful in the changing setting.

If one decides to use neural networks for forecasting, the recommendation is to use it with caution and to test it for the robustness of the result, given different network configuration and learning algorithms. It is advisable to use it, but it would be even more beneficial to combine it with other forecasts in order to get an overall perspective from various models (unless a network that consistently and significantly outperforms the other methods is found, using S- and Direction Accuracy tests).

11.2 Limitations

This research is limited to the applications of recurrent dynamic type of neural networks in forecasting stock returns, thus the interpretations should only be made with regards to the performance of these networks over a specified time period in USA. Moreover, various networks differ in their configuration, settings and trainings, therefore, this result applies only for the NAR and NARX recursive dynamic networks, since the results of different network settings could be different.

Geographical limit of this research was USA stock market. It is possible that other markets, especially the ones in emerging markets where the return behaves more abnormally, would be different. The performance of neural networks, being non-linear models, depends greatly on the chosen data, therefore, it is possible that the performance of ANN could be better using other set of data (that would have more non-linear associations between the return and explanatory variables).

With regards to the methodology, this research had some further limits. Firstly, this study is limited to investigation of ANN versus linear models, such as ARMA and VAR models. It is possible that when one compares ANN to non-linear forecasting models, the result could be different. Furthermore, the forecasting accuracy assessment was limited to the statistical and economic measures, such as MSPE, MAPE, Success ration and Direction prediction.

However, the financial significance of this result has not been determined (this could be done by simulating an investment strategy that would buy and sell index based on the predictions) and it is possible that a financial measure could choose an altogether different model than the other measures. Additionally, only one-step ahead forecasting has been performed. Some researchers found that non-linear models outperform linear counterparts only in the long-term forecasting (Kalliovirta et al. 2016), therefore, if one chooses to perform different time horizon forecasting, it is highly likely that the result will be different.

Moreover, the choice of variables that affect stock returns is somewhat an open area for the investigation as there are many potential factors that were once significant in some studies, but became insignificant, or do not hold for another time period, or the investors have learnt about it and adjusted their trading behavior thus reducing the return predictability. This means that there are many other variables, that one can test for in the further research, and the result might differ from this study.

Additionally, the question of AI usefulness in a financial area in this study was limited by the analysis of neural network application, however, this methodology is lacking the power to explain or answer the question fully. In order to do that, it is also necessary to perform the tests that would compare the performance of the companies that employ AI in their work versus the performance of the companies without it (for example, the returns of AI funds versus the returns of human-managed funds). However, this research is limited to the data availability and survivorship bias, as, firstly, companies that employ AI might not want to publish their results in order not to destroy the shareholder's value or brand image if the returns are low, but also not to attract competition by disclosing high profitability. Furthermore, the companies that will disclose the information are subjected to the survivorship bias, given the fact that the performance only of those companies that were willing to provide the information will be available, thus it would not be a representative sample of the population (=all companies that apply AI).

11.3 Suggestion for further research

The area of forecasting and Artificial Intelligence applications is infinite, especially with constant development of new methods and models. This opens up vast opportunities for further research.

First suggestion comes from the forecasting methods, as this research have utilized one-step-ahead forecasting method, however there are many other forecasts that one can test for, for example 5-steps or 20-steps ahead forecast. The desire for the accurate long-term forecasting is self-evidently very important to the finance professional. Therefore, the combination of one-step and multi-steps forecasts could be used for making a more profitable investment decision. For example, if one step ahead forecast predicts an up movement, but 5-steps-ahead forecast predicts a significant down movement, the investor will be less willing to follow the one-step-ahead forecast and will not rush into making an investment.

Secondly, this research has used unrestricted VAR model, which means that all variables had predetermined number of lags. Having a VAR model with 2 lags and 12 variables already makes it 48 lag coefficients to estimate. However, the time-series dynamics of each variable is not the same, therefore, some variables could use a longer lag order, to capture some of the dynamics left in the data, while the number of lags in other variables, that are not useful after a certain lag, could be reduced, thus minimizing the number of coefficients that the model has to estimate, consequently possibly improving the forecast accuracy. This could be done by using search algorithm with some information criteria function that would determine the optimal number of lags for each variable, and using this information one could create a restricted VAR model. It would be interesting to test the accuracy of the forecasts between these models.

From the neural network research, there are plenty of suggestions and recommendations for the further research. Firstly, it would be a good idea to use more observations. A larger data-set would allow for much more data points in a neural network, thus possibly improving its accuracy. Secondly, this research has used only NAR and NARX recurrent dynamic networks, but given the wide diversity of artificial neural networks, it would be interesting to use various types of ANNs to identify the best network for predicting the return and compare it with the traditional forecasting methods.

Moreover, neural networks in this study were trained with LM, SCG and BR training algorithms, but it is possible that other training algorithms could provide the network with better estimations and higher accuracy of the forecasts. For example, one promising training algorithm is Extreme Learning Machine which provides an exceptionally fast generalization of the network while yielding good results, which might be a good fit for the large data file.

Furthermore, the output for the network was specified to be in the actual level, however, the optional modification would be to change the transfer function of the output layer to create the output in the form of likelihood probability of the return change (from -1 to 1; or from 0 to 1, meaning that the higher value (closer to 1) would indicate a higher probability of the stock return to increase in the forecasting period), which could be then tested for the accuracy using Success Ratio or Direction ratio (how often it correctly predict the sign or movement of the return). Additionally, there has been a lot of promising research on the potential of the AI algorithms that do not require a supervised learning, thus allowing to keep human biases completely out of the data processing. Therefore, it would be worthwhile to create a network or machine learning algorithm that would use unsupervised learning and to analyze the forecasting accuracy of such algorithms.

During the writing of this thesis, the author got familiar with other non-linear forecasting methods for stock returns than artificial neural networks, such as regime switching, threshold models and LSTAR models (Smooth Transition Autoregressive model), all of which received a lot of attention with regards to their ability to forecast the time series with arguable accuracy. Therefore, comparing the performance of artificial neural networks not only with linear models but also with non-linear forecasting models, would provide an interesting dimension to the future research. Furthermore, some previous research papers have indicated that a hybrid of two models (ANN and a traditional one) is often received the best forecasting accuracy results (Zhang, 2003, Maia & Carvalho, 2011). Therefore, it is suggested that after the identification of the two best model, one can combine their forecasts (following Timmermann(2006) methodology) to check whether the hybrid surpasses the performance of the individual models.

Lastly, this thesis has tested the forecasting accuracy using statistical measures of forecast accuracy such as MSPE and MAPE, and economic measures such as Success ratio (or sign prediction) and Direction prediction. For future research, it would be beneficial to also include a financial measure of the forecasting accuracy such as a simulation of the investment strategy, which would involve investing a hypothetical sum of money into the stock when the return is forecasted to be positive and to sell when it is predicted to be negative, and compare the rate of return generated by this strategy (after the effect of the transaction costs) to returns from a buy-and-hold strategy or to returns generated using the forecasts of another models.

REFERENCES

- AlphaSense (2017). *Intelligent Search. Why AlphaSense?* [online]. Available at <https://www.alpha-sense.com> [Accessed on 20 August 2017]
- Andrawis, R.R., Atiya, A.F. and El-Shishiny, H. (2011). Forecast Combinations of Computational Intelligence and Linear Models for the NN5 Time Series Forecasting Competition. *International Journal of Forecasting*. Vol 27 (3) pp 672-688
- Baele L., Bekaert G., Inghelbrecht K. (2014). Flights to Safety. *Finance and Economics Discussion Series, Federal Reserve Board*. [online] Available at <https://www.federalreserve.gov/pubs/feds/2014/201446/201446pap.pdf>
- Bahrammirzaee, A. (2010). A Comparative Survey of Artificial Intelligence Applications in Finance: Artificial Neural Networks, Expert System and Hybrid Intelligent Systems. *Neural Computing & Application* 19: 1165-1195.
- Betterment (2017). *This is simply a smarter way to invest your money*. Available at <https://www.betterment.com>
- Binner, J.M., Kendall G. and Chen S.H. (2004). Introduction. *Applications of Artificial Intelligence in Finance and Economics* 19, 9-12
- Bollerslev, T., Engle R.F. and Wooldridge, J.M. (1988). A Capital Asset Pricing Model with Time-varying Covariances. *Journal of Political Economy* 96 (1) 116-131
- Box, G. E. P. and Jenkins, G. M. (1970, revised ed.1976). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.
- Brooks, C. (2008). *Introductory Econometrics for Finance*. 2nd edition. London: Cambridge University Press.
- Brown, G.W., Cliff, M.T. (2005). Investor Sentiment and Asset Valuation. *The Journal of Business*, vol. 78 (2), 405-440.
- Campbell, J. Y. and Thompson, S.B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies* 21(4) 1509-1531.

- Cerebellum Capital (2016). Our Mission. [online] Available at <http://www.cerebellumcapital.com/about.html> [Accessed on 4 April 2017]
- Ceron, A., Curini, L., Iacus, S.M. (2015). Using Sentiment Analysis to Monitor Electoral Campaigns: Method Matters – Evidence From the United States and Italy. *Social Science Computer Review*, vol. 33 (1), 3-20.
- Cheng, I. H., Hong H. , and Scheinkman J.A, 2015, “Yesterday’s Heroes: Compensation and Risk at Financial Firms,” *Journal of Finance*, 70, 839-879.
- Clements M.P, Franses P.H. and Swanson N.R. (2004). Forecasting economic and financial time series with non-linear models. *International Journal of Forecasting* 20, 169-183
- Coval, J.D. and Moskowitz, T.J. (1999). Home Bias at Home: Local Equity Preference in Domestic Portfolios. *The Journal of Finance* 54 (6) 2045-2073.
- Culp, S. (2017) *Artificial Intelligence Is Becoming A Major Disruptive Force In Banks' Finance Departments*. [online] Forbes. Available at: <https://www.forbes.com/sites/steveculp/2017/02/15/artificial-intelligence-is-becoming-a-major-disruptive-force-in-banks-finance-departments/#51484b5d4f62> [Accessed 5 Mar. 2017].
- Dataminr (2017) *About Us. The World’s Real-time Information Discovery Company*. [online]. Available at <https://www.dataminr.com/about> [Accessed on 12 August 2017]
- De Gooijer, J.G. and Hyndman, R.,J. (2006). 25 years of time series forecasting. *International Journal of Forecasting* 22, pp 443-473
- De Matos Neto et al (2017). A Perturbative Approach for Enhancing the Performance of Time Series Forecasting. *Neural Networks* 88. 114–124
- Diebold F.X. and Mariano R.S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13(3) 253-263
- Ding S., Zhao H., Zhang Y., Xu X. and Nie R. (2013). Extreme Learning Machine: Algorithm, Theory and Applications. *Artificial Intelligence Review* 44, 103-115

- Elliott G. and Timmermann A. (2016). Forecasting in Economics and Finance. *Annual Review of Economics* 8, 81-110
- Faggella, D. (2016). *Machine Learning in Finance – Present and Future Applications*. [online]. Available at <https://www.techemergence.com/machine-learning-in-finance-applications/>
- Fama E.F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *Journal of Finance* 25 (2), 282-417
- Fama, E.F., (1981). Stock Returns, Real Activity, Inflation, and Money. *American Economic Review* 71, 545 – 565.
- Fama E.F and French K.R. (1988). Dividend Yields and Expected Stock Returns. *Journal of Financial Economics* 22(1), 3-25
- Fama E.F and French K.R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics* 33, 3-56
- Fan, A. and Palaniswami, M. (2001). Stock Selection using Support Vector Machines. *Neural Networks*. Vol. 3, pp. 1793-1798.
- Federal Reserve Bank of St. Louis (2017). *Federal Reserve Economic Data. FRED Graph Observations*. [online]. Available at <https://fred.stlouisfed.org/series/AAA#0>
- Franses P.H. and van Dijk, D. (2003) *Non-linear Time Series Models in Empirical Finance*. Cambridge University Press
- Goyal, A. and Welch, I (2008). A Comprehensive Look at The Empirical Performance of Equity Premium Prediction. *Review of Financial Studies* 21(4) 1455- 1508
- Grossman, S. (1976). On The Efficiency of Competitive Stock Markets Where Trades Have Diverse Information. *Journal of Finance* 31, 573-585.
- Grossman, S. and Stiglitz, J. (1980). On the Impossibility of Informationally Efficient Markets. *American Economic Review* 70, 393-408.

- Harmon, D., Lagi, M., de Aguiar, M., Chinellato, D., Braha, D., Epstein, I. (2015) Anticipating Economic Market Crises Using Measures of Collective Panic. *PLoS ONE*, vol. 10 (7).
- Hodrick R.J. (1992) Dividend Yields and Expected Stock Returns: Alternative Procedures for Inference and Measurement. *The Review of Financial Studies* 5(3) 357-386
- Huang G. B., Zhu Q.Y. and Siew C.K (2006). Extreme Learning Machine: Theory and Applications. *Neurocomputing* 70, 489-501
- Hwarng (2001) Insights into Neural-network Forecasting of Time Series Corresponding to ARMA(p; q) Structures. *Omega 29 The International Journal of Management Science* 29, 273–289
- IDC (2016). *Worldwide Semiannual Cognitive/Artificial Intelligence Systems Spending Guide*. [online] International Data Corporation. Available at: <http://www.idc.com/getdoc.jsp?containerId=prUS41878616> [Accessed 5 Mar. 2017].
- iSentium (2017) *Breakthrough Sentiment Analytics*. [online] Available at <http://isentium.com> [Accessed on 10 June 2017]
- Kahneman D. and Tversky A. (1979). Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2), 263-291
- Kalliovirta, L., Meitz M. and Saikkonen P. (2016). Gaussian Mixture Vector Autoregression. *Journal of Econometrics* 192 (2) 485-498.
- Khashei M. and Bijari M. (2010) An Artificial Neural Network (p, d, q) Model for Time Series Forecasting. *Expert Systems with Applications* 37 479–489
- Kensho (2017) Global Analytics. [online] Available at <https://www.kensho.com> [Accessed on 28 August 2017]
- Kitco Metals Inc. (2017). *Historical Charts & Data. Gold*. [online]. Available at <http://www.kitco.com/charts/historicalgold.html>
- Kriesel, D. (2007). *A Brief Introduction to Neural Networks*. [online] Available at http://www.dkriesel.com/en/science/neural_networks

- Kumar S. and Goyal N. (2014). Behavioral biases in investment decision making – a systematic literature review. *Qualitative Research in Financial Markets* 7, 88-108
- Kumar, M. (2009). Nonlinear Prediction of the Standard & Poor's 500 and the Hang Seng Index under a Dynamic Increasing Sample. *Asian Academy of Management Journal of Accounting and Finance*, vol. 5 (2), 101-118.
- Kuorentzes, N., Barrow. D.K., Crone, S.F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41 (9), 4235-4244.
- Kutner, M. H., Nachtsheim, C. J., Neter, J. (2004). *Applied Linear Regression Models* (4th ed.). McGraw-Hill Irwin.
- Lo, A. (2004). The Adaptive Market Hypothesis: Market Efficiency from an Evolutionary Perspective. *Journal of Portfolio Management* 5 (30) 15-39
- Maia A.L.S and Carvalho F.A.T (2011). Holt's Exponential Smoothing and neural network models for forecasting interval-valued time series. *International Journal of Forecasting* 27 750-759
- Maskay B. (2007), "Analyzing the Relationship between change in Money Supply and Stock Market Prices. *The Park Place Economist* 15, 72-79
- Mastercard (2016). *Press Releases: MasterCard Rolls Out Artificial Intelligence Across its Global Network*. [online] Available at: <http://newsroom.mastercard.com/press-releases/mastercard-rolls-out-artificial-intelligence-across-its-global-network/>
- Moat, H.S., Curme, C., Avakian, A., Kenett, D.Y., Stanley, E., Preis, T. (2013). Quantifying Wikipedia Usage Patterns Before Stock Market Moves. *Scientific Reports*, 3 (1801)
- Moskowitz T.J., Ooi Y.H. and Pedersen L. H. (2012). Time series momentum. *Journal of Financial Economics* 104, 228-250
- MATLAB Neural Network Toolbox (2016).For Use with MATLAB;[user's Guide]. Math-Works
- Niaki S. T. A. and Hoseinzade S.(2013). Forecasting S&P 500 index using artificial neural networks and design of experiments. *Journal of Industrial Engineering International* 9 (1). 1-9

- Nomura Research Institute (2015). *Utilization of artificial intelligence in finance*. [online]
Available at :
<https://www.nri.com/~media/PDF/global/opinion/lakyara/2015/lkr2015227.pdf>
- Odean, T. (1998). Are investors reluctant to realize their losses? *The Journal of Finance* 53 (5) 1775-1798.
- Odean, T. (1999) Do investors trade too much? *The American Economic Review* 89 (5) 1279-1299.
- Olsen, R.A. (1998). Behavioural Finance and its Implications for Stock-Price Volatility, *Financial Analysts Journal*, vol. 54 (2), 10-18.
- Olson, D., and Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting* 19 (2003) 453–465
- Oztekin A., Kizilaslan R., Freund S. and Iseri A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research* 253, 697-710.
- Panda & Narasimhan (2007). Forecasting exchange rate better with artificial neural network. *Journal of Policy Modeling*. 29 (2) Pages 227–236
- Pannu, A. (2015). Artificial Intelligence and its Applications in Different Areas. *International Journal of Engineering and Innovative Technology*. Vol 4(10) 79-84
- Pesaran M.H and Timmermann A.G. (1992). A simple Non-Parametric Test of Predictive Performance. *Journal of Business and Economic Statistics* 10(4), 461-465
- Pesaran M.H and Timmermann A.G. (1995). Predictability of Stock Returns: Robustness and Economic Significance. *Journal of Finance* 50 (4) 1221-1228
- Ross, S. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory* 13, 341-360.
- Preis, T., Moat, H.S., Stanley, H.E. (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. *Scientific Reports*, 3 (1684).
- Russell S.J., and Norvig, P.(1995). *Artificial Intelligence. Modern approach*. New Jersey: Prentice Hall, Englewood Cliffs

- Schwab Intelligent Portfolios (2017). Investing made easier. [online]. Available at <https://intelligent.schwab.com> [Accessed 5 Mar. 2017].
- Sellin, P. (2001). Monetary Policy and the Stock Market: Theory and Empirical Evidence. *Journal of Economic Surveys*, 200 , 5 (4), pp. 49 - 54 .
- Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *Journal of Finance* 19, 425-442
- Sentient Investment Management (2017). Answering Critical Questions. [online]. Available at <https://www.sentient.ai/our-story/> [Accessed 16 Jun. 2017].
- Sheta, A.F., Ahmed, S.E.M., and Faris, H. (2015). A Comparison Between Regression, Artificial Neural Networks and Support Vector Machines for Predicting Stock Market Index. *International Journal of Advanced Research in Artificial Intelligence*, vol. 4 (7).
- Shiller, R. (2017). *U.S. Stock Markets 1871-Present and CAPE Ratio*. [online]. Available at <http://www.econ.yale.edu/~shiller/data.htm>
- Teräsvirta T., van Dijk D. and Medeiros M.C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting macroeconomic time series: A re-examination. *International Journal of Forecasting* 21, pp 755-774
- The Regents of the University of Michigan. (2017). *Surveys of Consumers*. [online] Available at <http://www.sca.isr.umich.edu/files/tbmics.pdf>
- Timmermann, A. (2006) *Forecast combinations*. Published in G. Elliott, C.W.J. Granger, A. Timmermann (Eds.), *Handbook of economic forecasting*, Elsevier Pub. 135–196
- U.S. Bureau of Labor Statistics (2017) Unemployment rate. [online] Available at <https://data.bls.gov/cgi-bin/surveymost?bls>
- Von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press
- Wang, X. and Yang, B. (2012). Yield Curve Inversion and the Incidence of Recession: A dynamic IS-LM Model with Term Structure of Interest Rates. *International Advances in Economic Research* 18 (2) 177-185

- Zarnoth, P. and Sniezek, J. A. (1997). Social Influence of Confidence in Group Decision Making. *Journal of Experimental Social Psychology* 33(4) 345-366
- ZestFinance (2017). *Machine Learning & Big Data Underwriting*. [online] Available at <https://www.zestfinance.com> [Accessed on 20 June 2017]
- Zhang (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50 159-175
- Zhongzhi, S. (2011). *Advanced Artificial Intelligence*. Singapore: World Scientific Publishing Company.

APPENDIX 1 FULL CORRELATION MATRIX FOR DAILY DATASET

[illegible][illegible]