CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Bachelor thesis

2018 **Barbora Veselá**

CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Barbora Veselá

Top Stocks: A Broad Analysis of Its Performance and Search for Hidden Relationships

Bachelor thesis

Prague 2018

Author: Barbora Veselá

Supervisor: Mgr. Petr Polák, MSc.

Academic Year: 2017/2018

Bibliographic note

VESELÁ, Barbora. *Top Stocks: A Broad Analysis of Its Performance and Search for Hidden Relationships* Prague 2018. 62 pp. Bachelor thesis (Bc.) Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor Mgr. Petr Polák, MSc.

Abstract

This thesis focuses on the exploration of the basic characteristics of the Czech equity fund 'TOP STOCKS - open mutual fund' and a detailed analysis of the portfolio performance with respect to industry classification. Uniquely collected data from the fund's establishment allow us to make analysis to the fullest extent possible. Additionally, our dataset may be the basis for further studies of this fund as it is the first of its kind and scope. The research question investigates the connection between the portfolio performance and its industrial structure. Various classification schemes are summarised in the first part of the thesis, including the Global Industry Classification Standard used in our study. Subsequently, appropriate tools for a regression and forecast analysis are presented, mainly the Box-Jenkins method used to fit the ARIMA model to historical data and forecast values of weekly NAV. The results show that a stock-picking strategy operates effectively and immediately reacts to the market development of industry groups, resulting in a protection of investors. Also, the exchange rate commitment of the Czech National Bank supported excessive cash inflow to the fund resulting in considerable performance stimulation during last years. Future research may build on these findings.

Keywords

fund analysis, GICS structure, industry groups, assets under management, financial markets, Box-Jenkins, ARIMA

Abstrakt

Tato práce se zaměřuje na zkoumání základních charakteristik českého investičního fondu „TOP STOCKS - otevřený podílový fond“ a detailní analýzu portfolia dle sektorového členění. Unikátně získaná data od počátku fungování tohoto fondu umožňují provést analýzu v nejhlubším možném rozsahu a soubor dat sestavený autorem této práce může být výchozím souborem pro další studie tohoto fondu, neboť je první svého druhu a rozsahu. Hlavním cílem této práce je hledání souvislostí mezi výkonností fondu a jeho průmyslovou strukturou. V první části práce shrnujeme historický vývoj sektorového členění, včetně vedoucího schématu GICS, kterým se řídíme v této studii. Následně jsou představeny vhodné nástroje pro naši analýzu, především Box-Jenkinsnova metoda pro volbu ARIMA modelu vzhledem k historickým datům, který je následně použit pro předpověď dalšího vývoje ceny podílového listu fondu. Bylo zjištěno, že stock-picking strategie je řízena efektivně a umožnuje flexibilně reagovat na tržní vývoj jednotlivých sektorů, čímž chrání investory před potencionálními ztrátami. Analýza dále odhalila pozitivní dopad intervencí České národní banky, které výrazně zvýšily atraktivitu i výkonnost fondu. Na těchto poznatcích může stavit budoucí výzkum.

Klíčová slova

analýza fondu, GICS členění, průmyslová struktura, objem aktiv, finanční trhy, Box-Jenkins, ARIMA

Declaration of Authorship

The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

The author hereby declares that all the sources and literature used have been properly cited.

The author hereby declares that the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University a permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, 11 May 2018

Signature

Acknowledgments

Here, I would like to show gratitude to Supervisor of this thesis, *Mgr. Petr Polák, MSc.*, for patient guidance from the very first idea, through data analysis to finalising this thesis. His valuable suggestions and knowledge helped me immensely during the completion of this thesis.

The Bachelor's Thesis Proposal

Author	Barbora Veselá
Supervisor	Mgr. Petr Polák, MSc.
Proposed topic	Top Stocks: A Broad Analysis of Its Performance and Search for Hidden Relationships

Research question and motivation

A unique stock picking strategy performed by professionals allows investors to evaluate savings on stock markets with lower risk and without the need for time-consuming analysis. The volume of managed assets of the Czech investment fund 'Top Stocks' has exceeded 10 billion Czech crowns and it is still growing. Even though there are many publicly available data sets, only a complex analysis with understandable interpretation of results can provide investors with reliable information to facilitate the investment decision-making process and help them not to get lost in the details. Currently, there is no comprehensive fund overview that would provide investors with enough information to select an appropriate investment strategy.

Contribution

The overall contribution shall be a creation of such an overview and finding hidden relationships in portfolio performance. Results and conclusions of this thesis will help both ordinary investors with no deeper knowledge of this issue, as well as investment specialists. Moreover, it could be transformed into an academic paper and further developed.

Methodology

Volatility of stock funds is common in the short run and is influenced also by random events, which makes short-run analysis complicated and relatively insignificant. Therefore, it is crucial to build on long-term developments and use the data since fund establishment in our analysis. Fund performance (on a daily, weekly and monthly basis) will be analysed with respect to geopolitical events, change in Global Industry Classification Standard (GICS) portfolio structure and currency diversification. Using suitable statistical tools, I will also look for possible trends in the number of traded units.

Outline

1. Introduction
2. Motivation and literature review
3. Theoretical background
4. Data description and methodology
5. Evaluation of results
6. Conclusion

Core Bibliography

1. BARRO, R. J. and URSÚA, J. F. (2009). Stock Market Crashes and Depressions. *Harvard University*.
2. CAMPBELL, J. Y., LO, A. W. and MACKINLAY, A.C. (1997). The Econometrics of Financial Markets. *Princeton University Press*, New Jersey.
3. FAMA, E. F. and FRENCH, K. R. (1988). Permanent and Temporary Components of Stock Prices. *Journal of Political Economy*, 96(2):246–273.
4. LO, A. W. and MACKINLAY, A. C. (1988). Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test. *Review of Financial Studies*, 1: 41-66.
5. MALKIEL, B. G. (2003). The Efficient Market Hypothesis and Its Critics. *Journal of Economic Perspectives*, 17(1):59-82.
6. LÁSTOR, L. and STAMBAUGH, R. F. (2008). Are Stocks Really Less Volatile in the Long Run?
7. THALER, R. H. (1987). Anomalies: The January Effect. Economic Perspectives, 1(1):197-201.

Contents

1 INTRODUCTION 2

2 GENERAL INFORMATION 4

3 LITERATURE REVIEW 7

3.1 INDUSTRY CLASSIFICATION SCHEMES 7
3.1.1 SIC Codes 7
3.1.2 NAICS Codes 9
3.1.3 Fama-French Industry Classifications 10
3.1.4 GICS Codes 11
3.2 APPROACH OF THIS THESIS 12

4 METHODOLOGY 13

4.1 DESCRIPTIVE STATISTICS 13
4.2 GRAPHICAL ANALYSIS 14
4.3 CALCULATIONS 17
4.4 REGRESSION AND FORECAST ANALYSIS 20

5 DATA 26

5.1 DATA AVAILABILITY AND COLLECTION 26
5.2 DATASET DESCRIPTION 28

6 RESULTS 31

6.1 DESCRIPTIVE STATISTICS AND GRAPHICAL ANALYSIS 31
6.2 REGRESSION AND FORECAST ANALYSIS 47
6.3 LIMITATIONS AND DISCUSSION 53

7 CONCLUSION 54

BIBLIOGRAPHY 57

LIST OF APPENDICES 62

1 Introduction

The Czech equity fund 'Top Stocks' managed by the Czech branch of Erste Asset Management GmbH, belongs to the most successful open-end mutual funds in the Czech Republic. It is actively managed using a unique *stock-picking strategy*, which is performed by a portfolio manager, Ing. Jan Hájek, CFA, and his team.

This strategy offers a continuous market analysis and a constant search for exceptional investment ideas. As oppose to passive investment strategies represented by conservative portfolio mixes, treasury bonds, etc. accompanied by almost zero yield potential today, actively managed funds are represented by highly concentrated portfolios and above-average returns.

At the time of low interest rates caused by the Czech National Bank's (CNB) exchange rate commitment effective from November 2013, people searched for alternative options how to appreciate their savings, as CNB interventions pushed interest rates on savings accounts to virtually zero. Mutual funds represent the easiest option how to enter the investment market without the need for deep knowledge of its environment, as the process is fully managed by clients' banks. In addition, they attract the attention of investors by having relatively high liquidity. Consequently, assets under management of 'Top Stocks' exhibit an exponential trend in recent years and its strategy seems to provide sufficient assurance of capital appreciation while carrying a reasonable degree of risk.

The objective of 'Top Stocks' is the long-term appreciation of mutual fund units (fund share unit price) by investing in a concentrated portfolio of equities traded on developed stock markets. Concentrated portfolio consists of 25 investments on average, where one investment idea usually means investing in one selected company. The proportion of shares and investment instruments carrying the risk of shares on the portfolio varies from 80% to 100%. In addition to shares, funds may also be invested in bonds, treasury bills and bank deposits.

This thesis focuses on a fund's performance since its establishment in 2006 to February 2018, thus it also covers the whole period of the Global Financial Crisis and the following recovery of financial markets. The primary objective of this research is to find the relationship between the fund's performance and a portfolio structure based on the *Global Industry Classification Standard (GICS),* as there should be some trade-off between industries to mitigate the risk accompanying a highly concentrated portfolio, which is especially important at the time of stock market crashes.

The challenge here is to create the very first dataset containing GICS sectoral breakdowns for the whole period covered by this thesis by analysing monthly reports requested from a portfolio manager. To the best knowledge of the author, this uniquely collected data have never been subject to testing before.

Furthermore, the *total volume of assets* of 'Top Stocks' is deeply analysed in order to find hidden relationships and potential anomalies. The *net asset value (NAV)* representing the fund's share unit price is the connecting element for all the tests performed in this thesis. Even though many data vendors provide a collection of historical development of NAV for Czech investment funds, there is no dataset covering the whole period since the fund's establishment to this day. Merging all collected data with a significant author's effort to search for missing values results in a complete dataset, whose analysis is expected to bring relevant findings and trustworthy conclusions.

The aim of this bachelor thesis is to analyse the data included in our dataset, comment on their implications for investors, and use them to perform a forecast analysis of weekly net asset value using the Box-Jenkins method by fitting the $ARIMA$ model to historical data. Also, we will construct a correlation matrix to capture inter-industry relationships in the portfolio structure to uncover details of a stock-picking strategy.

This thesis is the first of its kind and scope, as 'Top Stocks' has never been publicly analysed to this extent. Our findings can support further research and, if precise enough, our conclusions can serve as a benchmark for other studies of this topic. In addition, there will always be a space for deeper investigation and usage of more suitable and advanced tools, so, this research may proceed continuously over the fund's existence.

The bachelor thesis is organised in the following manner: First, general information about 'Top Stocks' are presented in Chapter 2, including a brief description of a stock-picking strategy and additional information about the fund's operation and opportunities for investors. Second, Chapter 3 presents a literature review, especially previous researches' findings and conclusions of the analysis of the four most widespread industry classification schemes used in financial research. This chapter is essential to understand the added value of a created dataset and reasons for using such data in this research. Subsequently, methodology and theoretical foundations are introduced in Chapter 4. Then, data availability, collection and a dataset are presented in Chapter 5 to provide a reader with a brief insight into following analysis. In chapter 6, general descriptive statistics and empirical results are shown and discussed. Finally, Chapter 7 evaluates these results, concludes our findings and gives suggestions for future research.

2 General information

The open-end mutual fund 'Top Stocks'[1] has been established based on the Czech National Bank's resolution from 27. 6. 2006 and started its operation two months later under the Czech branch of Erste Asset Management GmbH. Its mutual fund units are offered in the Czech Republic and other member states of the European Union through Investment Instrument Asset Account at 'SERVIS24 Internetbanking' online platform, which represents a comfortable way to manage the investment process and monitor the development of the fund's performance.

The idea behind 'Top Stocks' comes from Česká spořitelna's list of recommended titles prepared by Ing. Jan Hájek, CFA, a member of a strategic analysis department at that time. This report included recommendations for investment in individual stocks and there was a very good response from clients to such analysis. However, there has been a demand for a fund, which will follow this investment strategy and provide clients with more convenient alternative to individual investment.

As a result, Ing. Jan Hájek, CFA, got the opportunity to work as a portfolio manager of 'Top Stocks', a fund following his unique investment strategy known as a *stock-picking*. This philosophy lies in a highly concentrated portfolio with a limited number of investment ideas rather than a widely diversified portfolio with large number of companies that generally carries a lower risk for investors. The idea behind is quite obvious at first sight but immensely difficult to apply in asset management: one cannot have unlimited number of great investment ideas.

In order to find exceptional companies, a portfolio manager and his team have to continuously search through a large number of titles, deeply look into their investment stories and see what the rest of the market still cannot see. Only a deep analysis can uncover the potential of investment. Generally, they are interested in long-term growth investment stories that are very specific – related to unique company's products or services. Next, they focus on a growth of basic variables such as revenues, profits and free cash flows. A signal for deeper analysis is a significant difference between expected long-term growth of these variables calculated by this team and market expectation. The detailed process of market analysis is out of scope of this paper, and, at the same time, it will probably never be a subject of a public research, as it can be viewed as a secret 'know-how' of a portfolio manager and his team.

[1] Full name: 'TOP STOCKS – open-end mutual fund'

As a result of this procedure, 'Top Stocks' belongs to a family of *actively managed funds*. These funds offer flexibility and continuous research of managers who allocate significant resources to market analysis in order to find companies with an exceptional performance.

These funds, however, are more risky than broadly diversified portfolios. The portfolio's overall performance and a trust of investors can be highly affected by the inappropriate choice of an investment idea. Based on Hájek (2017, investujeme.cz interview), all investment ideas have a stake in the portfolio of approximately 4%, so the impact of each on a portfolio's performance is statistically significant. However, Elton and Gruber (1997) have shown that portfolios containing up to 20 companies have sufficiently benefited from diversification and with each additional company, the positive effect of diversification diminishes and acts more like averaging of returns. A portfolio of 'Top Stocks' contains 25 investment ideas on average, which ensure a reasonable diversification as no strong correlation occurs between investment stories of these companies. A detailed analysis of historical development of the number of investment ideas in a portfolio is performed in section 5.1 Data availability and collection.

Active management and expensive research implies higher fees for investors, which is equally important for investors as the risk behaviour of the fund. Initial charge and management fees are set to 3.00% and 1.95% respectively. Hájek (2017, Finlord interview) sees the amount of management fees as the investment risk of each investment company. The necessary but not sufficient condition for the fund to be successful is a long-term performance above its benchmark, which attracts new clients and generates resources for future research. Internal benchmark[2] representing a *passive investment strategy* consists of MSCI USA (80%) and MSCI Europe (20%) indices, both covering approximately 85% of the free float-adjusted market capitalization in the US and 15 developed countries across Europe respectively. A highly concentrated portfolio of 'Top Stocks' and its unique stock-picking strategy significantly outperform this benchmark. Based on 2016 annual report, one-off investment in 'Top Stocks' yields 8.42% p.a. since the fund's establishment, whereas internal benchmark reached 6.54% p.a.

Ongoing charges (total expense ratio, TER) are equal to the ratio of the total expenses of the fund to the average monthly value of the fund's assets for the previous accounting period, which is 2.15% for a financial year 2016[3]. Thus, TER indicates how

[2] Official benchmark is not set for 'Top Stocks' based on annual reports
[3] TER for a financial year 2017 will be available after the financial audit is performed

the assets in the mutual fund are burdened by all the costs of the fund. Based on Collins and Duval's (2017) study '*Trends in Expenses and Fees of Funds, 2016*', there is a downward pressure on TER for both actively managed and indexed funds from investors. They have also found an empirical evidence of a negative relationship between the expense ratio and the total fund's assets. Historical development of TER for 'Top Stocks' shows the same dependence.

Soon after its establishment in August 2006, 'Top Stocks' had to face the Global Financial Crisis that burst out in the United States and significantly affected stock markets. The collapse of global financial markets did not only affect economic performance and financial stability, but the public's trust in banks, their products and the financial system itself has strongly declined. (Cruijsen et al. (2013)) Therefore, consequences of bank bailout have affected people's decision-making process and investment funds' performance for several years.

Such environment has essentially forced a portfolio manager to offer an investment alternative with attractive returns on investment, and, carrying a reasonable degree of risk at the same time, to fight a worldwide risk-aversion of ordinary investors. Knack and Keefer (1997) analysed the relationship between economic performance and trust. They found a positive correlation between these two variables based on the analysis of 29 market economies. Therefore, the process of regaining a trust of investors has been crucial for 'Top Stocks'. This has been further supported by the CNB exchange rate commitment in 2013, which increased the attractiveness of 'Top Stocks' offering significantly higher returns compared to traditional investment products suffering from low interest rates.

The last important characteristics of 'Top Stocks' is its investment horizon reported to be at least five years. The average turnover of the portfolio is once every six years. Naturally, some exceptions exist, especially when the company has been taken over and withdrawn from trading. As a result of a stock-picking strategy, there are approximately five changes in the portfolio each year.

The internal limit of 25 investment ideas in the portfolio is strictly adhered to by the portfolio manager. Therefore, when he comes up with a new investment idea and a portfolio has a full capacity, he has to exclude one company in order to include another investment idea. Even though he can easily justify having both in the portfolio, he has never done it. The reason is obvious: when he can enforce having one more over the limit, he can do the same for an unlimited number of great investment ideas. Soon, the strategy of a concentrated portfolio would fall. A detailed analysis is performed in a Section 5.1.

3 Literature review

3.1 Industry classification schemes

To make capital market researches more accurate and precise, there is a need for industry classification that allows us to create homogeneous groups of firms that can be compared across or within these groups with specific characteristics. When using such a classification, we can greatly scale down overall scope of the research by reducing a sample size of our dataset or eliminate bias that evolves when analysing two or more slightly different business concepts in order to detect industry effects. Then, we can obtain results that are more robust and can serve as a benchmark for future empirical studies.

General goal of any industry classification scheme is to sort entities into homogeneous groups in terms of financial, organisational and other business-related characteristics based on clearly defined rules for these sectors. However, the optimal level of aggregation of similar firms and differentiation between industries is highly questionable.

Naturally, there are various schemes available for different types of data analysis, as they do not always have same implications for financial research and practical purposes. They have evolved concurrently with industry development, as older schemes were not able to keep up with rapidly growing industries during last decades. In the following section, there is an overview and historical development of four major classifications and a literature review related to them.

3.1.1 SIC Codes

Interdepartmental Committee on Industrial Classification, established by the Central Statistical Board of the United States, introduced Standardized Industry Classification (SIC) in 1939 as a direct application of a List of Industries for (non)-manufacturing industries issued by the Board over the past two years.

SIC code is a 4-digit numerical code that enables us to divide all economic activities into a specific industry group. When simplified, first two digits are related to the major industry group, third digit identifies the industry group and the fourth digit determines the industry.

According to Pearce (1957), there was a need to classify financial and statistical data by industries and ensure a wide adoption of the system across the US in order to promote uniformity and comparability of collected data. During the first decades, SIC has become a primary tool to classify data and as a standard tool of the Federal Government, it has benefited from the highest availability.

Clarke (1989) performed a detailed study of SIC usefulness and limitations. His main goal was to answer the question whether SIC design and methodology leads to homogeneous groups of firms in terms of business-related characteristics. In his research, Clarke analysed the Compustat North America database over the 8-year period starting in 1975. Using a regression model, he came to conclusion that SIC cannot explain profit ratios, net asset value changes and changes in sales of companies well. However, he shows that a four-digit code can work more effectively when using just first two digits, as the major industry group acts homogenously, whereas more precise 3- or 4-digit grouping does not always meet homogeneity assumption.

Guenther and Rosman (1994) performed other test for homogeneity of industry groups by looking at stock returns. In their research, they focused on Compustat and the Center's for Research is Security Prices (CRSP) SIC codes. The vast majority of studies from the second half of the 20th century used SIC codes reported by these two databases because they were widely distributed and available for research purposes. In contrary to Clarke (1989), they stick to a full 4-digit code and calculated a Pearson correlation coefficient[4] between monthly stock returns of firms having identical 4-digit code. Based on their findings, Compustat SIC codes appeared to be more homogeneous than the other part of their dataset, CRSP codes. Based on a 2-digit SIC code, which refers to a major industry group, there is a 62% match in the above-mentioned databases.

Analysis of Compustat and CRSP databases has been performed also by Kahle and Walking (1996), who showed that just little over one fifth of firms have identical full codes in these two databases and there is almost 80% match in a one-digit code. They address these differences to the fact that many firms change their primary SIC code over time and this change is only reflected in CRSP codes, while Compustat codes remain unchanged. Based on Weiner's (2005) analysis, every fourth firm has changed a SIC code

[4] *Pearson correlation coefficient* is a measure of the linear correlation between two variables calculated as a ratio of the covariance of the two variables and product of their standard deviations

at least once during 1973 to 1994. There has been an improvement in Compustat database in 1987, as there are available also historical *primary SIC codes*[5].

In 1987, there has been another important event related to SIC codes. The Office of Management and Budget did a major revision of SIC codes as it had to keep pace with an industry development that changes the economy a lot compared to SIC beginnings in 1930s. Ten years later, NAICS codes were introduced as a full replacement for SIC Codes.

3.1.2 NAICS Codes

Contrary to SIC 4-digit codes, North American Industry Classification System (NAICS) offer more detailed groupings based on 6-digit codes which can be reduced for more homogeneous 2-digits in financial research. As a number of digits increases, codes are becoming more specific with a 6-digit full industry specification.

The reason for creation of a new global industry classification comes from the fact that SIC codes fail to keep pace with industry development and soon seemed to be insufficient, resulting in not proper fit of firms to SIC definitions. "*The NAICS improves upon the SIC by using a production-based framework throughout to eliminate definitional differences; identifying new industries and reorganizing industry groups to better reflect the dynamics of our economy; and allowing first-ever industry comparability across North America, addressing the monitoring provision of NAFTA.*" (Saunders (1999))

As NAICS gradually replaces SIC codes, it has become a standard classification scheme for classifying businesses and a necessary tool for proper analysis of economic and statistical data. However, there are both classification codes preserved in most databases to support continuity of ongoing research and to ensure data comparability.

Naturally, there was an opportunity to compare these two classifications during the years after NAICS introduction in 1999. Samuelson (1999) wrote an article where he described the need for better classification and presented plans for future regular updates. Later, he was quoted by Krishnan and Press (2002), who performed a detailed analysis of SIC and NAICS groupings and their implications for accounting research. In a part of their study, they also built on a research of Guenther and Rosman (1994), which is described in the previous section, and they focus on intra-industry homogeneity

[5] A *primary SIC code* refers to a company's core industry, at most five *secondary codes* representing other industries not covered by a main industry are allowed for each company

comparison of SIC and NAICS. Based on their findings, NAICS performed better in transportation, service and manufacturing industries.

Despite the fact that NAICS outperforms SIC in some industries, they are also very similar in some characteristics. Both were at least partly affected by the governmental agencies responsible to collect economic and statistical data from a great range of industries, and NAICS build on the same foundations as SIC and only serve as a superstructure in some problematic fields described above. Therefore, their hierarchical structures are very close to each other and both perform similarly in most financial and statistical applications.

3.1.3 Fama-French Industry Classifications

Fama and French (1997) were financial academics who developed their own classification system on SIC basics. They reclassified SIC codes into 48 groups, each representing a specific industry, in order to achieve more homogeneous grouping with similar risk characteristics. Across these groups, there are statistically significant differences in the cost of capital. This adjustment of SIC methodology leads to simplification of a research process and its results were not affected by 'outliers' - industries with a small number of firms, that has to be excluded from many empirical studies as they did not produce relevant findings for corresponding industry group.

Even though this specific FF 4-digit industry classification appears in other academics' researches (namely Lee, Myers and Swaminathan (1999) or Gebhardt, Lee and Swaminathan (2001), Purnanandam and Swaminathan, (2003), Loughran and Marietta-Westberg (2005)), this whole approach to industry classification based on a grouping of selected SIC codes has never been tested for completeness and operating effectiveness. However, its application is straightforward as the grouping can be represented by 48 explanatory variables in a regression model, whose coefficients are easy to interpret even for interaction terms (with binary variables used for the analysis of seasonal effects) included in the model equation. Of course, not all variables have to be included in the model as having 48 explanatory variables could produce strange results and not necessarily lead to higher adjusted R^2 of the model. R^2 will rise with each additional explanatory variable (adj. $R^2 \leq R^2$ from simple calculus), but it does not tell us any useful characteristic of the model as it does not penalise for too many variables.

3.1.4 GICS Codes

Another classification from the end of the 20th century is a Global Industry Classification Standard (GICS) developed by Standard & Poor's (S&P) and Morgan Stanley Capital International (MSCI). As both were main providers of stock indices and benchmark-related products, the reason for research and development of a new classification system is evident.

Compared to a 4-digit SIC and a 6-digit NAICS codes, GICS assigns 8 digits to each company. As a result of collaboration of S&P and MSCI, data availability differs. For companies that belong to S&P 1500, data are available from the end of 1994, for non-S&P companies; there is GICS information since June 1999. The main criterion is a principal business activity of the firm, which is directly observed by S&P and MSCI analysts from company's reports and financial statements. As opposed to SIC and NAICS, GICS is unique in taking into account sources of firms' revenues and earnings, which play a major role in proper classification. Compared to two main classifications mentioned above, this has been the key improvement over the obsolete and inadequate approach taking into account only supply and production characteristics when assigning firms to industry groups. Another advantage over SIC and NAICS is the detailed manual for firms with diversified business activity to get the most relevant GICS code.

Bhojraj, Lee and Oler (2003) performed the most comprehensive study of all of the above-mentioned classification systems. Their findings and conclusions are reflected in the vast majority of studies where an industry classification plays a role in data analysis. They have used descriptive statistics to find general relationships between systems and they have performed a regression analysis. Based on their findings, GICS performs as a superior industry classification system in terms of explaining co-movements in stock prices and cross-sectional variations in forecasted growth rates and valuation metrics. All of these are crucial for academic and financial research. Also, GICS methodology of sorting stocks creates materially different industry samples than when sorting by the other three classifications.

Important implications from Bhojraj, Lee and Oler's (2003) study came from the percentage match between classification systems. The oldest of the four, SIC codes, were set as a base group. Not surprisingly, NAICS reached an 80% match with SIC. This result is expected when taking into account the NAICS purpose to build on successful SIC codes and fix some issues. Fama-French industry codes that re-arrange SIC into 48 groups agree

to SIC even at higher rate of 84%. However, GICS map to SIC only at a rate of 56%, which can produce significantly different results when comparing findings and conclusions of studies using distinct industry classification schemes.

GICS outperforms other classifications in most tests that Bhojraj, Lee and Oler have performed, whereas the other three classifications act similarly. They concluded that this is caused by a "*financial-oriented nature of the industry categories themselves, and a consistency of the firm assignment process.*" (Bhojraj, Lee and Oler (2003)) Especially the fact that the assignment of codes to firms is done by S&P and MSCI specialists and does not rely on data vendors' judgement should be considered as a main advantage of GICS improving the overall consistency of industry classification.

3.2 Approach of this thesis

After a deep look into existing literature and academic papers, it is clearly evident that some type of industry classification appears in the vast majority of them as an instrument for reducing sample size or a tool for observing industry-related relationships. This is not a surprising fact as reasons for their usage were deeply discussed in this chapter. However, just few empirical studies investigate relationship and differences between classification schemes and provide recommendations for their application in financial research, and, the vast majority refers to a comprehensive study of Bhojraj, Lee and Oler (2003). This can be an inspiration for future theses, which can greatly contribute to the overall development and improvement of financial research.

Most recent studies listed above verified that there is a significant difference between currently used industry schemes in terms of methodology manuals and structure schemes. Each require a different dataset modification and one cannot easily compare results of researches that use different schemes in order to avoid an error. All these potential issues would disappear if there would be only one industry scheme. This situation will never happen in real world; however, GICS has a good potential to be the global leading classification. Data vendors provide several classification codes in their databases and the overall data availability is still improving with IT development.

A broad analysis of 'Top Stocks' can produce unexpected findings, however, its performance has to be analysed regularly and only a long-lasting research can provide robust results and trustworthy conclusions. Keeping this in mind, we will use GICS in our analysis to make this thesis comparable with ongoing research in a long-run perspective.

4 Methodology

A statistical analysis includes research planning, a long-term process of data collection, using proper tools to analyse them, application of relevant theoretical concepts and finally, reporting findings of a study.

A *planning section* was carried out a year ago and results are summarised in the Bachelor's Thesis Proposal. This was followed by a *data collection process*, which lasted until February 2018. As mentioned above, no comprehensive study of this topic has been made before. Therefore, this period can be seen as the most important part of our statistical analysis. Without meaningful data containing information we searched for after a detailed literature review and their transformation to a form that allows us to make the analysis to the fullest extent possible, we cannot draw trustworthy conclusions and report findings that would have any added value. In order to get the data from a verified resource, a portfolio manager of 'Top Stocks' has been asked for collaboration on this research. During the data-collecting period, we have requested all the data needed to perform a broad analysis of the portfolio's performance. As expected, some data were missing and additional search for these values was performed with intention to make our dataset complete. For more details, see Chapter 5 – Data.

This chapter refers to the third stage of a statistical analysis – *search for appropriate tools*. Due to the scope of this thesis, research design consists of two main parts. Firstly, basic descriptive statistics and graphics analysis are presented in order to understand the nature of the data, their distributions and historical development of the fund's performance. This section should give a meaning to collected data and prove the reason why we chose them to have in our analysis. Then, a regression and forecast analysis is performed using various testing procedures, whose outcomes are finally compared and discussed.

4.1 Descriptive statistics

Before we perform any analysis, we have to distinguish between different types of variables in order to choose the most suitable statistical method to analyse them. Our dataset contains quantitative (numerical) data represented by both discrete and continuous variables. For the whole analysis, we have used a free statistical software R and Excel 2016, especially the Analytics Toolbar add-in for dataset manipulation.

Such data can be generally analysed using measures of central tendency.[6] These summary statistics include mean, median and mode; basic statistics that do not need further description. Their implications for our dataset and portfolio performance are described in detail in the results' section 6.1 Descriptive statistics and graphical analysis.

Measures of variability include variance, standard deviation, extreme values, range of variables, and shape statistics are represented mainly by skewness. With this simple analysis, we can easily find the shape and spread of our variables.

These elementary calculations help us to not only understand the data, but also to verify some of the key characteristics about the portfolio. As an example, I would mention the analysis of a number of investment ideas in a portfolio that have a strict limit of 25, ensuring a highly concentrated portfolio with reasonable volume of companies, which is an essential assumption of a stock-picking strategy. This analysis is supported by a graphic representation of results in a section 5.1 Data availability and collection.

Even though the methods used for this analysis belong to the most basic statistical methods, they help us to verify a long-term stability of the portfolio in terms of a compliance with internal regulations and verification of necessary assumptions for the proper functioning of the fund.

4.2 Graphical analysis

A graphical representation of results in graphs and tables is a way to present outcome of a statistical analysis in a form that maximises the amount of information while minimising amount of ink at the same time. (Tufte (1983)) This is a guiding principle for every graphical element included in our research. Basic features implying the unambiguity of information are clarity of message passed to readers, simplicity of design without unnecessary elements that can interrupt the flow of information, clarity of words and integrity of intentions and action. (Bigwood and Spore (2003))

Each chart or table should have a title that clearly defines the relationship represented by the element. Also, where applicable, the number of observations can help readers to create a complete picture of concrete data used in particular analysis. We will follow other recommendations of Tufte (1983, 2001) presented in his book '*The Visual Display of Quantitative Information*' when including charts a tables in this thesis.

[6] Also known as measures of central location

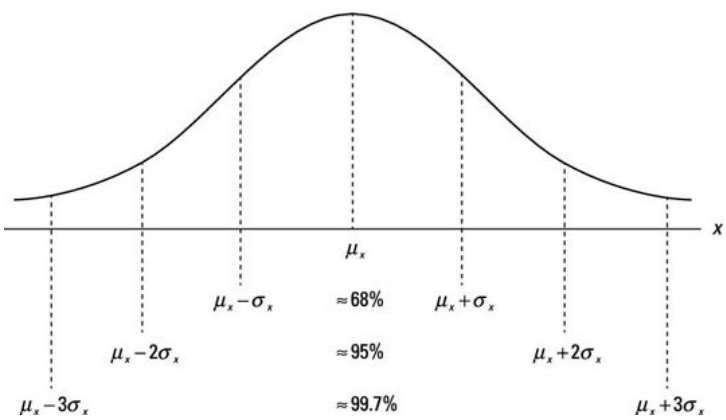
Charts are used to describe relationships between variables, show a visual comparison of distributions and long-term trends in development of a portfolio's performance. Modern statistical software provide us with various possibilities of graphical analysis, including distribution charts (histogram, scatter plot), composition charts (pie chart), bar/line charts and many others. With these figures included in this thesis, one can easily comment on the data distribution and trends, which will be otherwise not that evident when described in words.

In statistics, Gaussian distribution (Normal distribution) can be seen as one of the most important theoretical probability distributions. Where applicable - mostly for the analysis of returns and errors, we will try to find its basic features and comment on their relevance and implications to a corresponding analysis. A graphical representation of data distribution allows us to comment on symmetry about the mean, equivalence of descriptive statistics and a probability of random variable *x* lying between limits defined by the two parameters mean (μ, location parameter) and standard deviation (σ, scale parameter). These probabilities, represented by a *probability density function* of a normal distribution

$$f(x \mid \mu, \sigma^2) = P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ , x \in R \ ,$$

as well as a characteristic bell-shaped distribution symmetric around the mean are presented in Figure 1.

Figure 1: Normal distribution with mean μ and standard deviation σ



Source: Recognizing Usual Variables - Normal Distribution, available at: www.dummies.com [7]

[7] www.dummies.com/education/economics/econometrics/recognizing-usual-variables-normal-distribution

Trends in historical development are the second thing we can easily identify from a simple graphical analysis. We have to keep in mind that stock markets, prices of individual stocks and mutual fund unit price (net asset value) tend to trend in a long-term; therefore, we have to use appropriate tools for de-trending the data in order to perform a complex analysis unaffected by the trends.

The reason for covering the widest possible time range of a fund's performance by our dataset is to distinguish between substantial net asset value volatility in the short run, seasonal effects and a long-term trend pattern. The analysis of volatile stock market environment in a short-term influencing mutual unit price of 'Top Stocks' is far beyond the scope of this thesis as the portfolio consists of 25 companies on average, that are affected by various factors. However, a deep analysis can uncover seasonal trends and a long-term development of selected variables.

As for seasonal trends, we will focus on a net asset value and the total volume of fund's assets using historical development charts with *trendlines* connecting troughs and peaks (often called as tunnel[8]). Both variables will also be analysed in a long-term perspective, which will allow us to forecast future development when identifying sufficiently strong trends. There is no added value of predictions based on short-term evolution of 'Top Stocks', as any attempt to forecast market development (and a portfolio performance) is merely a guess. This estimate can only be improved by taking into account permanent chart trendlines and analysing a performance from a long-term perspective, as the importance of trends rises with their duration (time). A complex analysis requires a large dataset and advanced statistical tools, however, nobody can predict stock market trends with sufficiently high guarantee.

Similarly as descriptive statistics mentioned in the previous section, a simple graphical analysis can provide us with very important findings, even though the tools used for this part of our research belong to the most basic statistical methods. The challenge here is to treat the portfolio's performance since the fund's establishment in August 2006 from a long-run perspective to eliminate a short-run volatility of stock prices, uncover seasonal effects of both price and total volume of assets and finally, search for long-term trends identified by trendlines. While the results are relatively easy to obtain from descriptive statistics and graphical analysis, their proper interpretation is very demanding.

[8] The combination of two trendlines, one for peaks and one for troughs

4.3 Calculations

For the purpose of monthly analysis, the average of weekly net asset value per share (NAV) for the corresponding month has been calculated as follows:

$$avgNAV = \frac{p_{w_1}+p_{w_2}+\cdots+p_{w_n}}{n},$$

where p_w denotes the mutual unit price and indices $1, 2, \ldots, n$ denote the number of weeks related to particular month. Based on the calendar arrangement and weekly data schedule,

$$4 \leq n \leq 5 \, , n \in N$$

has to hold for the tested period. Any value outside this interval would indicate missing data, therefore, a need for manual addition of the corresponding value to the dataset.

When it comes to *returns* representing a change of mutual unit price from one time period to another, we have to consider two possible ways how to calculate these values. A *simple return* at time i, also known as arithmetic return, is calculated as follows[9]:

$$R_i = \frac{p_i}{p_j} - 1 = \frac{p_i - p_j}{p_j}, \; j \equiv (i-1) \, ,$$

where p_i, p_j denote prices in time i, j respectively. Using this formula, we can calculate a return from time j to time i and both weekly (original) and monthly ($avgNAV$ calculated above) data can serve as an input data for the same formula. The obvious advantage of this approach is its straightforward interpretation without the need for calculus, therefore, simple returns are widely used in studies and market analyses intended for the public.

Another possibility is to use *log returns*. We will denote log return at time i with lower-case r_i, calculated as follows:

$$r_i = \ln\left(\frac{p_i}{p_j}\right) = \ln(p_i) - \ln(p_j),$$

where ln stands for the natural logarithm and we apply general logarithm rule for quotient. The most important property of logarithm in this approach to calculating returns is that the graph of a function $y = ln(x)$ passes through point [1;0] of the Cartesian coordinate system. The slope of the function at this point equals 1 and therefore, it is similar to a straight line $y = x - 1$ around point [1,0], which follows from the theory of Taylor series.

[9] Assuming no dividends are paid during the period

This similarity implies that

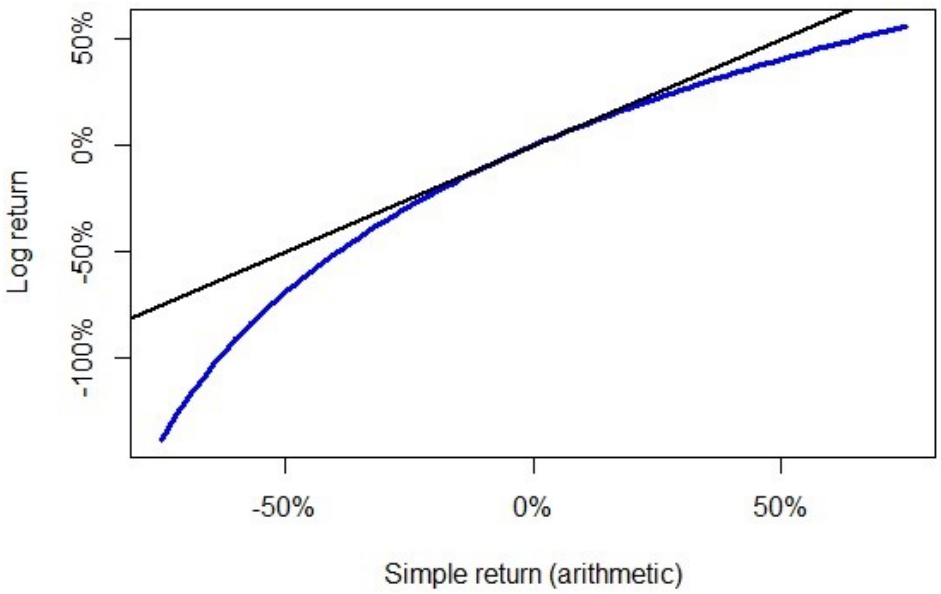
$$\ln(1 + r) \approx r$$

for r much smaller than 1 ($r \ll 1$). To explain it more precisely on a real example, let p (price of the asset) increase by a small percentage (e.g. r = 0.01), so we have $p(1 + r)$ and take a natural logarithm $\ln(p(1 + r))$. Based on the log rule for product, we can split this formula to $\ln(p) + \ln(1 + r)$, and using the above mentioned relationship, the second addend approximately equals r for a sufficiently low r. This process can be summarised as follows:

$$\ln\big(p(1 + r)\big) = \ln(p) + \ln(1 + r) \approx \ln(p) + r,$$

which says that increasing a price of the asset by a sufficiently small percentage (r = 0.01) approximately equals adding 0.01 to $\ln(p)$. The lower is the percentage change in price; the more precise is this approximation, as can be directly observed from Figure 2.

Figure 2: Equivalence of arithmetic and log returns



Source: Author's own computations using a free statistical software R

This theoretical concept, supported by a short explanation to uncover the mathematic logic behind, has numerous advantages over arithmetic returns. In theory, we often assume that returns are log-normally distributed in long-term, which is in most cases proved by positivity and right skewness of stock prices, even though the advanced

methods would be more suitable to use in research. Then, $log(1 + r_i)$ follows a Normal distribution described in detail in Section 4.2 Graphical analysis. As a result, we can benefit from the log symmetry, which in fact means that a growth today and the same decline tomorrow bring us back to the *initial value*. In contrast, when using simple (arithmetic) returns, this equality does not hold.

Both approaches, however, have some disadvantages as well. Several studies have analysed the relationship between arithmetic and log returns and their applicability in financial research. Aas (2004) performed a study called '*To log or not to log: The distribution of asset returns*', where he focused on comparison of distributions of both types of returns. He came to conclusion that even though we often assume returns to follow a Normal distribution, this is not necessarily true, especially when volatility of price increases. This is consistent with the fact that both approaches produce approximately equal results for sufficiently small r. Therefore, a deep analysis of distribution of weekly and monthly returns has to be performed in order to identify possible accuracy threats.

Quigley (2008) supported Aas's (2004) findings and conclusions about the independence and identical normal distribution of returns. In this research, he identified that the assumption of Normal distribution is often violated when using real data from stock markets. The behaviour of investors can cause heavy tails, which cannot be covered by the Normal distribution. Also, there is often a clear evidence of skewness in the distribution of returns, which also affects the validity of our assumptions. In his empirical study, Quigley identified a volatility pattern that is not consistent with iid assumption.

Other implications of these methods for a financial research are directly associated with numerical logic behind and calculus. When it comes to aggregation, there is a significant difference between arithmetic returns that aggregate across assets (a weighted sum of simple returns of corresponding investment ideas in the portfolio), and log returns, which aggregate across time.

To conclude, with an overview of previous and current literature, both approaches bring advantages and disadvantages to the analysis of a portfolio's performance. Fortunately, previous sections (4.1 Descriptive statistics, 4.2 Graphical analysis) would provide us with sufficient evidence of returns' behaviour and we would be able to decide which method will be more appropriate and useful for the purpose of our analysis, keeping in mind what is the group of readers for whom this study is intended. Results are presented in Chapter 6.

4.4 Regression and forecast analysis

The last method used to perform a deep analysis of a portfolio's performance is a regression analysis. Generally, this statistical method allows us to find the relationship between one dependent and several independent variables. Also, additional tests will be performed in order to confirm or refute our claims from Sections 4.1 – 4.3.

Firstly, descriptive and graphical analysis of returns and a fund share unit price (NAV) will be supported by simple regression models and statistical tests, which will verify our comments on general data characteristics. The primary software to analyse our data is a free statistical software R supplemented by MS Excel 2016 in some cases[10].

Before we use our data in regression models, we have to verify their basic properties and consider which tools are appropriate for their analysis. For returns, *stationarity* is often assumed, whereas stock prices represented by NAV in our thesis often show signs of non-stationarity and usually follow a random walk. The reason for the analysis of stationarity is that certain types of statistical models (ARMA, etc.) require stationary time series as an input data, and of course, stationarity is required for the Law of Large Numbers (LLN, for large sample estimation) and the Central Limit Theorem (CLT, for large sample inference) to hold. If any trend or seasonality is present, it has to be removed by a specific procedure before the analysis is performed. Also, Ordinary Least Squares (OLS) method relies heavily on stationarity and improper usage can produce misleading results. A spurious relationship, firstly mentioned by Granger and Newbold (1974), is represented by significant estimates with no logic meaning and high value of a coefficient of determination. To identify spuriousness, one can use a 'rule of thumb', which confirms this issue when $R^2 > DW$ (Durbin-Watson) statistic.

The following definition of stationarity may seem a little abstract, but in general, it requires data distribution to be stable over time:

"The stochastic process $\{x_t : t = 1,2, \dots\}$ *is stationary if for every collection of time indices* $1 \leq t_1 < t_2 < \cdots < t_m$ *, the joint probability distribution of* $(x_{t_1}, \dots, x_{t_m})$ *is the same as the joint probability distribution of* $(x_{t_{1+h}}, \dots, x_{t_{m+h}})$ *for all integers* $h \geq 1$*." (Wooldridge (2012))*

[10] data analysis add-on for MS Excel 'XLSTAT' (statistical software), which offers a wide range of tools for a time-series analysis including visualisation, descriptive analysis, DW test, tests for homogeneity and heteroscedasticity of data, ARMA, unit root and stationary tests etc.

As a direct implication of this definition, statistical properties like mean and variance are stable over time, which ensure accuracy and relevance of our time series in a regression testing. Also, there is weaker form of stationarity, known as a covariance stationary process, with following properties:

(i) $E[x_t] = \mu$,

(ii) $Var(x_t) = \sigma^2$,

(iii) For any $t, h \geq 1$, $Cov(x_t, x_{t+h})$ depends only on h and not on t,

where $\{x_t: t = 1, 2, \dots\}$ is a stochastic process with finite second moment, i.e. $E[x_t] < \infty$, μ and σ are constants and $Cov(x_t, x_{t+h}) = f(h) \neq f(t)$, where t is time. In other words, $\{x_t\}_{t=1}^{T}$ is covariance stationary if its mean and variance do not change with time and the value of covariance between time periods depends only on distance (lags) h between two corresponding time periods, but does not depend on time t itself. Important to mention here is that, if a stationary process has a finite second moment, it has to be also covariance stationary, but not vice versa. (Wooldridge (2012))

Stationarity will be tested for both returns and NAV time series using several statistical tests, which can be divided into two groups based on their null hypothesis. Firstly, there are *tests of stationarity* (e.g. KPSS test), which test stationarity (H_0) against H_A non-stationary time series. Secondly, there are unit root tests (ADF test, PP test), for which stationarity is the alternative hypothesis to a unit root process (H_0). It is a common standard in academic research to provide results of at least two tests, because each has some type of limitations and, if we perform more tests, it can be seen as a sufficient verification of the results. Even though non-stationarity can sometimes be directly observed from a graphical analysis described in Section 4.2, when a time series plot shows a clear trend or seasonality, stationarity often requires using statistical software.

At first, the autocorrelation function (acf) will be used to confirm theoretical assumptions of stationary returns and non-stationary NAV by plotting 'correlograms' for 50 lags[11] with default setting of 95% confidence interval (CI), which seem sufficient for the majority of statistical tests. Also, this is in line with previous researches' approaches, where 95% CI represents a desired level of confidence set by the authors. The output plot shows a correlation of time series data with their lagged values and a dashed line

[11] Set by the author's judgement to maintain consistency and uniformity of testing approach when comparing correlograms of returns and NAV, where 50 lags are appropriate to show the time development of correlation coefficients

represents bounds of CI and therefore clearly identifies correlation outside the threshold. It is important to mention that the lag 0 autocorrelation is fixed at 1 by the convention. Numerical values of all correlation coefficients at different lags can be extracted and further analysed in R using a special command.

Following the graphical analysis, stationarity will be tested by above mentioned hypotheses testing procedures. The most common test for stationarity is the Augmented Dickey-Fuller test[12] (ADF test) by introduced by Dickey and Fuller (1979, 1981), which belongs to the group of unit root tests. Therefore, it identifies a presence of a unit root process (hence non-stationarity of time series), which is the null hypothesis, whereas the alternative hypothesis stands for a (trend)-stationarity. The unit root process is a non-stationary process that often raised up issues in statistical inference of time series data, because it makes OLS method invalid and what is more, possible spurious regression can dramatically affect our findings and conclusions.

To preserve simplicity with regard to the target group of readers of this thesis, the general version of a Dickey-Fuller test will be introduced in detail, as it is more straightforward to understand. Both the general and augmented versions are available in all statistical software with pre-defined logical processes, so the test procedure itself is fully automatized. However, to understand the main idea of this widely-used unit root test, we will stick to a description of the non-augmented version.

The autoregressive unit root process can be written as follows:

$$x_t = \alpha + \rho x_{t-1} + \varepsilon_t,$$

with ε_t being a stationary error process (assumed iid) and where α should be either equal to zero (a random walk) of should be different from zero representing a random walk with a drift (or a stochastic process with a time trend). The hypotheses of a DF test are

$$H_0: \rho = 1$$
$$H_A: |\rho| < 1,$$

where the null hypothesis stands for a unit root, whereas the alternative indicates a stationary process. As mentioned above, under the null, both x_t and x_{t-1} are non-stationary, so estimating whether ρ is different from one can produce misleading

[12] Augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models, which allows testing of higher orders of autoregressive processes

results as a t-ratio does not follow a t-distribution. To deal with this issue, x_{t-1} will be subtracted from both sides of the autoregressive equation:

$$x_t = \alpha + \rho x_{t-1} + \varepsilon_t \qquad / -(x_{t-1})$$
$$x_t - x_{t-1} = \alpha + (\rho - 1)x_{t-1} + \varepsilon_t$$
$$\Delta x_t = \alpha + \delta x_{t-1} + \varepsilon_t \ .$$

Obviously, under the null hypotheses of a DF test, $\delta = 0$ and the non-stationary x_{t-1} disappears from the right hand side of the equation. However, CLT does not hold and $\hat{\delta}$ does not have a t-distribution. Therefore, the value of a t-statistic of $\hat{\delta}$ has to be compared with specific critical values calculated by Dickey and Fuller[13]. With $t < DF_\tau$, we will reject the null hypothesis of a unit root. Under the alternative, a stationary x_{t-1} would be included in the equation.

The ADF test provides improvement of a very simplifying assumption of AR(1) process when allowing testing of higher orders of autoregressive processes (*n*). The process is the same as for simple DF test, but the procedure is applied to the model

$$\Delta x_t = \alpha_0 + \alpha_1 t + \delta y_{t-1} + \sum_{i=1}^{n} \alpha_i \Delta x_{t-1} + \varepsilon_t.$$

Based on Černý and Kočenda (2007), ADF test exhibits a high chance of Type II error of not rejecting incorrect null hypothesis. As a result, ADF test has been modified several times by other researchers[14] to provide more precise results. As mentioned earlier in this chapter, we will perform other tests (KPSS, PP) to verify results of the ADF test for both returns and NAV time series.

Finally, there is a brief confirmation of non-stationarity of AR(1) process with $\rho = 1$ (a random walk), based on properties of a covariance stationary process. Consider

$$x_t = x_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim iid\ (0, \sigma^2)$. Then we substitute x_{t-1} on the right hand side of the equation by corresponding value obtained from the general equation above and repeat this step until we obtain the final equation:

$$x_t = x_{t-2} + \varepsilon_{t-1} + \varepsilon_{t-2}$$
$$x_t = x_0 + \sum_{i=0}^{t-1} \varepsilon_{t-i}$$

[13] Critical values were calculated via Monte Carlo simulation, denoted as τ statistics
[14] For details, see Bell, Dickey, Miller (1986); Dickey and Pantula (1987); Perron and Phillips (1988)

Now, we will show that a random walk process does not meet the criteria of a covariance stationary process, keeping in mind that $\varepsilon_t \sim iid\ (0, \sigma^2)$:

i. $E[x_t] = E[x_o] + \sum_{i=0}^{t-1} E[\varepsilon_{t-i}] = E[x_0] = 0$

ii. $Var(x_t) = \sum_{i=0}^{t-1} Var(\varepsilon_{t-i}) = t\sigma^2$

iii. $Cov(x_t, x_{t+h}) = Cov(x_t, x_t + \sum_{i=0}^{h-1} \varepsilon_{t+h-i}) = Var(x_t) = t\sigma^2$

Clearly, both variance and covariance are functions of time; therefore, their values change with time and do not meet the criteria of a covariance stationary process.

Next, we will use a Box-Jenkins (1970) method to construct the *ARIMA*[15] model in order to forecast future points in the time series of weekly NAV based on past values. This model offers improvement over ARMA by including differencing to ensure stationarity of the series. General ARIMA model with y_t' representing a differenced series can be written as follows:

$$y_t' = c + \phi_1 y_{t-1}' + \cdots + \phi_p y_{t-p}' - \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t\ .$$

However, more common is to use a *backshift operator B* representing the following relationship between lagged values of y_t:

$$By_t = By_{t-1}$$
$$y_t' = (1 - B)y_t$$
$$y_t^d = (1 - B)^d y_t\ .$$

This relationship is the core of a *backshift notation* of $ARIMA\ (p, d, q)$ model, with p, q representing the order of autoregressive and moving average parts respectively, and d the degree of first differencing involved in a model.

$$\underbrace{\left(1 - \phi_1 B - \cdots - \phi_p B^p\right)}_{AR(p)}\ \underbrace{(1 - B)^d}_{d \text{ differences}} y_t = c + \underbrace{(1 + \theta_1 B + \cdots + \theta_q B^q)}_{MA(q)} e_t$$

[15] AutoRegressive Integrated Moving Average model (ARIMA)

To find the optimal values of p, d, q needed to map the historical development of weekly NAV and forecast future points of this time series, we will use auto.arima function in R which fits the best ARIMA model to univariate time series based on historical values. As every automated process, it brings a certain degree of risk to our analysis. Testing all possible combinations of p, d, q can often be very time demanding, therefore, there are several short-cuts included in default setting of auto.arima. This simplification cannot necessarily produce the best fit at the end; however, it can sometimes be the only possible solution for large datasets. Nevertheless, we will try to disallow any simplification and let R to run the code with no stepwise and approximation.

To estimate parameters of a model based given past observations, the Maximum Likelihood Estimation (MLE) is used. The aim of MLE is to maximise the likelihood function, i.e. to find values of parameters (called MLE estimates) which maximise the probability of getting the most precise match between our model and observed data. In ARIMA models, these estimates are similar as those obtained using OLS method, which minimises the sum of squared residuals $\sum_{i=1}^{T} e_t^2$.

Our testing approach is as follows: Firstly, we will use auto.arima to identify the best fit according to the historical development of NAV, which will be confirmed by a time series plot of actual and fitted development. Secondly, we will use a forecast function to find future values of our time series and obtain corresponding confidence intervals. Then, we will analyse residuals of our model by performing a Ljung-Box test, checking normality of residuals by various graphs and acf and finally, we will calculate average difference between actual and forecasted values during the last year.

This methodology approach is in line with similar studies where NAV has been forecasted. More specifically, Lawrence et at. (2015) presented various time series models including ARIMA to forecast NAV when analysing a mutual fund performance. Also, Dalrymple (1978) studied Box-Jenkins techniques to fit the sales development and used this ARIMA model to forecast future values. Other studies focused not only to a forecast analysis of NAV or other prices, but they also compared Box-Jenkins method assuming no particular pattern in the past data of the time series with other methods that can be used. In many research papers, automatic Box-Jenkins forecasting is performed and values of p, q are verified retrospectively using acf and pacf functions in a statistical software. Based on Hill and Woodworth's (1980) study, automatic modelling yields the same results as manual analysis.

5 Data

A broad analysis of the fund's performance does not require collecting the widest possible dataset. Based on a detailed literature review, one can focus on specific information provided by data vendors, which can reduce the possibility of a 'too much information problem' with large databases, which has been studied by Talbert et al. (2013) and others. This problem is quite common nowadays as we face almost unlimited data availability thanks to a worldwide technology development.

It is, however, very important to test the longest possible period to capture a long-term development of the fund's performance and uncover seasonal anomalies and eliminate their excessive effects in the short-run. Therefore, the final dataset should not only consist of desired information needed for our analysis, but these data have to cover the whole period subjected to testing. In addition, there should be some process of verification of the data to prove their validity.

5.1 Data availability and collection

Even though there are some information that has to be publicly available based on both Czech and EU laws[16], these information do not seem sufficient for our analysis. Either they do not cover the whole period of the fund's existence, or, they do not contain information needed for a complex analysis. Therefore, with a deep knowledge of previous and ongoing research and a detailed overview of a methodology for this thesis, the author of this thesis has created the whole dataset itself using both public and private data provided by the portfolio manager.

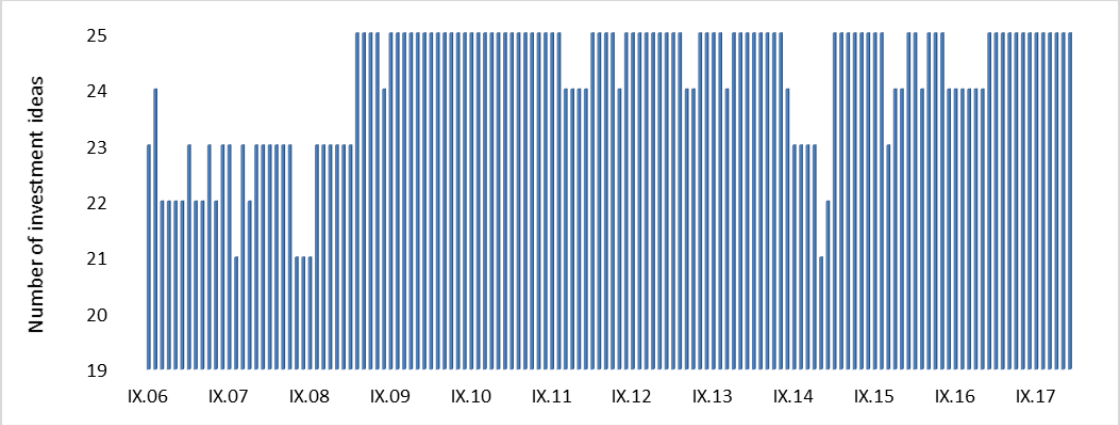
Firstly, the *net asset value per share* has been obtained from 'Kurzy.cz, s.r.o.'. This free online database contains weekly data for the net asset value, total volume of fund's assets and returns from previous week and month. Other data such as 3-, 6- and 12-months returns are missing in the majority of cases and do not cover a period subjected to testing. This period was set by the author's judgement to 9/2006 – 2/2018. However, Kurzy.cz does not cover the first 8 months of the fund's operation. This issue will be discussed in detail in the following section 5.2 Dataset description.

Secondly, based on a literature review, Global Industry Classification Standard (GICS) sectoral breakdown was needed for our analysis, which is included in monthly reports. Seven latest reports are available online at a webpage 'Investiční centrum' run by

[16] The European Commission regulation # 583/2010: ‚The Key Information Communication Regulation'

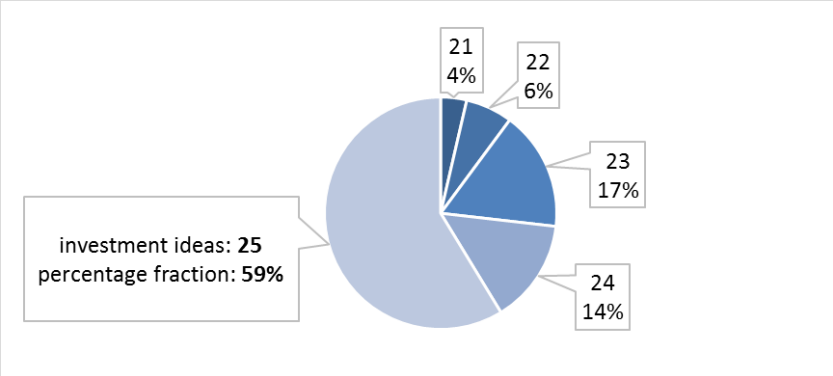
Česká spořitelna. We have requested older monthly reports needed to cover a 138-month period directly from a portfolio manager, Ing. Jan Hájek, CFA. These reports contain not only a GICS sectoral breakdown, but also information about the number of investment ideas and total assets volume of the fund related to particular month. A quick analysis of historical development of the number of investment ideas and their percentage fraction has been performed and results are summarised in Figure 3 and Figure 4.

Figure 3: Historical development of the number of investment ideas



Description: Figure 3 shows the historical development of the number of investment ideas included in the portfolio of 'Top Stocks'. By this chart, we have verified that the maximum limit of 25 ideas has not been exceeded during the twelve-year history of the fund's operation. This is the key assumption of a stock-picking strategy to ensure its operating effectiveness.

Figure 4: Number of investment ideas and percentage fraction



Description: Figure 4 depicts the percentage fraction of the number of investment ideas included in the portfolio and serves as an auxiliary graph to Figure 3.

Source (Figure 3, Figure 4): Author's own computations using monthly reports

As there are usually 25 investment ideas and their fluctuation is materially insignificant in a less-than-month intervals, monthly reports seem sufficient to cover major industry changes. Since there are always seven reports available online with an unlimited public access, a continuous process of data collection for any future analysis of 'Top Stocks' should be markedly easier and faster as the dataset prepared ty the author of this thesis contains historical information from the fund's establishment in 2006.

Finally, there is a need for annual reports to obtain information about monthly statistics related to the number of issued and redeemed mutual fund units, and, a detailed structure of the portfolio including information about the country of issuer of a security. Also, 'Ernst & Young Audit, s.r.o.' as an external auditor provides the auditor's report on financial statements, which has to be included in the annual report of 'Top Stocks' according to the Czech law.

Upon request, we have obtained annual reports covering the period 2011 – 2016 from a portfolio manager. In addition, there is also a mid-year evaluation available for 2017. At the time of completion of this thesis, annual report for 2017 has not yet been available due to financial statements procedures and audit purposes. Based on information provided by a portfolio manager Ing. Jan Hájek, CFA, older reports are not accessible.

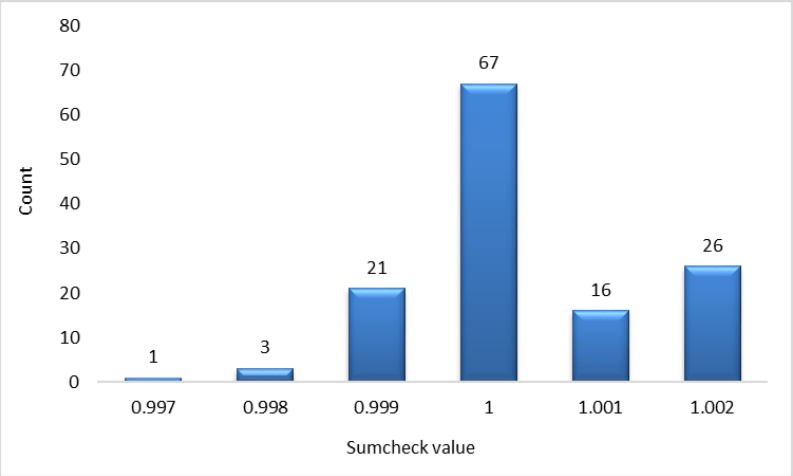
5.2 Dataset description

The final dataset created by the author of this thesis is the very first of its kind and scope as no complex public research of 'Top Stocks' has been conducted during its twelve-year history.

GICS sectoral breakdown is available for the full tested period (9/2006 – 2/2018). Originally, monthly reports consist of 27 sub-industries and a cash 'sector'. However, the majority of them had not been used for the whole period subjected to testing. This can be caused by industry development, sectoral performance evolution or different preferences of investors, which is not clearly visible at first glance. Data were adjusted for GICS structural analysis by creating corresponding *industries* (22 industries + cash) and *industry groups* (11 industry groups + cash) from original *sub-industry* data in monthly reports. This dataset modification does not affect homogeneity of groups or relevance of the analysis based on the previous academic research mentioned in the literature review as we strictly followed GICS methodology manual. See Appendix A for a detail description of this procedure.

Naturally, there are some blank cells, as a portfolio cannot necessarily cover all of these categories in a particular month. To check the dataset for potential errors due to the manual transcription of all monthly reports with GICS sectoral breakdown, which are available in PDF format only, a *sumcheck* [17] of each line (month) was performed. The range of all values equals 1 ± 0.003 and slight differences are caused by a two-decimal rounding for the reporting purposes in monthly reports. This difference is statistically insignificant and data validity check to eliminate typing errors has been performed successfully.

Figure 5: Sumcheck distribution



Description: This figure represents the output of the verification procedure performed by the author of this thesis to eliminate typing errors that might occur during the extensive manual transcript of the data. For each month, the sum of percentage fractions of sub-industries should equal to one. The rounding of the numbers for reporting purposes in monthly reports caused slight deviations, which are materially insignificant.

Source: Author's own computations using monthly reports

Other information from monthly reports such as the total volume of assets, number of investment ideas and the expected consensual dividend yield of the companies included in the portfolio were added in the dataset with no additional modification.

Average of the net asset value has been calculated for the purpose of monthly analysis by using *averageif* [18] function in MS Excel 2016. Arithmetic mean is the best indicator of the average price over the month as there are four prices for each month

[17] A *sumcheck*: sum of each line to verify that the percentage breakdown equals 1 (100 %) for each month
[18] *Averageif* function: finds arithmetic mean for the cells specified by a given condition or criteria

(weekly prices) and there is no trend in net asset value over the period of one month. Obviously, this calculation brings a certain degree of approximation to this study; however, it will not produce materially different results with respect to length of a period subjected to testing. Also, weekly analysis with originally obtained data is performed as well.

As described in the previous section, a free online database of stocks' prices 'Kurzy, s.r.o.' contains NAV per share data since 5/2007[19], while the fund has been operating since the end of 8/2006. In order to perform analysis to the fullest extent possible, the author manually added values of NAV for the period 9/2006 – 4/2007 from the fund's official webpage, where is an interactive graph from which these values were extracted. This 8-month period consisting of 24 weeks is short enough to be added manually to our dataset without a significant risk of typing errors. This adjustment is necessary for analysing the full period, comment of findings and formulate credible and trustworthy conclusions of this thesis.

Returns and other computations are included in the original dataset from a data vendor as well, however, to verify their correctness, they have been recalculated directly in a statistical software R when used in a calculation. The reason for this approach is to verify the accuracy of the data provided by a third-party before they are used in a study. A simple calculation minimizes a statistical error, which can occur when using unverified data. Other variables are directly specified in R when added to a model or used in a formula.

To conclude, a dataset used for the analysis of a fund's performance since its establishment to February 2018 is complete with no missing values that would otherwise cause problems for this analysis. Extensive manual transcript of the GICS sectoral breakdown has been checked for typing errors using a sumcheck tool in MS Excel 2016. Another manual adjustment of NAV for a 24-week period from the very beginning of the fund's operation does not bear a significant risk of errors that will cause materially different results.

[19] https://www.kurzy.cz/podilove-fondy/eamcr/top-stocks-reinvesticniretail/statistiky/?page=29

6 Results

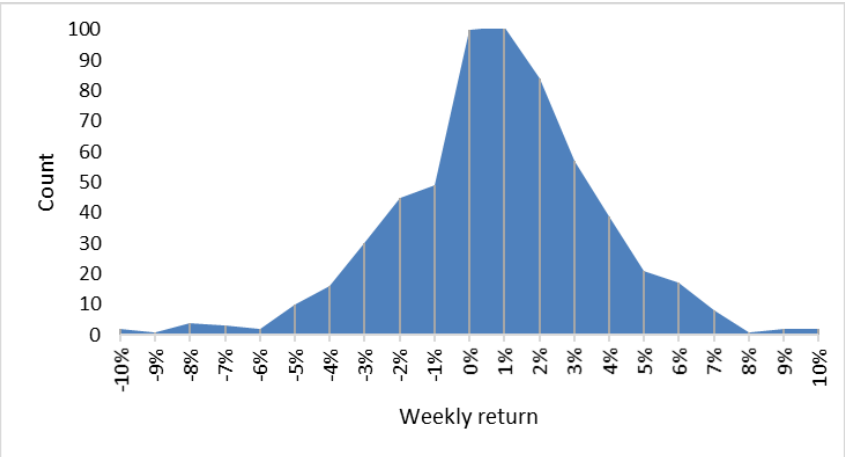
This chapter is divided into three parts. At first, we describe basic characteristics of our dataset using descriptive statistics and graphical analysis according to the methodology approach introduced in Chapter 4. This section should provide us with sufficiently strong evidence of validity of theoretical assumptions needed for deeper analysis of a portfolio's performance. Secondly, we focus on a regression and forecast analysis and perform other related procedures. Then, we discuss the outcomes with respect to past researches and limitations of our test approach.

6.1 Descriptive statistics and graphical analysis

Returns

Following general methodology principles described in sections 4.1, 4.2 and 4.3 of this thesis, we performed a detailed analysis of weekly and monthly returns. By plotting a simple histogram (Figure 6) of weekly returns, we can directly observe that they approximately follow the Gaussian distribution[20]. This corresponds to previous studies of Fama (1976), Brown and Warner (1985) and others, and is in line with our expectations as no reason for different behaviour of returns has been identified yet.

Figure 6: Distribution of weekly returns



Description: Figure 6 represents the distribution of weekly simple returns of 'Top Stocks'. Six outliers higher than 10% in absolute value are excluded, N=594.

Source: Author's own calculations

[20] Depicted in Figure 1, section 4.2 Graphical analysis of this thesis

Basic characteristics of weekly and monthly returns denoted r_w and r_m respectively, are shown in Table 1. A comparison of mean, median and mode values, confirmed by a direct calculation, indicates a moderate skewness of the distribution of returns.[21] The surprising fact here is the negative sign of the coefficient of skewness, as we expect rather positive coefficient because of the fact that returns cannot take values smaller than -100 % as a portfolio cannot lose more than 100 % of its value.

Table 1: Basic statistics of returns

	r_w (%)	r_m (%)
Mean	0.23	0.96
Mode	0.33	-2.90
Median	0.36	1.18
Variance	0.11	0.51
Min	-23.62	-37.05
Max	18.10	34.31
Skewness (x)	-0.95	-0.80
Kurtosis (x)	8.48	4.49
IQR	3.32	6.96

Description: Table 1 summarizes basic statistics of weekly and monthly simple returns of 'Top Stocks'. Notation *(x)* indicates that coefficients of skewness and kurtosis are pure numbers with no units, whereas other statistics are in percent. The unexpected negative skewness of both sets of returns is further analysed.

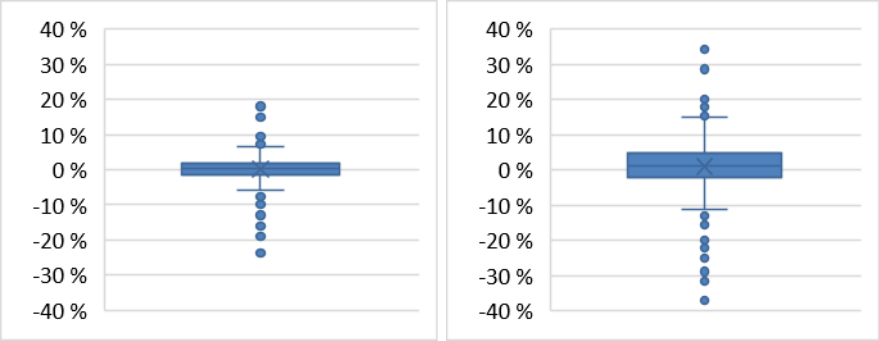
Source: Author's own calculations

Usually, a right tail (positive skewness) representing a maximum gain should be considered unlimited with theoretically no upper bound. "*Since the skewness is very sensitive to the extreme value and the maximum gain is unlimited while the maximum loss is 100%, the effect of maximum gain dominates that of maximum loss.*" (Gong (2014)) Hence, we have constructed boxplots (Figure 7) for both sets of returns to identify a position of outliers with respect to the distribution. Obviously, there are several outliers on the left side of a distribution that cause a sign of the coefficient of skewness to be negative. Our data exhibit a so-called 'fat left tail', which together with positive mean value show properties of a statistical profile of a *Taleb distribution*, which in general implies higher probability of payoff of small positive returns and low probability of losses carrying a significant risk for investors.[22]

[21] Generally, negatively skewed distribution has following properties: $mean < median < mode$
[22] For more insight, see Taleb (2007): *The Black Swan: The Impact of the Highly Improbable*

Features of such a distribution of both weekly and monthly returns are consistent with a stock-picking strategy as a team of professionals is constantly analysing the market in order to maximise the portfolio's performance, resulting in steady and relatively stable positive returns.

Figure 7: Boxplots of weekly and monthly returns, respectively



Description: These charts represent a graphical visualisation of weekly and monthly simple returns of 'Top Stocks' respectively. Outliers are plotted as individual points outside the interquartile range (IQR). The same scale of vertical axes allows us to compare distributions of both sets of returns in percent.

Source: Author's own calculations

Also, we should consider previous researches' findings and conclusions about the presence of negative skewness in the distribution of returns. Chen, Hong and Stein (2001) performed a study of firms' size impacts on a distribution of returns. The reason for taking size of a business into account is that there is a different policy for the release of business-related news, which can affect stock prices and returns. They found an empirical evidence of right skewed distributions for small companies, which can afford to release positive information immediately and do not hurry up with sharing news with negative impact on their business, because there is lower pressure on the management.

Other studies have supported this empirical evidence of negative skewness, as oppose to positive skewness, which occurs much often in the case of large firms with different internal regulations for information release. Ekholm and Pasternack's (2007) empirical research focused on a business size impact in detail. They conclude that non-scheduled news force stock returns to form a negatively skewed distribution, and differences between small and large firms' distributions are caused by *"asymmetries in the news disclosure policies of firm management"*. (Ekholm et al. (2007))

Taking into account a portfolio structure of 'Top Stocks' consisting of large companies including Microsoft Corporation, PANDORA, Starbucks Corporation, Vertex Pharmaceuticals, Apple Inc. etc., the distribution of returns will probably face a moderate negative skew based on previous studies.

What is more, data exhibit an excessive kurtosis[23] (leptokurtic distribution with coefficient of kurtosis equal to 8.48 for weekly returns and 4.49 for monthly returns) based on Table 1. According to the previous research conducted by Fama (1965) and further studied by Choudhry (2001), Beaulieu et al. (2003) and others, kurtosis often appears in the distribution of asset returns. Pesaran (2010) found strong evidence of non-normality of asset returns in period of bubbles and crashes on a stock market, and he verifies this claim by analysing four indices[24] for the period 2000 – 2009. As the excess kurtosis has been present even in short time intervals, he rejected the null hypothesis of Gaussian distribution based on Jarque-Bera test on sufficiently high significant levels.

Our complex analysis of distribution of weekly and monthly returns has raised several issues to our assumption of Normal distribution of returns. It is often assumed that this holds based on the *Efficient Market Hypothesis (EMH)* proposed by Samuelson (1965) and Fama (1965). This well-known theoretical concept is supported only by plotting a histogram in many studies; however, further testing can uncover very important features of a distribution that can differ quite a lot from Normal. Theoretical assumptions of EMH are frequently subject to criticism (e.g. Lo and MacKinley (1988), Plerou et al. (2001), Pesaran (2010)) and even though this concept is respected and subject to testing in both previous and current financial researches, it is not worshipped as a valid and trustworthy theory built on sufficiently strong foundations.

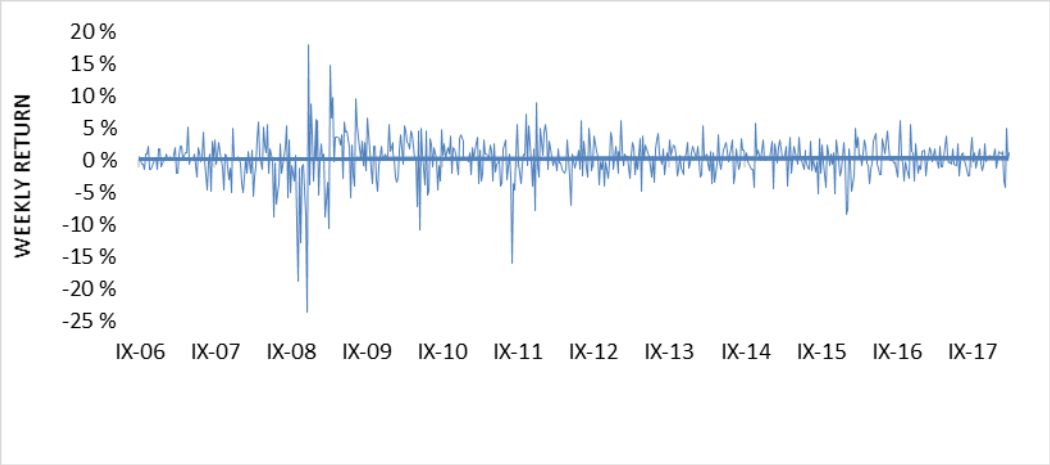
To sum it up, both weekly and monthly returns of 'Top Stocks' are characterised by non-Gaussian behaviours and their distributions are strongly leptokurtic. Therefore, we will stick to simple returns instead of continuously compounded returns (both approaches are described in detail in Section 4.3).

Finally, Figure 8 shows a time-series plot of returns. Compared to asset prices (Figure 9), which usually follow a random walk and can be described by I(1) process, returns are often assumed to be stationary, therefore integrated of order zero, I(0). Having said that, the reason for using stock returns instead of prices in financial research is evident as I(0) process has much more convenient properties for the analysis, especially

[23] 'peakedness' of a distribution
[24] S&P, FTSE 100, DAX, NIKKEI 225

thanks to a stability of moments over time. Also, using OLS requires a stochastic process to be stationarity, as non-stationary time series data can lead to spurious correlation, as moments change with time. In such case, a statistical package can produce significant results; however, they will be completely irrelevant and can bring serious issues to our findings and conclusions. When using proper model and relevant tools, these false relationships would disappear immediately. More testing is done in the following section using the Augmented Dickey–Fuller test.

Figure 8: Time-series plot of weekly returns



Description: Time series plot of weekly simple returns of 'Top Stocks' confirms the theoretical assumption of stationarity, N=600. See the next section for related hypotheses testing using both tests of stationarity (KPSS) and unit root tests (ADF, PP).

Source: Author's own analysis

Net asset value

Apart from returns, a fund share unit price (net asset value, NAV) has been analysed in detail in order to uncover long-term trends and potential seasonal effects, which are, however, more relevant for stock returns. Fama (1965) found an evidence of higher variance in stock returns on Mondays, which has been later confirmed by Gibbons and Hess (1981), who performed an empirical study of S&P 500 data (1962 – 1978) and based on their findings, they were able to reject the null hypothesis of identical distribution of stock returns during a trading week. Anomalies related to months were firstly examined by Rozeff and Kinney (1976), who have found a sufficiently strong evidence of higher returns in January compared to the rest of the year.

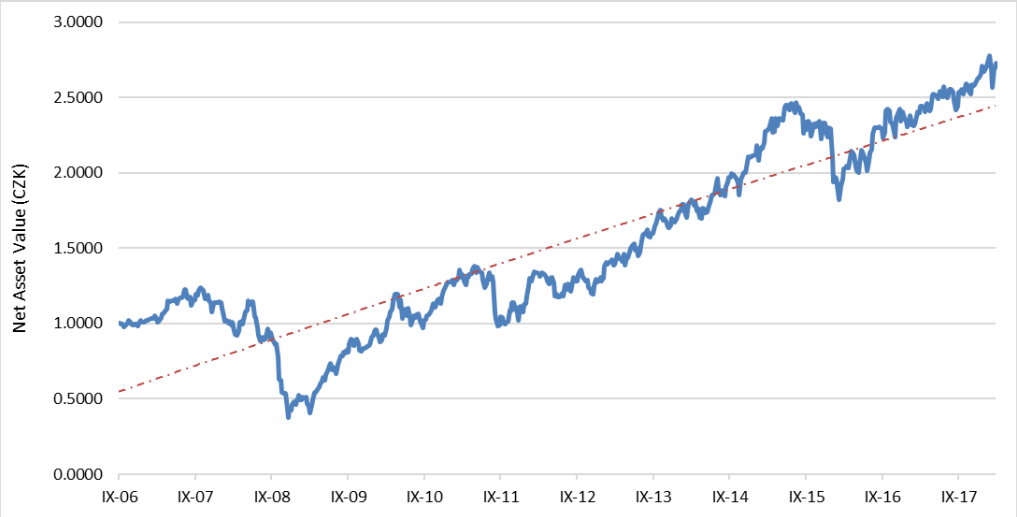
A line chart (Figure 9) shows a historical development of weekly net asset value (NAV) since the fund's establishment in August 2006 to February 2018. It is evident that there is a long-term upward trend and that a fund share unit price significantly deviates from this trend only in cases of sudden market crashes and booms.

Linear and logarithmic trendlines appear to be very similar (they are undistinguishable at first sight), whereas other trends (including exponential, polynomial etc.) are represented by slightly different curves and corresponding R^2 is substantially lower. NAV historically follows these equations:

Linear trend: $y = 0.0005x - 17.136$ with R^2 equal to 0.8209

Logarithmic trend: $y = 18.558\ ln(x) - 195.62$ with R^2 equal to 0.8140

Figure 9: Historical development of a fund share unit price of 'Top Stocks'



Description: Time series plot of weekly net asset value per share of 'Top Stocks' in CZK confirms non-stationarity of our data. A trendline depicted by a red dashed line represents a historical trend specified by the equations mentioned above. Clearly, at the time of the Global Financial Crisis, European Debt Crisis and during the major fall of stock and commodity prices at the beginning of 2016, NAV falls under the long-term trend.

Source: Author's own analysis of data from Kurzy.cz

A quick look at this chart confirms non-stationarity of NAV, which is in line with our expectations built on previous researches. For example, Murthy, Washer and Wingender (2011) performed an empirical study of (non-)stationarity of stock prices, when analysing the Dow Jones Industrial Average, NASDAQ composite and S&P 500

indices for a statistically significant period (1971 – 2009). They conclude that all three major data sets of daily closing values show a non-stationary behaviour, therefore, "*trading strategies that simply rely on mean reversion of stock prices are valueless*". (Murthy et al. (2011))

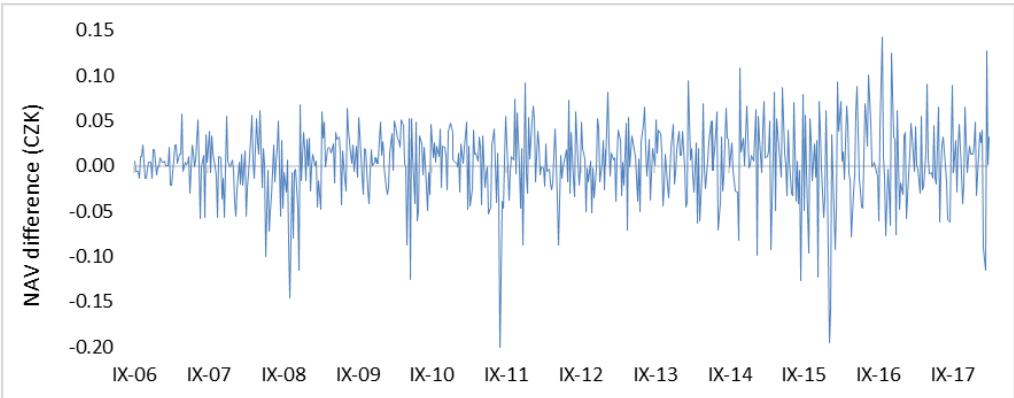
Following the definition of stationarity cited in Section 4.4 (Wooldridge (2012)), the joint probability distribution function of our data does not meet the criteria and we face a non-stationarity issue of unstable mean and variance over time. To deal with it, we need to stabilize moments by differencing the data, i.e. we calculate the change between each two consecutive periods, resulting in a series of changes between time periods.

This adjustment eliminates the upward trend present in our time-series, enables us to find meaningful values of mean and variance, which are irrelevant for non-stationary data as having no reporting value, and thus makes our series applicable for further testing. For this purpose, we used a statistical software R to follow this pattern for the whole series of weekly NAV:

$$p'_t = p_t - p_{t-1} \,,$$

where p denotes a fund share unit price and t stands for a time period. Obviously, a differenced series will always have $t-1$ values as we are not able to calculate a difference for the first observation. Therefore, we have 599 observations for weekly differenced data. A de-trended series is depicted in Figure 10. We can conclude that weekly NAV is stationary in difference, i.e. integrated of order 1, I(1). Random walk, which has these properties, has been described in detail in Section 4.4 of this thesis.

Figure 10: A de-trended series of weekly NAV



Description: A differenced series of weekly net asset value of 'Top Stocks' (in CZK) confirms that NAV is an I(1) process, i.e. it is stationary in first differences. Of course, the differenced series has one less observation than the original one, therefore, N=599.

Source: Author's own analysis

Total volume of assets

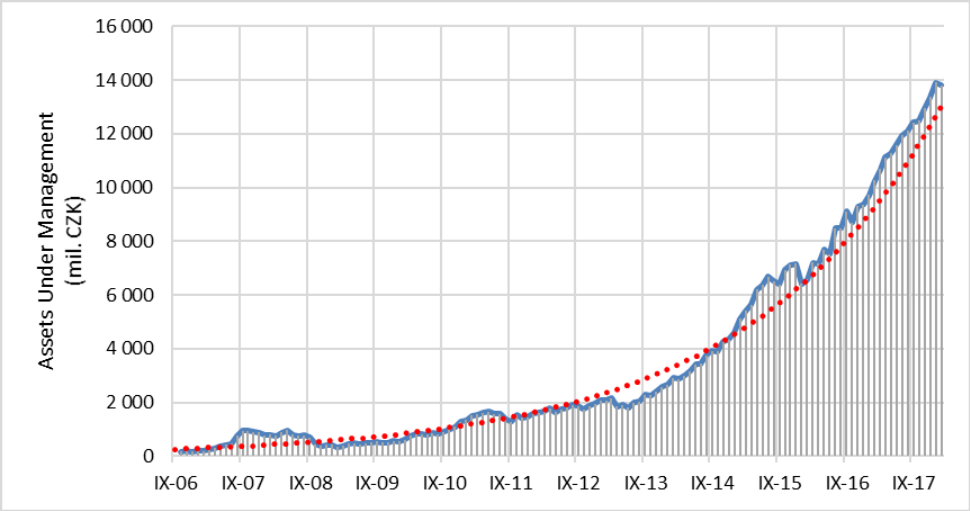
Total volume of assets, also known as '*asset size*' of a fund or '*assets under management*' (AUM), refers to the total market value of financial assets being managed by a fund, investment company etc. It is often considered as an indicator of funds' size; therefore, it enables us to compare individual investments products by the volume of total funds they control. Apart from statistical, reporting and managerial purposes, it can also be seen as marketing tool attracting new investors and providing them with easily accessible information to compare various products on the financial market. However, it is important to see beyond pure numbers when considering whether to invest or not, and pay more attention to investment strategy of the fund, which materially affects the capital appreciation and a fund performance.

Even though AUM and a fund's performance are not highly related to each other, we should consider the historical development of asset size before we invest our funds. 'Top Stocks' is an actively managed fund, which markedly outperforms its internal benchmark represented by a passive investment strategy. In other words, it offers more by a constant market search performed by the team of professionals, who can pick 25 investment ideas at maximum, which are likely to produce high returns and carries sufficiently low risk that is acceptable for investors. The index of AUM, especially its historical development, could be seen as a trust-indicator of investors.

There are two sources of a growth (decline) of AUM. Firstly, it is affected by capital inflow/outflow, i.e. volume of funds provided by investors during a selected time period. Based on our analysis (see Appendix B), there have been two major periods with negative net cash flow to the fund, i.e. cash outflow has been higher than cash inflow. According to our expectation, the Global Financial Crisis was accompanied by the lack of trust of investors to financial institutions, and 'Top Stocks' has been suffering from very low net cash flows (close to zero) during years 2008 and 2009. The most dramatic fall occurred during the economic recession in the Czech Republic, where more than 300 mil. CZK fled from the fund in April 2013 based on our analysis of monthly reports.

Secondly, AUM rises/falls based on the market value of already invested funds, therefore a well-functioning investment strategy generates stable growth of AUM. Important to mention here is that this source affects AUM development substantially less than a direct cash inflow to the fund. See Figure 11 for details of historical development of AUM.

Figure 11: Historical development of AUM of 'Top Stocks'



Description: Figure 11 shows a historical development of assets under management of 'Top Stocks' in mil. CZK. The long-term exponential trend is depicted by a red-dotted line. This trend supports the effectiveness of a fund management and a stock-picking strategy. Note that gridlines are kept to ensure better orientation.

Source: Author's own analysis of monthly reports

An exponential trend in AUM of 'Top Stocks' is represented by a red-dotted line with corresponding R^2 equal to 0.9357. It is caused mainly by the excessive cash inflow to the fund in reaction to the CNB exchange rate commitment (see Appendix B), and it also supports the uniqueness of a stock-picking investment strategy. There has been slight decline during 7th and 8th year of the fund's existence (economic recession in the Czech Republic), however, the market value of total assets rises sharper than a long-term historical trend since then. Based on this graph, we can obviously conclude that a portfolio of 'Top Stocks' performs above expectations in terms of asset management and its investment strategy seems to be well-functioning. Also, high value of R^2 of the exponential trendline allows us to perform a forecast analysis with sufficiently high reporting value. Results for different time horizons are presented in the following table:

Table 2: Forecast Analysis of AUM in mil. CZK, 95% CI

period	2/2019	2/2020	2/2021
Forecasted value	16 882	19 962	23 042
Upper limit	20 528	28 618	37 953
Lower limit	13 236	11 306	8 132
range CI	7 292	17 312	29 821

Description: Results of a forecast analysis of AUM development in mil. CZK for the following 3 years using the confidence interval of 95%.

Source: Author's own analysis

Using the most frequent confidence interval of 95%, we have obtained forecasts of total volume of assets of 'Top Stocks' for three following years. For the period ending in February 2019 (note that we use data from the fund's establishment in 2006 till February 2018 in this thesis), the forecasted value of 16 882 mil. CZK has the *lower confidence bound* approximately equal to the value of AUM in February 2018. Therefore, we can expect a rise of AUM with 95% confidence during the following year.

This information is very important for both current and future investors. With sufficient accuracy, we can say that the size of 'Top Stocks' will increase and more funds would be available to invest via stock-picking strategy, that historically yields higher returns than passive investment products. Nevertheless, we are not able to distinguish between market appreciation of current AUM and a new cash inflow from investors in this forecast. However, based on a monthly cash inflow[25] analysis, we can conclude that cash inflow generates a larger share of AUM growth during last years. Therefore, it is more likely that the considerable increase of AUM during the following period(s) will consist of positive cash inflow to the fund, which overshadows the capital appreciation when analysing the entire AUM development. Also, we can expect a downward pressure on the total expense ratio (TER), which will attract the attention of new investors and further support cash inflow to the fund.

GICS summary

Finally, there is a short overview of a GICS structure of 'Top Stocks'. Following the approach presented in previous chapters, sub-industry data were assigned to industries and industry groups according to the GICS methodology manual – see Appendix A for details of this procedure.

The core of this analysis is represented by Table 3 with basic statistics related to GICS industry groups (11 groups + cash 'sector'), and a correlation matrix depicted in Table 4, which captures relationships between a GICS structure of 'Top Stocks' and its performance. Both tables provide us with synoptic overview of a main part of our dataset. Note that for reporting purposes, we present results for industry groups only.[26]

[25] Corresponding analysis is performed in Appendix B
[26] For further details about industries and sub-industries, please contact the author of this thesis.

Table 3: GICS basic statistics – a percentage share in a portfolio

Statistic	Energy	Transportation	Retailing	Food & Staples Retailing	Household & Personal Products	Pharma & Biotech	Software & Services	Technology Hardware	Cash	Semiconductors	Consumer Durables & Apparel	Consumer Services
Mean	7.7	3.9	12.8	4.5	4.1	30.8	5.4	11.9	1.9	16.1	7.8	9.8
Median	7.3	4.0	11.9	3.8	4.1	31.7	5.3	11.9	1.8	15.4	7.5	9.2
Mode	6.9	4.0	11.6	3.7	3.7	33.0	3.2	13.3	1.9	18.4	8.6	7.6
Range	9.2	4.1	16.9	6.0	2.7	27.9	9.0	12.3	18.2	16.1	12.2	9.4
Minimum	3.3	1.7	6.6	2.7	3.1	17.9	2.8	6.8	-6.5	9.3	2.3	6.7
Maximum	12.5	5.8	23.5	8.7	5.8	45.8	11.8	19.1	11.7	25.4	14.5	16.1
Kurtosis (x)	0.49	0.64	1.41	1.05	1.62	-0.42	0.11	0.14	1.77	-0.78	-1.21	0.04
Skewness (x)	0.29	-0.28	1.03	1.54	0.74	0.10	0.86	0.16	0.34	0.42	0.17	0.96
Count (x)	96	68	124	46	77	138	26	138	133	138	129	122

Description: Table 3 presents basic statistics of industry groups' share in a portfolio of 'Top Stocks'. Notation *(x)* indicates that coefficients of skewness and kurtosis are pure numbers with no units, whereas other statistics are in percent. Various levels of green fill indicate the highest values in the table, which are commented in detail below.

Source: Author's own analysis of monthly reports

This table contains 11 industry groups, which were constructed in accordance to the GICS methodology manual by analysing original sub-industry data from monthly reports. Important to mention here is that companies from 8 out of 11 industry groups were not included in a portfolio for the whole period subjected to testing. This is in line with different development of industries and preferences of investors that are built on current sectoral performance expectations. However, as the corresponding value of descriptive statistics is highly influenced by the number of observations, we have to be very careful with interpretation of results.

As an alternative procedure providing a comparison to Table 3, we have also performed the analysis where all blank cells, i.e. months with no companies from corresponding industry groups, has been replaced by zero. See Appendix C for details.

At first sight, there are three core industry groups where 'Top Stocks' gathers its assets for the whole period of its operation:

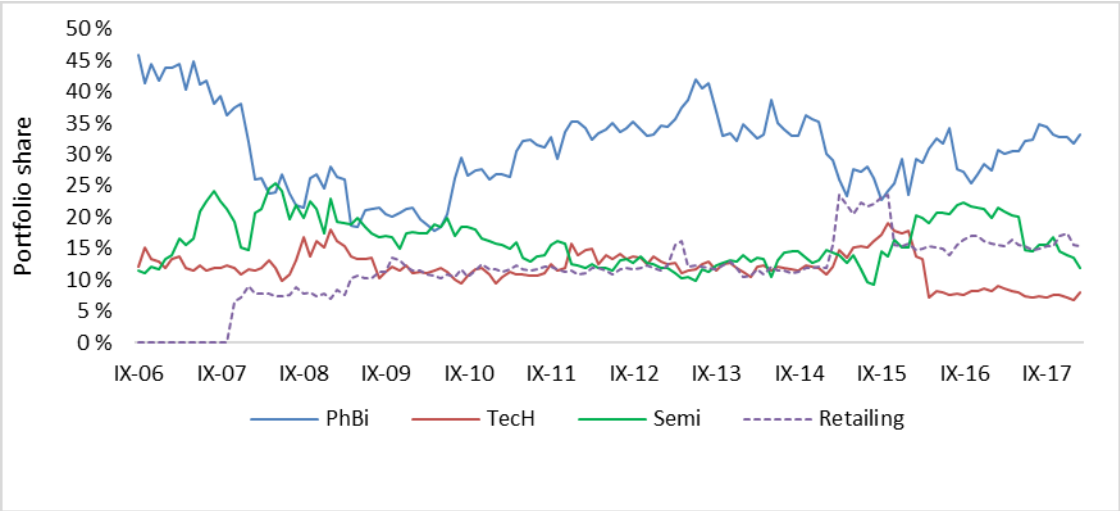
i) *Pharmaceuticals, Biotechnology and Life Sciences (PhBi)*

ii) *Semiconductors and Semiconductor Equipment (Semi)*

iii) *Technology Hardware & Equipment (TecH).*

Additionally, *Retailing* seems to be comparable with these core groups as its statistics are very similar to *TecH* industry group, and 124 observations seem stable enough to be compared with the top three groups listed above.

Figure 12: Historical development selected industry groups' share in a portfolio



Description: Figure 12 represents a historical development of four industry groups' percentage share in the portfolio of 'Top Stocks', which were selected based on the previous analysis. With the exception of *Retailing*, they are included in the portfolio for the whole period of the fund's operation.

Source: Author's own analysis of monthly reports

Clearly, *PhBi* represents the core of the portfolio. With mean and median values around 30%, this industry group has a significant impact on a portfolio's performance. Additionally, even its minimum value is above mean values of all other industry groups, and, calculations based on the total number of observations are valid as companies from *PhBi* industry group occur in a portfolio of 'Top Stocks' ever since. A historical development of *PhBi's* stake in a portfolio exhibits a seasonal behaviour. This can have two causes that are highly correlated. Either the seasonality is present on the market, or, purchases and sales of shares follow some cycle.

To uncover market forces, we have compared *EURO STOXX® TMI Pharmaceuticals & Biotechnology Index* and *S&P Pharmaceuticals Select Industry Index*. As expected, both were negatively affected by the Global Financial Crisis during 2008 – 2009. Market value of shares in the portfolio of 'Top Stocks' has declined and based on monthly reports' analysis, no significant purchase or sale related to *PhBi* industry group occurred during that period. Conversely, both indices exhibit a sharp

increase in 2015, whereas *PhBi's* stake in a portfolio falls. This points out that at least one investment idea from this industry group has been excluded from the portfolio. By a detailed look at Figure 12, we can identify that three ideas were excluded and soon replaced by another during 2014 – 2015 period. Reason for this is clear: *PhBi* faced a so called 'market overheating' represented by unsustainable growth, which brings significant risks for investors. Based on a flexible stock-picking strategy, some changes in a portfolio structure were necessary to prevent potential losses.

We have selected one moment from the above mentioned period and examined in detail, what changes in a portfolio structure influencing *PhBi* industry group have occurred. In December 2014, shares of Threshold Pharmaceuticals were sold due to the fact that they were not fulfilling the original investment expectations. Instead, shares of a biotechnology company Relypsa were bought, after its product Veltassa has been approved by the US regulator (FDA) for the treatment of hyperkalemia[27] having a considerable potential, which has not been valued properly by the market at that time.

A cash 'sector', which is obviously not included in a GICS scheme, represents free funds, i.e. total assets held in cash or cash equivalents that are not invested in a concrete investment idea during particular month. Also, some level of cash should be held by the fund for liquidity purposes.

In the short term, cash level can rise due to cash inflows from investors. A stock-picking strategy requires a detailed analysis, which is very time-consuming, therefore having free cash is a pre-step for the upcoming investment, which could be realised in a horizon of several months. If only limited amount of cash is available, the fund loses the chance to take the advantage of investment opportunities immediately.

An interesting feature visible from a historical development of cash levels of 'Top Stocks' is that their average is rising during last years of the fund's operation. Due to current high valuations, there is a trade-off between investing the cash and facing the risk of losing a performance power in the case when financial markets will continue to rise. Also, overheated markets might experience a sudden crash sooner or later, as their growth is unsustainable in the long-term. This will be accompanied by the period of recovery of stock prices. Therefore, staying in cash and waiting for a market calm-down is the safer option than investing and facing serious risks.

[27] increase of potassium levels in blood affecting patients with chronic kidney disease and heart failure

Table 4: Correlation matrix of GICS industry groups' stake in a portfolio

Industry Group	Energy	Transportation	Retailing	Food retailing	Household products	Pharma. & Biotech.	Software & Services	Technology Hardware	Cash	Semiconductors	Consumer Durables	Consumer Services
Energy	1											
Transportation	0.627	1										
Retailing	-0.747	-0.375	1									
Food & Staples Retailing	0.603	0.358	-0.778	1								
Household & Personal Products	0.630	0.370	-0.661	0.618	1							
Pharma., Biotech. & Life Sciences	0.158	-0.204	-0.433	0.116	-0.108	1						
Software & Services	-0.514	-0.317	0.327	-0.283	-0.487	0.058	1					
Technology Hardware & Equip.	0.245	0.249	-0.092	0.210	0.123	-0.092	-0.603	1				
Cash	-0.178	-0.258	-0.097	0.062	-0.130	0.198	0.075	-0.215	1			
Semiconductors & Equipment	0.035	-0.015	-0.209	0.265	0.273	-0.410	0.092	-0.219	-0.093	1		
Consumer Durables & Apparel	-0.781	-0.612	0.737	-0.667	-0.580	-0.216	0.153	-0.067	-0.009	-0.138	1	
Consumer Services	-0.334	0.066	0.605	-0.528	-0.095	-0.663	0.102	-0.174	-0.345	-0.063	0.347	1

Description: Table 4 represents a correlation matrix of industry groups included in the portfolio of 'Top Stocks' used to analyse inter-industry relationships and uncover the trade-off between industry groups that a portfolio manager faces to mitigate the risk accompanying a highly concentrated portfolio. Read the text below to properly understand the specific relationship captured by this table.

Source: Author's own analysis of monthly reports

A correlation matrix (Table 4) has been used to analyse inter-industry relationships and a portfolio performance. Colours of the cells indicate the strength of a correlation, which will be further analysed. Of course, the value of a correlation between two identical industry groups is equal to one.

Before we proceed to a detailed description of individual coefficients, we have to emphasize what relationship this matrix captures as its improper interpretation would lead to creating misleading connections between GICS structure and the fund's performance.

Input data consist of sectoral breakdowns obtained from monthly reports for each month covered by our analysis. The processes of manual transcription and related procedures to eliminate typing errors have been described in the previous chapter. The percentage stake of each industry group in the portfolio indicates not only changes in a portfolio structure due to purchases and sales of shares, but it takes market development into account at the same time. If one industry group faces economic recession, market value of firms having corresponding GICS code will decline and so will its stake in a portfolio of 'Top Stocks'. Conversely, if other industry group experiences a boom, its portfolio share can overshadow the purchase of shares of companies from different industry groups.

Having said that, the relationship presented in Table 4 becomes clearer, as both changes in the portfolio structure caused by purchases/sales of investment ideas and market development are included in input data. Therefore, this table can be interpreted as desired relationships between industry groups by a portfolio manager and his team with respect to market development. Important to mention here is that a cash 'sector', which is obviously not covered by the GICS scheme, is included in our analysis to verify its independence represented by correlation coefficients close to zero, which is confirmed by this analysis.

To analyse market development, we have used *S&P Select Industry Indices* that consist of stocks included in the S&P Total Market Index classified[28] into corresponding GICS industries to provide overall performance measurement for the whole sector. Based on performance evolution of these indices, we are able to uncover reasons for changing a portfolio structure and properly interpret significant correlation coefficients in the table above. As all performance-related statistics and graphs are available online[29], we do not include this information in our thesis and focus only on their implications for a portfolio structure of 'Top Stocks'.

Firstly, we will focus on *Software & Services (SoftS)* industry group that is a part of the portfolio since May 2015, starting with 2.8% and closing the period subject to testing with 11.8%. Corresponding S&P Index exhibits a clear upward trend and since May 2015, its value becomes approximately 1.6 times higher. Also, *SoftS* seems to be relatively resistant to negative shocks and its long-term appreciation (10 Yr. ann. returns +14.47%) since the Global Financial Crisis confirms why a concentrated portfolio of 'Top Stocks' includes shares from this industry group and their stake in a portfolio rises each month.

As oppose to expanding *SoftS* market, a negative correlation identifies sectors suffering from contraction at the time when *SoftS* rises. For example, if we look at *Energy* industry group's S&P Index, we will identify a period of performance dropout starting in August 2014 caused by low commodity prices. Not surprisingly, this sector has been completely excluded from the portfolio of 'Top Stocks' in September 2014. This confirms that a constant market analysis that is necessary for a stock-picking strategy responds to market changes immediately to ensure positive returns and preserve investors' trust.

---

[28] For details, see S&P Dow Jones Indices: S&P Select Industry Indices Methodology, February 2018
[29] www.us.spindices.com

Energy industry group is accompanied by the strongest correlation coefficients in the correlation matrix. Apart from *SoftS*, *Retailing* and *Consumer Durables & Apparel* industry groups are negatively correlated with *Energy*, which implies upward trend in their performance. A detailed look at S&P Retail Select Industry Index would explain all five strong correlations in the table, as market expansion is clearly evident from the chart. Interesting finding related to *Retail* industry group is that it has been considered safe enough during the Global Financial Crisis as it firstly occurs in a portfolio of 'Top Stocks' in November 2007 and its portfolio share is still growing to this day, as well as the value of S&P Retail Index, which is six times higher than at the time of Lehman Brothers' crash in 2008. As explained above, *PhBi* industry group has been influenced by the Crisis heavily and its portfolio share declined substantially, represented by a negative correlation coefficient of -0.433 with *Retailing*.

An example that cannot be described by a market development is a positive correlation of 0.627 between *Energy* and *Transportation* industry groups. Therefore, we have searched through monthly reports to find the reason directly at a company-level for this unexpected relationship. In March 2013, shares of Avis Budget Group (*Transportation*) were sold as their current market valuation fully reflected the company's growth potential. In September 2014, a company Dresser-Rand (*Energy*) has been excluded from the portfolio because of the announcement of the firm's takeover. After that, no shares from these two GICS industry groups were included in the portfolio, which results in a positive correlation between these two groups, even though corresponding S&P Select Industry Indices exhibits a negative relationships in a long-term performance development.

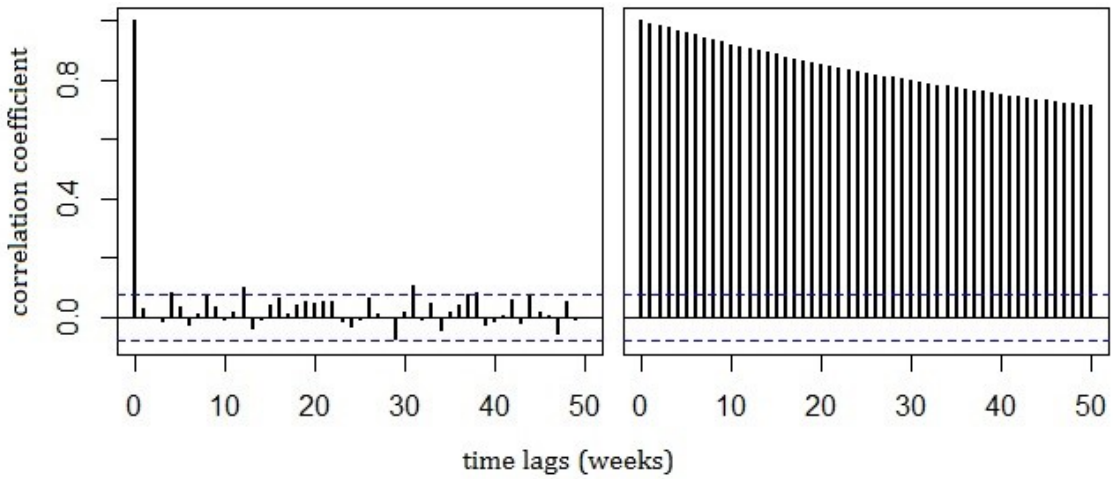
This reveals the potential weakness of the interpretation of coefficients from this table. Because of a highly concentrated portfolio consisting of 25 investment ideas at maximum, the relationship cannot necessary reflect the links between industry groups, but it rather refers to individually-specific firm level, i.e. it shows a correlation between individual investment ideas sometimes. Therefore, it is necessary to compare S&P Industry Select Indices with our market development assumptions and when finding unexpected relationships, search for details in monthly report to directly identify moments that are behind these numbers.

Limitations of our testing procedures and discussion for refining our analysis are presented in Section 6.3.

6.2 Regression and forecast analysis

Following the methodology approach described in Chapter 4, results of statistical tests and regression models will be presented in this section. Firstly, we will comment on 'correlograms'[30], which represent the output of a statistical software R, when performing a graphical analysis of correlation between different lags of returns and a fund share unit price. On these simple charts, the concept of stationarity will be discussed and we will compare results with our theoretical assumptions for returns and stock prices based on the previous research. Secondly, stationarity will be tested in detail using proper statistical tools. Finally, results of regression and forecast analyses will be presented.

Figure 13: Correlograms of weekly returns and weekly NAV, respectively



Description: Autocorrelation function calculates correlation coefficients between time lags of a univariate time series and offers a graphical plot shown above. From these correlograms, we can directly observe that weekly simple returns (left) are uncorrelated and satisfy the definition of stationarity, whereas weekly NAV (right) obviously exceeds the threshold depicted by horizontal dashed lines.

Source: Author's own analysis using R

In line with our theoretical assumptions and literature review, autocorrelation of returns and a fund share unit price differs substantially, which can be directly observed from Figure 13. Weekly returns exhibit correlation between limits with exception of 12th and 31st lag, however, there is no reason to analyse the source of such behaviour, because there is no rational theory behind this performance and we can consider the length of 12,

[30] The plot of the autocorrelation function (acf) versus time lag is called a correlogram

resp. 31 weeks long enough to have an insignificant impact on today's portfolio return. Based on this chart, weekly returns can be considered stationary. On the other hand, net asset value clearly breaks the upper limit (95% CI) and no negative correlation occurs during the time of 50 weeks (50 lags). This indicates a non-stationary behaviour of stock prices, and we can suggest that further testing would prove the presence of a random walk. Additionally, a very slow decline of correlation coefficients' values supports our analysis of trends from the previous section.

While the results of descriptive and graphical analysis are in line with our findings and conclusions built on autocorrelation of returns and NAV, we have to use special statistical tests to support our claims by methods and procedures developed by experienced statisticians, who specialised in time series. We performed the Augmented Dickey-Fuller (ADF) test and obtained following results:

Table 5: Augmented Dickey-Fuller test

	r_w	r_m	NAV	$avgNAV$
Dickey-Fuller	-7.29	-7.20	-2.94	-2.73
p-value	<0.01	<0.01	0.18	0.27
result	**stationary**	**stationary**	**non-stationary**	**non-stationary**

Description: Table 5 summarizes outcomes of the Augmented Dickey-Fuller test, which has been used to verify our claims about (non)-stationarity of simple returns and NAV from the Section 6.1.

Source: Author's own analysis using R

As mentioned in Section 4.4, ADF test belongs to a group of unit root tests, because its null hypothesis is a unit root (non-stationary), and the alternative hypothesis stands for stationarity. Therefore, small p-value implies rejection of H_0, which is the case of both weekly and monthly returns. When using the most common confidence interval across statistical researches (95%), first-differencing is required for p-values greater than 0.05, which holds for the net asset value and its monthly averages.

Based on Černý and Kočenda's (2007) indication of higher chance of Type II error of not rejecting incorrect null hypothesis when using ADF test, we will compare the results with the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test, which belongs to the group of stationarity tests, because level or trend stationarity is its null hypothesis, whereas H_A is a unit root process, i.e. it reverses hypotheses of the ADF test. Results are summarised in Table 6.

Table 6: Kwiatkowski–Phillips–Schmidt–Shin test

	r_w	r_m	NAV	$avgNAV$
KPSS (level)	0.67	0.14	8.83	4.15
p-value	0.016	>0.01	<0.01	<0.01
result	**non-stationary**	**stationary**	**non-stationary**	**non-stationary**

Description: Table 6 summarizes the outcomes of the KPSS test, which has been performed in order to verify results obtained from the ADF test. For weekly returns denoted r_w, we have obtained a different result. As the KPSS test suffers from a high frequency of Type I error, i.e. it rejects the null hypothesis too often, we have to perform additional testing to support either ADF or KPSS result.

Source: Author's own analysis using R

If one-sided LM-statistics, which is used for this test, is greater than a corresponding critical value (see Kwiatkowski et al. (1992), Table 1), H_0 is rejected and time series is non-stationary. For 95% CI, the critical value equals 0.413 for a linear stationarity. Compared to results of the ADF test, which are summarised in Table 5, we have obtained a different result for weekly returns (r_w). This is likely caused by the fact that the KPSS test with reversed hypotheses suffers from a higher frequency of Type I errors, as it tends to reject stationarity too often.

Therefore, we have also carried out the Phillips-Perron (PP) unit root test, which has same hypotheses as the ADF test. For both weekly and monthly returns, we obtained a p-value smaller than 0.01, which obviously rejects H_0 of unit root and confirms stationarity, same as the ADF test. For NAV and $avgNAV$, we got p-values equal to 0.22 and 0.29 respectively, which corresponds to both ADF and KPSS statement of non-stationarity.

Following the approach described in Section 4.4, we performed a regression analysis of weekly NAV using the Box-Jenkins method to fit the ARIMA model. Results presented above provide us with sufficient knowledge about our dataset, therefore, we have enough information about our time series and we can proceed to a deeper analysis of the portfolio's performance. Using auto.arima function in R, we have identified the most appropriate values for p, d and q of the ARIMA model and obtained following results:

Table 7: Fitting ARIMA model to historical NAV per share data

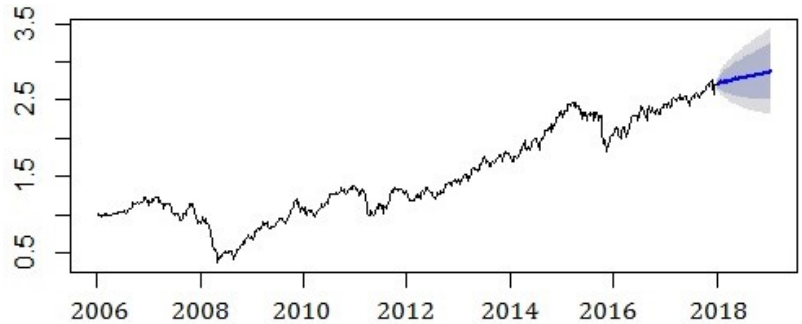
identified model	**ARIMA (3,1,2) with drift**				
	Coefficient	Standard error		t-ratio	Significance
ar1	-0.2703	0.0505	-	5.3525	***
ar2	-0.9522	0.0257	-	37.0506	***
ar3	-0.0664	0.0424	-	1.5189	
ma1	0.2679	0.0310		8.6419	***
ma2	0.9509	0.0220		43.2227	***
drift	0.0029	0.0016		1.8125	
sigma_sq_est	0.0017			*AIC*	-2122.84
log-likelihood	1068.42			*BIC*	-2092.07

Description: Results of the Box-Jenkins method used to find the best fit to historical values of weekly NAV are presented in Table 7. We have disallowed all short-cuts in the procedure (stepwise, approximation) to obtain the most precise fit to past data. The accuracy of this fit has been verified by a graphical analysis.

Source: Author's own analysis using R (Box-Jenkins method)

For comparison, when allowing short-cuts in the automated procedure of fitting the model, we would obtain $ARIMA(0,1,0)$ with a drift with non-significant drift estimate. Also, we would end up with lower value of log-likelihood and higher values of Akaike's Information Criterion (AIC)[31], whereas the goal of the model is to minimise the AIC value. Therefore, by restricting simplification of auto.arima function, we ended up with a more precise model that better fits our data. To verify this claim, we have compared both actual and fitted values on a time series plot. As they almost overlap, we can conclude that this fit corresponds to real development of historical NAV very well.

Figure 14: Forecast of weekly NAV per share from $ARIMA$ (3,1,2) model with drift



Description: Figure 14 shows forecasted values of weekly NAV in CZK for the next 52 weeks. Also, confidence intervals (80% and 95%) are depicted.

Source: Author's own analysis using R (forecast package)

[31] $AIC = -2\log(L) + 2(p + q + k + 1)$, where L is the log-likelihood, k=1 for models including a drift

Then, we performed a deep analysis of forecasted model's residuals. At first, we have constructed a Q-Q plot to check whether residuals follow a Gaussian distribution. Majority of the data lies on a desired diagonal line, which is the result we expected. Secondly, we have verified that the autocorrelation of residuals is within a threshold using the acf function in R.

After that, we have excluded last 52 observations from the regression and set the forecast to 104 weeks, i.e. two years. This procedure allows us to compare actual and forecasted values. On average, there has been a 3.4% difference between actual and forecasted values during the last year. Taking into account determinants of stock prices and unexpected shocks that accompany the world's largest stock markets, the calculated difference is acceptable and we can conclude that $ARIMA\,(3,1,2)$ captures the development of NAV very well. Therefore, we can consider its forecast as reliable.

Finally, results of the Ljung-Box test (1978) for detecting whether residuals of our model resemble 'white noise', i.e. a random process with zero mean, constant variance over time and no autocorrelation between time lags, are summarised in the following table:

Table 8: Ljung-Box Test for fitted residuals

df	5	10	15	20	25
X_sq	1.3242	4.8545	7.4862	13.046	13.672
p-value	0.9324	0.9007	0.9427	0.8754	0.9672

Description: Table 8 summarizes outcomes of the Ljung-Box test for fitted residuals. With very high p-values, we fail to reject the null hypothesis of independently distributed data. Therefore, residual diagnostics have been successful and support the effectiveness of our model.

Source: Author's own analysis using R

We fail to reject the null hypothesis of independently distributed data, as all p-values are very close to one. This is exactly what we were looking for when performing deep residual diagnostics, which also support the suitability and relevance of our fitted model.

With additional testing that supports the fitted $ARIMA\,(3,1,2)$ model and a forecast analysis, we can now interpret the added value of our findings for investors. Obviously, the upward trend in NAV could be seen as the reason for immediate purchase of mutual fund units. Of course, this will be preceded by a comparison of expected returns of other investment instruments, however, a forecasted value of 2.8703 CZK, representing

a 5.14% increase of NAV from the last week of February 2018 during the following year, indicates that investing in 'Top Stocks' should outperform returns from other products available at the financial market, especially those affected by low interest rates or passively managed funds.

Important to mention here is that every forecast is accompanied by confidence intervals and the final decision whether to invest or not depends on risk preferences of every individual investor. Risk-seeking investors would focus on forecasted values of NAV and upper confidence bound, whereas risk-averse individuals would still see the threat of potential loss. As everyone has its own investment strategy and future plans, a forecast analysis performed using the Box-Jenkins method does not necessarily cover all investors' requirements and more testing with the inclusion of individual characteristics should be done. Therefore, to satisfy various levels of risk preferences and expectations of investors, we have performed a *Monte Carlo Simulation* of weekly NAV development for the period of next 600 weeks.[32] As the theory behind these computational algorithms built on a random sampling is outside the scope of the bachelor's degree curriculum, brief description and results are presented in Appendix D.

[32] The same length as covered by our analysis (approx. 11.5 years)

6.3 Limitations and Discussion

During our research, we have to face many limitations and ambiguity of interpretation, which can lead to inaccurate conclusions. In this section, we will describe some of the major issues and suggest alternative methods to avoid them in future studies.

The first drawback is related to the data collection and processing. As no previous study of this topic exists, all the data had to be collected in sufficient advance from the portfolio manager of 'Top Stocks' and a manual transcript from .pdf format to .xlsm and .csv formats was necessary. Even though we performed a sumcheck to minimize the risk of typing errors, an automation of this procedure would be a great advantage that would not only lower the risk of errors, but also the time needed to perform this procedure. For the analysis of larger time period of the fund's operation, this would be essential.

Next drawback is connected with the GICS structure and information contained in monthly reports. For reporting purposes, we have to restrict[33] our analysis to industry groups, as more detailed analysis would significantly expand the volume of this thesis. Also, it would not allow us to perform various other tests related to our research question. Even though we have followed the GICS methodology manual, we have lost some specific information related to industries and sub-industries when merging them into industry groups. To prevent this issue, future studies and empirical theses should focus on GICS structure only and do not perform a broad analysis of the portfolio's performance.

Next, we had to exclude GICS data from a regression analysis and perform alternative procedures. Even though we have collected the data since the fund's establishment in 2006, having 138 observations seems insufficient for regression models. Many studies, including Green (1991), Morgan et al. (2007) and others have analysed the minimum acceptable sample size needed to obtain results with adequate reporting value. Our sample size would pass some of the suggested methods; however, more data will definitely increase the accuracy of constructed models. In combination with automation of data transcript from .pdf to .xlsm format mentioned at the beginning of this section, this would make the regression analysis more precise.

Finally, we have to be careful with interpretation of industry and firm-specific relationships, which often coincide in concentrated portfolio analyses.

[33] See Appendix A for more details

7 Conclusion

This thesis provides literature and methodology overview and results of a broad analysis of the Czech mutual fund 'Top Stocks' and its performance. As no previous public research of this fund has been performed during its twelve-year history, the author collected all the data herself and merged both publicly available data with uniquely collected reports requested directly from a portfolio manager. This results into a complex dataset of weekly and monthly statistics including a *Global Industry Classification Standard* (GICS) sectoral breakdowns, information about assets under management, the fund's cash flow, net asset value per share etc. This dataset covers the whole period subject to testing from the start of the fund's operation in 2006 to February 2018 with no missing values. An extensive manual transcript of the data has been checked for potential typing errors and data validity has been verified using several procedures.

At first, we focused on descriptive and graphical analysis of returns and net asset value. Using the Augmented Dickey-Fuller test, supported by the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test and the Phillips-Perron (PP) unit root test to control for errors (Type I, Type II) accompanying hypotheses testing, we came to conclusion that returns are stationary and the net asset value is integrated of order one, I(1), i.e. stationary in first differences. These results are in line with our assumptions built on previous researches of similar topics, and, they were used later in a regression and forecast analysis, where the concept of stationarity plays a major role.

Next, we have used a historical development of the total volume of assets to forecast its future development and comment on implications it brings to investors. With 95% confidence, we can say that assets under management of 'Top Stocks' will rise during the following year and the exponential trend indicates that the stock-picking strategy of this actively managed fund works well in terms of capital appreciation and building investors' trust. Also, investors can benefit from economies of scale accompanied by the downward pressure on the Total Expense Ratio, which will attract the attention of potential investors and support the cash inflow to the fund.

The analysis of the GICS monthly sectoral breakdowns uncovers three main industry groups, in which 'Top Stocks' allocates its funds for the whole period of its operation. These are Pharmaceuticals, Biotechnology and Life Sciences; Semiconductors and Semiconductor Equipment and Technology Hardware & Equipment. The

performance of these industry groups can be easily compared to S&P Select Industry Indices available online that offer a good benchmark when taking into account the GICS structure, which is currently seen as the most appropriate industry classification supported by the broad comparison of the available classification schemes performed by Bhojraj et al. (2003).

Then, the correlation matrix indicates that both market development and a firm-specific level could affect the desired structure of the portfolio and a trade-off between industry groups. We have identified both expected relationships, which correspond to the development of S&P indices, and unexpected relationships, which were studied in detail on a firm-specific level. In contrast with similar studies that use a correlation matrix to identify inter-industry relationships, we have to be careful with interpretation as having a highly concentrated portfolio can easily reduce the industry relationship to a specific correlation between two investment ideas. Based on the analysis of monthly reports, most unexpected coefficients occur when investment ideas were excluded from the portfolio at the same time from different reasons. Then, the correlation coefficient does not reflect the true relationship between industry groups, but is influenced by random events. Also, we have verified that cash and cash equivalents are uncorrelated with all industry groups. Last but not least, the correlation matrix has been used to prove the flexibility and operating effectiveness of a stock-picking strategy on a market dropout of Energy industry group in August 2014, when this sector has been immediately excluded from the portfolio of 'Top Stocks'.

Another contribution of this thesis is the regression and forecast analysis of weekly net asset value. Using a statistical software R, we came through all stages of the Box-Jenkins test and fit $ARIMA\,(3,1,2)$ model with drift. These stages are: data preparation, model selection, estimation of MLE parameters, residuals diagnostics and finally a forecast analysis. Results are summarised graphically, and, we have also calculated the average difference between actual and forecasted values during the last year. Having a 3.4% difference in weekly NAV on average during the last 52 weeks, we came to conclusion that both fitted and forecasted values are very accurate. Of course, our analysis is accompanied by confidence intervals of 80% and 95%.

Furthermore, we have performed a *Monte Carlo Simulation* of possible development of weekly net asset value in the next 12 years (the period subject to testing in this thesis). As the set of computational algorithms built on a random sampling exceeds the technical level of a bachelor's thesis, results are shown in Appendix D.

To summarise, we have performed a complex analysis of the performance of 'Top Stocks'. The uniqueness of this thesis lies in creation of a complex dataset, which is the first of its kind and scope. The emphasis is put on the GICS sectoral breakdown and inter-industry relationships that form the core of a stock-picking strategy. The overall scope of this topic provides wide opportunities for future research, which can build on our findings and conclusions, and our dataset covering the whole period of the fund's operation till February 2018 is a great asset for future studies of 'Top Stocks'.

Bibliography

AAS, K. (2004). To log or not to log: The distribution of asset returns. *Norwegian Computing Center*.

BEAULIEU, M. C., DUFOUR, J. M. and KHALAF, L. (2003). Exact Skewness-Kurtosis Tests for Multivariate Normality and Goodness-of-Fit in Multivariate Regressions with Application to Asset Pricing Models. *Oxford Bulletin of Economics and Statistics*, 65(1):891–906.

BHOJRAJ, S., LEE, C. M. C. and OLER, D. (2003). What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research*, 41:745–774.

BIGWOOD, S. and SPORE, M. (2003). Presenting Numbers, Tables, and Charts. *Oxford University Press*.

BOX, G. and JENKINS, G. (1970). Time Series Analysis: Forecasting and Control. *Holden-Day, San Francisco.*

BROWN, S. J. and WARNER, J. B. (1985). Using Daily Stock Returns - The Case of Event Studies. *Journal of Financial Economics*, 14(1):3–31.

ČERNÝ, A. and KOČENDA, E. (2007). Elements of Time Series Econometrics: An Applied Approach. *Karolinum Press*.

CHAUDHRY, M. and CHRISTIE-DAVID, R. (2001). Coskewness and cokurtosis in futures markets. *Journal of Empirical Finance*, 8(1):55–81.

CHEN, J., HONG, H. and STEIN, J. C. (2001). Forecasting crashes: Trading volume, past returns, and conditional skewness in stock prices. *Journal of Financial Economics*, 61(3):345–381.

CLARKE, R. N. (1989). SICs as Delineators of Economic Markets. *The Journal of Business*, 62(1):17–31.

COLLINS, S. and DUVALL, J. (2017). Trends in the Expenses and Fees of Funds, 2016. *ICI Research Perspective*, 23(3).

CRUIJSEN, van der C., HAAN de J., and JANSEN, D. J. (2013). Trust and Financial Crisis Experiences. *De Nederlandsche Bank*, Working Paper no. 389.

DALRYMPLE, D. J. (1978). Using Box-Jenkins techniques in sales forecasting. *Journal of Business Research*, 6(2):133–145.

DICKEY, D. A. and FULLER, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of American Statistical Association*, 40(366):427–431.

DICKEY, D. A. and FULLER, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*, 49(4):1057–1072.

EKHOLM, A. and, PASTERNACK, D. (2007). Overconfidence and Investor Size. *European Financial Management*, 14(1):82–98.

ELTON, E. J. and GRUBER, M. J. (1997). Modern Portfolio Theory 1950 to Date. *Journal of Banking & Finance*, 21:1743–1759.

FAMA, E. F. (1965). The Behaviour of Stock Market Prices. *Journal of Business*, 38(1):34–105.

FAMA, E. F. (1976): Efficient Capital Markets: Reply, *The Journal of Finance*, 31(1):143–145.

FAMA, E. F. and FRENCH, K. R. (1997). Industry costs of equity, *Journal of Financial Economics*, 43:153–193.

Finlord.cz (2017). Jan Hájek: aktivní fondy nikdy nevymřou. *Finlord.cz*, interview [online]. Available at: https://finlord.cz/2017/10/jan-hajek-aktivni-fondy-nikdy-nevymrou/ [Accessed April 6, 2018]

GIBBONS, M. R. and HESS, P. (1981). Day of the week effects and asset returns. *Journal of Business*, 54(1):579–596.

GONG, X. (2014). Technical Trading Strategies: Skewness-Seeking Games. *Tilburg University*, Department of Economics and Management.

GRANGER, C. W. and NEWBOLD, P. (1974). Spurious Regressions in Econometrics. *Journal of Econometrics*, 2(2):111–120.

GREEN, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariete Behavioral Research,* 26(3):499–510.

GUENTHER, D. A. and ROSMAN, A. J. (1994). Differences between COMPUSTAT and CRSP SIC Codes and Related Effects on Research. *Journal of Accounting and Economics,* 18:115–128.

HILL, G. W. and WOODWORTH, D. (1980). Automatic Box-Jenkins Forecasting. *The Journal of the Operational Research Society*, 31:413–422.

HYNDMAN, R. J. and ATHANASOPOULOS, G. (2013). Forecasting: principles and practice. *OTexts*. [Accessed on April 15, 2018]

KAHLE, K. M. and WALKING, R. A. (1996). The Impact of Industry Classifications on Financial Research. *The Journal of Financial and Quantitative Analysis*, 31:309–335.

KNACK, S. and KEEFER, P. (1997). Does Social Capital Have an Economic Payoff? A Cross-Country Investigation. *Quarterly Journal of Economics*, 112(4):1252–1288.

KRISHNAN, J. and PRESS, E. (2002). The North American Industry Classification System and Its Implications for Accounting Research. *Temple University*, Working Paper.

KWIATKOWSKI, D., PHILLIPS, P., SCHMIDT, P., SHIN, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. *Journal of Econometrics*, 54(1-3):159–178.

LAWRENCE K. D., KLEINMAN G. and LAWRENCE S. M. (2015). Time Series Models to Predict the Net Asset Value (NAV) of an Asset Allocation Mutual Fund VWELX. *Handbook of Financial Econometrics and Statistics.* Springer, New York.

LJUNG, G. M. and BOX, G. E. P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika,* 65(2):297–303.

LO, A. and MacKINLEY, A. C. (1988). Stock Prices do not follow Random Walks: Evidence from a Simple Specification Test. *Review of Financial Studies*, 1:41–66.

MANIKANDAN, S. (2011). Measures of central tendency: Median and mode. J Pharmacol Pharmacother, 2:214–215.

MORGAN, B. L. and VanVOORHIS, C. R. V. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials in Quantitative Methods for Psychology*, 3(2):43–50.

MURTHY, V., WASHER, K. and WINGENDER, J. (2011). Are stock prices in the US nonstationary? Evidence from contemporary unit root tests. *Applied Financial Economics*, 21(22):1703–1709.

PEARCE, E. (1957). History of the Standard Industrial Classification (SIC). *Executive Office of the President*, Office of Statistical Standards, U.S. Bureau of the Budget. Washington 25, D. C.

PESARAN, M. H. (2010). Predictability of Asset Returns and the Efficient Market Hypothesis. *Institute for the Study of Labor*, IZA Discussion Papers.

PLEROU, V. et al. (2001). Price Fluctuations, Market Activity and Trading Volume. *Quantitative Finance*, 1:262–269.

QUIGLEY, L. (2008). Statistical Analysis of the Log Returns of Financial Assets.

ROZEFF, M. S. and KINNEY, W. R., Jr. (1976). Capital market seasonality: The Case of stock returns. *Journal of Financial Economics*, 3(1):379–402.

SAMUELSON, P. A., (1965). Proof that properly anticipated prices fluctuate randomly. *Industrial management review*, 6(2):41–49.

SAUNDERS, N. C. (1999). The North American Industry Classification System: Change on the Horizon. *Occupational Outlook Quarterly*, 34–37.

TALBERT, S. and SOLE, M. L. (2013). Too much information: research issues associated with large databases. *Clin Nurse Spec.*, 27(2):73–80.

TALEB, N. N. (2007). The Black Swan: The Impact of the Highly Improbable. *Random House*.

TUFTE, E. R. (1983, 2001). The Visual Display of Quantitative Information. *Graphics Press*.

WEINER, CH. (2005). The Impact of Industry Classification Schemes on Financial Research. Available at SSRN: https://ssrn.com/abstract=871173

WOOLDRIDGE, J. M. (2012). Introductory Econometrics: A Modern Approach. *South-Western College Pub*, 5th edition.

ZÁMEČNÍK, P. (2017). Ján Hájek: Dostal jsem příležitost řídit fond Top Stocks podle vlastní investiční filozofie. *Investujeme.cz,* interview [online]. Available at: https://www.investujeme.cz/clanky/jan-hajek-dostal-prilezitost-ridit-fond-top-stocks-podle-vlastni-investicni-filozofie/ [Accessed: April 6, 2018]

List of Appendices

Appendix A: GICS structure of 'Top Stocks'

Appendix B: Graphical analysis of monthly cash inflow

Appendix C: Descriptive statistics of GICS sectoral breakdowns

Appendix D: *Monte Carlo Simulation* of weekly net asset value per share

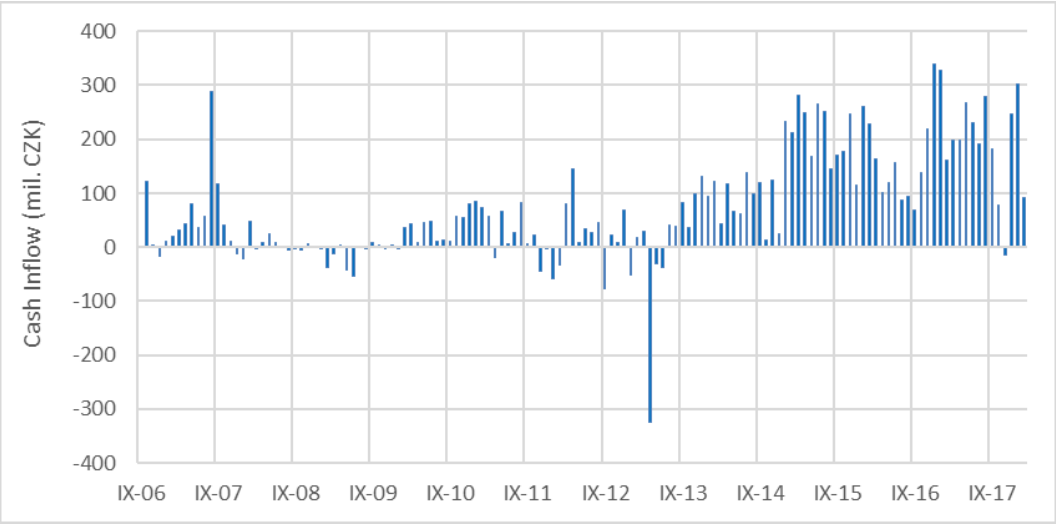
Appendix A: GICS structure of 'Top Stocks'

Description: The chart below represents a GICS structure of 'Top Stocks'. We have grouped sub-industries obtained from monthly reports to industry groups according to the GICS methodology manual. This adjustment allows us to perform a deep analysis of a portfolio performance and report results in a structured form. Numbers in brackets represent frequency of occurrence of companies from corresponding sub-industries in a portfolio of 'Top Stocks' on a monthly basis.

Industry group	Industry	Sub-Industry
Energy	Energy Equipment & Services	Oil & Gas Equipment & Services (88)
	Oil, Gas & Consumable Fuels	Oil & Gas Storage and Transportation (12)
Transportation	Marine	Marine (7)
	Road & Rail	Trucking (60)
Consumer Durables & Apparel	Textiles, Apparel & Luxury Goods	Apparel, Accessories & Luxury Goods (129)
Consumer Services	Hotels, Restaurants & Leisure	Restaurants (122)
Retailing	Distributors	Distributors (36)
	Multiline Retail	Department Stores (32)
	Specialty Retail	Apparel Retail (124)
		Specialty Stores (107)
Food & Staples Retailing	Food & Staples Retailing	Drug Retail (46)
		Food Distributors (10)
Household & Personal Products	Household Products	Household Products (15)
	Personal products	Personal products (62)
Pharmaceuticals, Biotechnology & Life Sciences	Biotechnology	Biotechnology (138)
	Pharmaceuticals	Pharmaceuticals (138)
	Life Sciences Tools & Services	Life Sciences Tools & Services (19)
Software & Services	Internet Software & Services	Internet Software & Services (13)
	IT Services	Data Processing & Outsourced Services (22)
	Software	Systems Software (5)
Technology Hardware & Equipment	Communications Equipment	Communications Equipment (112)
	Technology Hardware, Storage & Peripherals	Technology Hardware, Storage & Peripherals (137+28)
	Electronic Equip., Instruments & Components	Electronic Manufacturing Services (1)
		Technology Distributors (36)
Semiconductors & Semiconductor Equipment	Semiconductors & Semiconductor Equipment	Semiconductor Equipment (6)
		Semiconductors (138)

Appendix B: Graphical Analysis of Monthly Cash Inflow

Description: Appendix B is a graphical representation of monthly cash inflow to the fund in mil. CZK. It is clearly visible that 'Top Stocks' has been suffering from almost zero cash inflow during the Global Financial Crisis. Additionally, a deep fall occurred in April 2013, however, there is a positive cash flow since then, which supports the exponential growth of assets under management (AUM). Since July 2013 to February 2018, there is an excessive cash inflow to the fund, which is likely caused by the exchange rate commitment of the Czech National Bank. At the times of low interest rates, 'Top Stocks' has become more popular than ever as its average annual yield of 8.51% p.a. since its establishment is a way above traditional savings and investment products available on the financial market.



Appendix C: Descriptive statistics of GICS sectoral breakdowns

Description: In addition to descriptive statistics of GICS sectoral breakdowns presented in Table 3 of this thesis, we also attach results of the same analysis, however, we replaced blank cells in our GICS dataset with zero. By this adjustment, we have numerically specified that some industry groups were not included in the portfolio of 'Top Stocks' for particular months. Obviously, several statistics have significantly changed.

Statistic	Energy	Transportation	Retailing	Food retailing	Household products	Pharma. & Biotech.	Software & Services	Technology Hardware	Cash	Semiconductors	Consumer Durables	Consumer Services
Mean	5.3	1.9	11.4	1.5	2.3	30.8	1.0	11.9	1.8	16.0	7.3	8.6
Median	6.9	0.0	11.7	0.0	3.6	31.7	0.0	11.9	1.6	15.3	7.3	8.6
Mode	0.0	0.0	0.0	0.0	0.0	33.0	0.0	13.3	1.9	18.4	0.0	0.0
Range	12.5	5.8	23.5	8.7	5.8	27.9	11.8	12.3	18.2	16.1	14.5	16.1
Min	0.0	0.0	0.0	0.0	0.0	17.9	0.0	6.8	-6.5	9.3	0.0	0.0
Max	12.5	5.8	23.5	8.7	5.8	45.8	11.8	19.1	11.7	25.4	14.5	16.1
Kurtosis (x)	-1.2	-1.7	1.0	1.1	-1.9	-0.4	4.7	0.1	1.8	-0.8	-0.8	0.9
Skewness (x)	-0.3	0.2	-0.5	1.4	-0.1	0.1	2.3	0.2	0.4	0.4	-0.1	-0.9
Count (x)	138	138	138	138	138	138	138	138	138	138	138	138

Appendix D: Monte Carlo Simulation of Weekly Net Asset Value

Description: Taking into account different risk preferences of investors during the forecast analysis of weekly NAV, we have performed a *Monte Carlo Simulation* showing several possibilities of future development based on computational algorithms following a repeated random sampling. As this procedure exceeds the scope of the thesis, related methodology is not presented in detail.

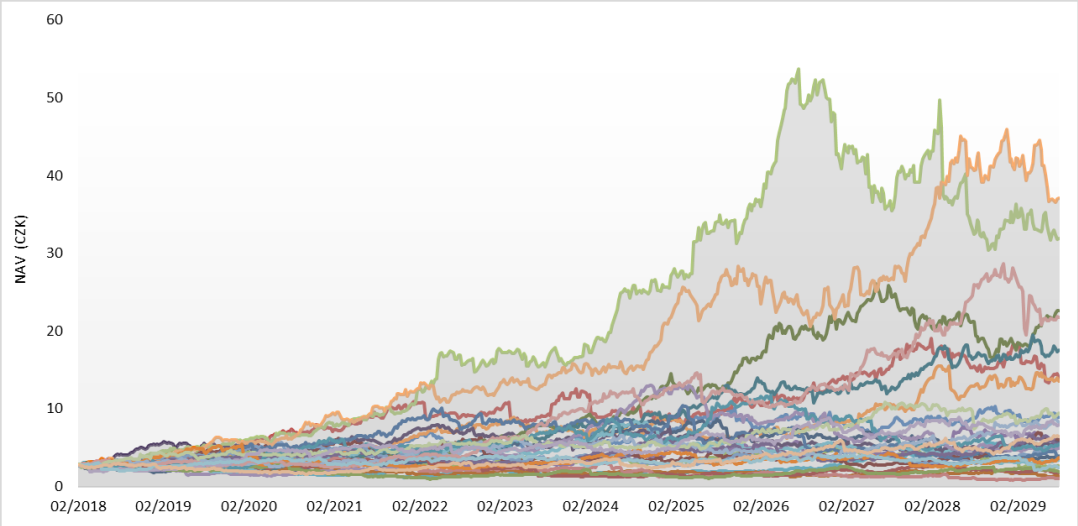
Please refer to the following literature for more details:

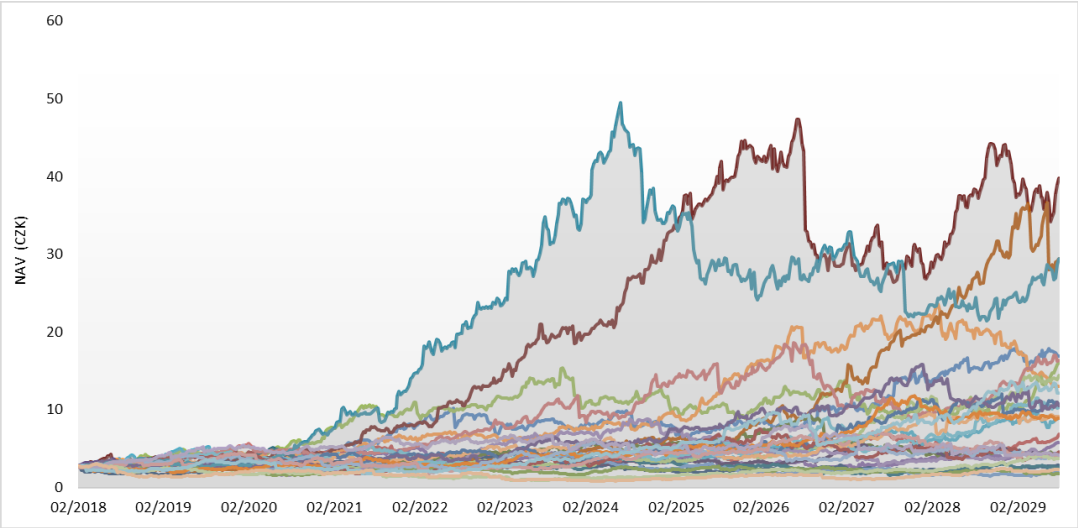
BOYLE, P., BROADIE, M. and GLASSERMAN, P. (1997). Monte Carlo Methods for Security Pricing. *Journal of Economic Dynamics and Control*, 21(8-9):1267–1321.

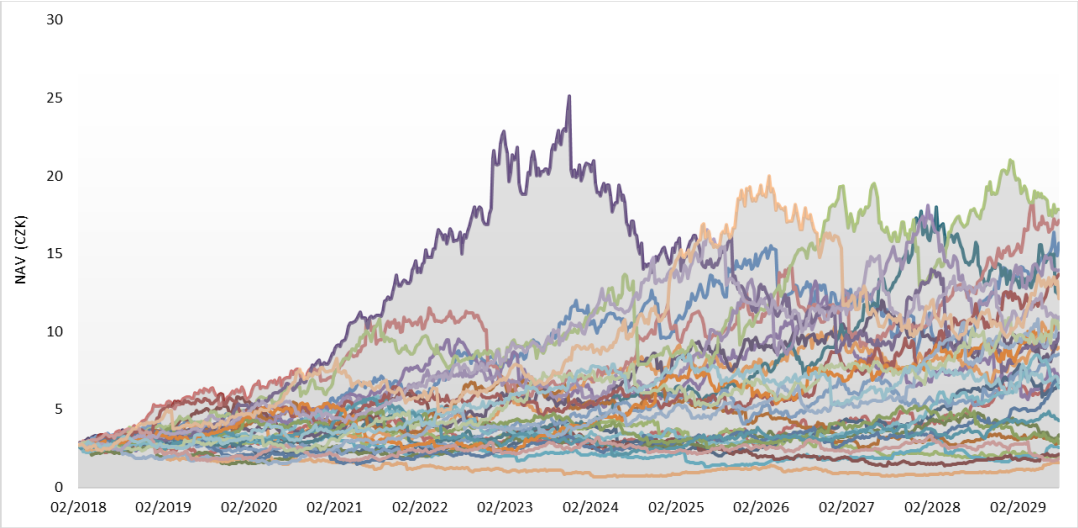
HARRISON, R. L. (2010). Introduction to Monte Carlo Simulation. *AIP Conference Proceedings*, 1204:17–21.

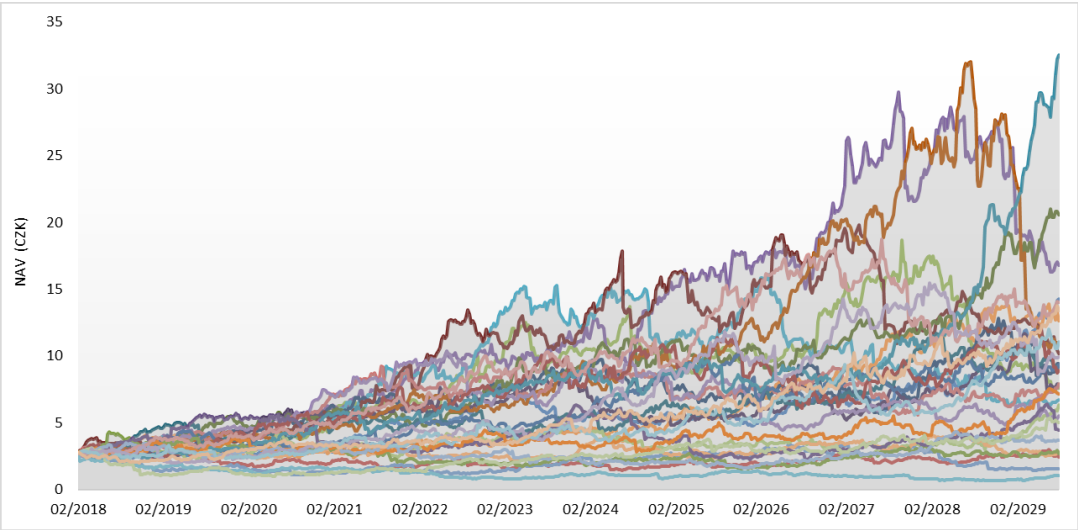
RAYCHAUDHURI, S. (2008). Introduction to Monte Carlo Simulation. *Winter Simulation Conference*, 91-100.

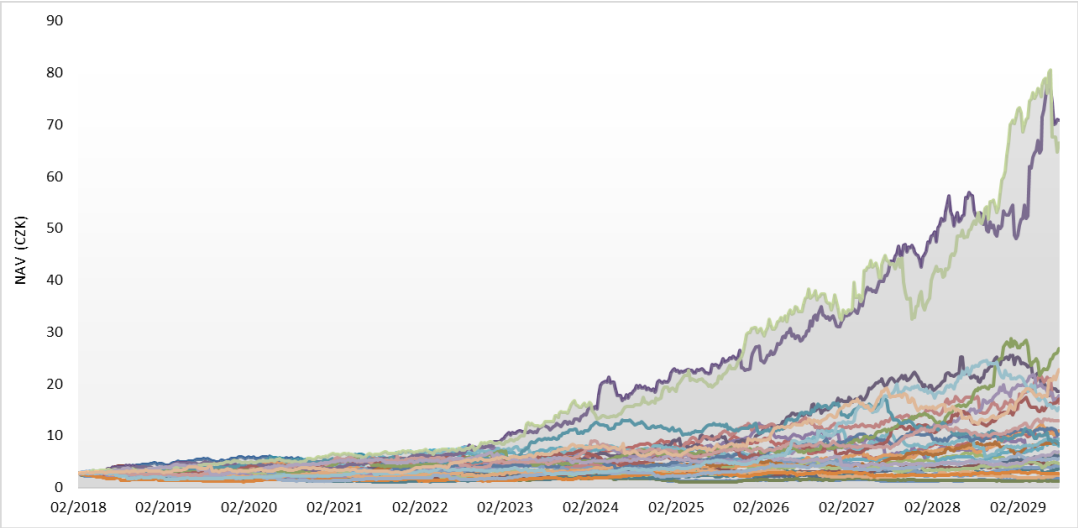
The contribution of this part of our research is that we provide a 'forecast' of the development of weekly NAV with respect to individual risk preferences, i.e. we allow readers to decide which one of simulated situations is the best according to their risk behaviour.

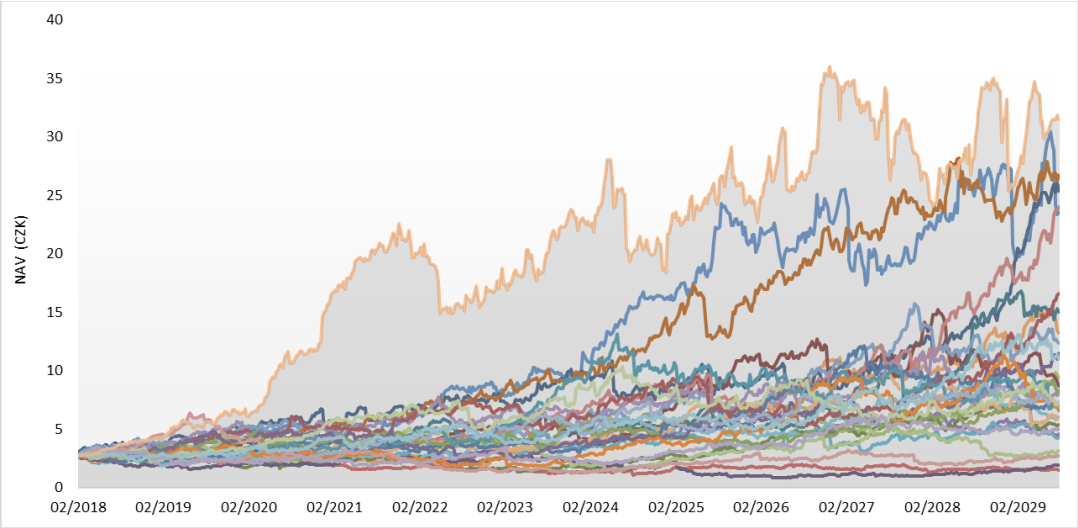


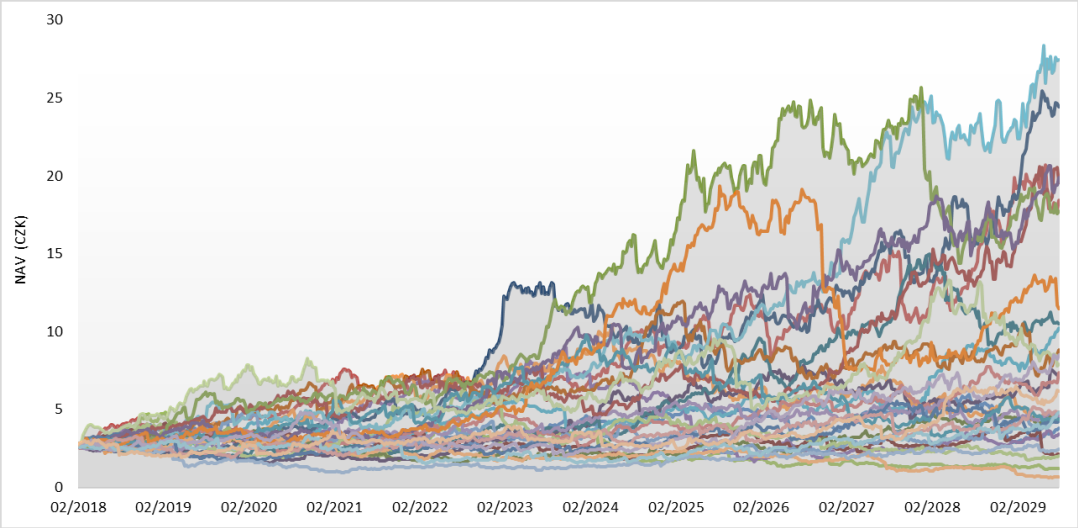

60
50
40
30
20
10
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029


30
25
20
15
10
5
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029


35
30
25
20
15
10
5
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029


90
80
70
60
50
40
30
20
10
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029


40
35
30
25
20
15
10
5
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029


30
25
20
15
10
5
0
NAV (CZK)
02/2018
02/2019
02/2020
02/2021
02/2022
02/2023
02/2024
02/2025
02/2026
02/2027
02/2028
02/2029