FEDERAL CENTER FOR TECHNOLOGICAL EDUCATION OF RIO DE JANEIRO

Investment Portfolio Diversification: the Minimum Correlated Portfolio Problem

Lawrence dos Santos Fernandes

Advisor:

Pedro Henrique González Silva, D.Sc.

Rio de Janeiro,

December, 2020

FEDERAL CENTER FOR TECHNOLOGICAL EDUCATION OF RIO DE JANEIRO

Investment Portfolio Diversification: the Minimum Correlated Portfolio Problem

Lawrence dos Santos Fernandes

Final paper in accordance with the standards of the Department of Higher Education of FEDERAL CENTER FOR TECHNOLOGICAL EDUCATION OF RIO DE JANEIRO (CEFET/RJ) as part of the requirements for a Bachelor's Degree in Computer Science.

Advisor:
Pedro Henrique González Silva, D.Sc.

Rio de Janeiro,
December, 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

F363 Fernandes, Lawrence dos Santos
Investment portfolio diversification: the minimum correlated portfolio problem / Lawrence dos Santos Fernandes — 2020.
x, 65f. + apêndice: il.(algumas color). ; enc.

Projeto Final (Graduação) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca , 2020.
Bibliografia : f. 55-65
Orientador: Pedro Henrique González Silva

1. Computação. 2. Investimento – Análise. 3. Programação linear. I. Silva, Pedro Henrique González (Orient.). II. Título.

CDD 004

Elaborada pela bibliotecária Tania Mello – CRB/7 nº 5507/04

I dedicate this work to my parents Antonio and Elisabete, to my fiancée Vanessa, and to my grandparents (in memoriam), for their love, support, and constant encouragement.

No, his mind is not for rent
To any god or government.
Always hopeful, yet discontent
He knows changes aren't permanent –
But change is

Neil Peart

ACKNOWLEDGMENT

I cannot begin to express my thanks to my advisor, Dr. Pedro Henrique, for guiding and supporting me through this journey. You have set an example of excellence as a researcher, mentor, and instructor.

I'm deeply indebted to Matheus Delgado for his invaluable contributions. The completion of this dissertation would not have been possible without your support.

I am also grateful to Dra. Julliany Salles Brandão and Victor Louvisse Lamanna, both responsible for valuable contributions to this research.

Special thanks to the members of my dissertation committee: Dr. Eduardo Ogasawara, who introduced me to the subject of this research, and Dr. Eduardo Bezerra for generously offering their time, guidance and good will throughout the preparation and review of this document.

Many thanks to Dr. Gustavo Paiva Guedes for his assistance during the initial phases of this research.

I cannot leave CEFET/RJ without thanking to all the amazing professors I've had and the colleagues I've made there over these years.

Finally, thanks should also go to the makers of Bialetti Moka Express, and the members of the StackOverflow and StackExchange communities for their invaluable assistance in many late at night coding sessions and theoretical elucidations.

ABSTRACT

Diversification is a highly explored risk management technique, which focus on reducing the risks associated with investment portfolios by mixing a wide variety of assets to compose them. Although extremely useful, this technique proves to be a daunt task, since extracting a subset from a large set of assets that presents the best result regarding the risk is a combinatorial optimization problem. That means this problem requires the use of specialized computational techniques to ensure a balance between the quality of the result, the execution time and memory expenditure involved in its processing.

This work presents a novel approach to investment portfolio diversification, named Minimum Correlated Portfolio Problem (MCPP), which seeks to reduce the risk of a diversified portfolio by minimizing the sum of correlations of the stocks that make it. This portfolio selection model is based on the Modern Portfolio Theory concept of reduction of correlation between assets, coupled with the 1/N rule, known as naive investing strategy. We propose an approach to solve MCPP based on Integer Programming, and show that this diversification strategy could improve the risk-adjusted returns of stock portfolios.

Keywords: Investment Diversification, Portfolio Optimization, Maximum diversity, Integer Programming, Correlation Matrix

Contents

1 Introduction **1**

2 Theoretical Framework **5**

2.1 Financial Assets and Instruments 5

2.2 Market Participants 8

2.3 Market Indexes 11

2.4 Portfolio Selection 14

2.4.1 The Mean-Variance Model 16

2.4.2 Diversification and Rebalancing 18

2.5 Portfolio Risk and Return 20

2.5.1 Standard Deviation 22

2.5.2 Alpha and Beta 24

2.5.3 R-squared 28

2.5.4 Sharpe and Information Ratios 30

2.6 Covariance and Correlation between stocks 32

2.7 Distance between stocks 35

3 Minimum Correlated Portfolio Problem **37**

3.1 Formal Definition 37

3.2 Mixed-Integer Nonlinear Programming Formulation 38

3.3 Linear Programming Formulation 39

4 Computational Experiments **42**

4.1 Instances 42

4.2 Computational Results 45

4.3 What-If Analysis 47

5 Conclusions **53**

Bibliographic References 54

Appendix A Figures **67**

Appendix B Tables **70**

List of Figures

FIGURE 1: Google trends for search term "day trading" in the United States, from 1/9/2019 to 11/9/2020 15

FIGURE 2: Security characteristic line, source: Wikipedia [b] 24

FIGURE 3: Top five rows from Apple stock price 44

FIGURE 4: Top five rows of the MCPP instance 44

FIGURE 5: Cumulative returns of MCPP portfolios vs. benchmark, full time period 49

FIGURE 6: Cumulative returns of MCPP portfolios vs. benchmark, 01/01/2018 to 01/05/2019 50

FIGURE 7: Cumulative returns of MCPP portfolios vs. benchmark, 23/01/2020 to 07/05/2020 50

FIGURE 8: Graphic of the stocks correlation matrix 68

FIGURE 9: Scatter plots of the cumulative returns of MCPP portfolios vs. benchmark 69

List of Tables

TABLE 1:	Financial instruments classification	8
TABLE 2:	Interpretations of correlation values	35
TABLE 3:	Linearization rules	39
TABLE 4:	Summarized results of the branch-and-bound exact method.	46
TABLE 5:	MCPP Portfolio Statistics	49
TABLE 6:	List of S&P 500 stocks used on the what-if analysis.	70
TABLE 7:	List of S&P 500 stocks removed from the what-if analysis.	85
TABLE 8:	List of S&P 500 stocks removed from MCPP portfolios	86

List of Abbreviations

CAPM	*Capital Asset Pricing Model*	2
CSV	*Comma-separated values*	43
CVM	*Comissão de Valores Mobiliários*	10
DJIA	*Dow Jones Industrial Average*	12
DJTA	*Dow Jones Transportation Average*	12
ETF	*Exchange-traded funds*	13
HDD	*Hard disk*	45
IAS	*International Accounting Standards*	7
IR	*Information Ratio*	31
MCPP	*Minimum Correlated Portfolio Problem*	3
MPT	*Modern Portfolio Theory*	2
NASDAQ	*National Association of Securities Dealers Automated Quotations*	10
NYSE	*New York Stock Exchange*	9
OLS	*Ordinary Least Squares*	28
ROI	*Return On Investment*	5
SCL	*Security Characteristic Line*	24
SCM	*Stock Correlation Matrix*	42
SEC	*U.S. Securities and Exchange Commission*	10
SIM	*Single-Index Model*	25

Chapter 1

Introduction

Diversification is a risk management technique widely used to reduce the risks associated with investment portfolios. This technique consists in allocating capital in a way that exposure to any specific asset or risks are mitigated. Broadly speaking, that means instead of investing in only one asset, investors would select different assets to compose their investment portfolios.

Possibly one of the most widely accepted principles in finance, diversification dates as far back as to the ancient times. It's credited to Rabbi Issac bar Aha in the 4^{th} century, the Talmudic passage that individuals should keep an equal proportion of their wealth on land (real state), merchandise (business), and cash (liquid holdings). The Talmud is the compendium of Jewish law which was compiled between the 3^{rd} and 5^{th} centuries [Dimitrovsky and Silberman, 2018].

Over the years, Aha's advice would lead to what is known naive strategies, namely: the 1/3 rule, which advises to split one's wealth equally between "user assets" (home, cars, boats), illiquid assets (rental properties, business holdings) and liquid assets (cash, stocks, bonds, retirement plans); and the 1/N rule or heuristic, also known as the uniform investment strategy or asset class diversification, which suggests to diversify uniformly between all available investment possibilities, since different assets (stocks, bonds, cash, real estate, commodities, and more recently cryptocurrencies [Guesmi et al., 2018]) are going to perform differently in various economic environments.

Although those strategies may seem very logical and are common sense between investors, they don't quantitatively address the selection of assets in order to compose a diversified investment portfolio. For example an investor could imagine to hold a diversified portfolio by following these allocation suggestions, just to discover later on a crisis period that his portfolio suffered a huge devaluation, due to the assets being perfectly positive correlated. Thus, if used alone, these rules could lead to the adoption of a bad diversification strategy which fails to mitigate the idiosyncratic (diversifiable) risk, and worse — keep investors with a false sense of security [Hicks, 2017].

Despite those diversification strategies that were known by the time, until the 1950s investors looked at securities individually for decision-making, given that the predominating

models were based on the idea of the existence of an optimal portfolio obtained through the maximization of portfolio returns, more often composed of only a single asset [Gartner, 2012]. It was Harry Markowitz's pioneering work [Markowitz, 1952] that pointed out this approach is high risk, because it does not take market risk into account.

The modern science of diversification is usually traced to Markowitz [1952], which led to the *Modern Portfolio Theory* (MPT) and was expanded upon in great detail in Markowitz [1991]. The mean-variance objective function (also known as mean-variance analysis, mean-variance framework, or mean-variance optimization), proposed by MPT, provides a method that can quantitatively determine the investment weight [1] of stocks to form an efficient stock portfolio, which maximizes the expected return for a given level of risk.

However, as pointed by Fletcher [2011], the major stumbling block of the implementation of the mean-variance framework is the estimation risk problem, where the true expected returns and covariance matrix are unknown and have to be estimated. As discussed in Choueifaty and Coignard [2008], much effort has gone into developing MPT, and while variance (risk) has proven to be estimated with a fair level of confidence, returns are so much more difficult to estimate. This evidence explains why most of the popular successor models of the mean-variance framework, such as the *Capital Asset Pricing Model* (CAPM) [Treynor, 1961, 1962; Sharpe, 1964; Lintner, 1965a,b; Mossin, 1966] [2] and *Black-Litterman* [Black and Litterman, 1992], have put the estimation of returns almost completely aside.

Perhaps one of the major contributions of the MPT, extensively addressed in Markowitz [1971], is the observation that investors can reduce risk of investment losses by reducing the correlation between the returns of the assets that compose their portfolios. This can be formulated as a combinatorial optimization problem, where the objective function is to minimize the sum of the correlation coefficients of the stock's returns that compose the portfolio.

Aiming to take advantage of this concept, some strategies have been proposed to provide a higher degree of diversification, for instance: Industry Diversification, where securities are selected from different industries in order to avoid industry specific risks; and Geographic Diversification, also known as International Diversification [Grubel, 1968; Levy and Sarnat, 1970], where the diversification is achieved by allocating capital on assets from foreign markets.

[1] Here weight is the same as percentage. Given some stocks to be analyzed, the output of the mean-variance framework is the individual percentages of each stock in order to obtain the Tangency Portfolio - the most efficient portfolio in terms of the risk-return relation.

[2] Prof. Craig W. French produced an interesting research about the origins of CAPM [French, 2003], revealing the importance of the works of Mr. Jack L. Treynor to the development of the model.

Regarding the benefits and weaknesses of those aforementioned techniques, in Ehling and Ramos [2006] it was found that unconstrained Geographic and Industry diversification are statistically equivalent. When introducing short-selling constrains, whilst Geographic diversification presents better results, the authors stated that the statistical evidence for a difference is weak. Another studies [Jorion, 1985; Bai and Green, 2010] have discussed whether international diversification is beneficial.

In spite of the 1/N rule seeming simplistic at first sight, after been compared against several portfolio optimization strategies, it was found it outperforms most of the other strategies [DeMiguel et al., 2009; Chan et al., 1999; Jagannathan and Ma, 2002]. In fact even Harry Markowitz, the father of MPT, managed his own funds with an application of the 1/N rule, by splitting his investments fifty-fifty between bonds and equities [Pflug et al., 2012].

For many years, Markowitz's principles for portfolio selection were poorly implemented, not because of disbelief as to their validity, but because of the excessive number of estimates needed to calculate portfolio risk [Freitas, 2008]. As the number of stocks to be evaluated grows, the computational resources and time required to obtain a solution becomes prohibitively, making an exhaustively search infeasible. In addition to this problem, the number of assets in a portfolio makes a direct impact on it's diversification.

As observed by Juan Zhan et al. [2015], the number of assets required to form a diversified portfolio has increased over the years from 8 [Evans and Archer, 1968] to 164 [Domain et al., 2007]. Chong and Phillips [2013] has noted that several studies converge to 30 [Fisher and Lorie, 1970; Statman, 1987] as the ideal number of stocks in a portfolio in order to obtain the benefits of diversification, whereas other works cite numbers ranging up to 300 [Statman, 2000]. After analyzing the scenario, Chong and Phillips [2013] concluded that the ideal number of stocks would be between 20 and 65 depending on the investor's objectives, while Benjelloun [2010] concluded that 40-50 stocks is all that is needed to achieve diversification in the US stock market.

Faced with the problematic involving the adoption of MPT's concept of asset's correlation reduction, exploring combinatorial optimization techniques to find a portfolio which maximizes the diversity of stocks is a natural path to follow. Therefore, this research defines the *Minimum Correlated Portfolio Problem* (MCPP), that is the portfolio which minimizes the sum of the correlation coefficient between the assets that compose it. We also propose a computational method to solve this problem.

In addition to this introduction (Chapter 1), this research is divided into four more sections. Chapter 2 presents theoretical concepts and a review of the literature used in this work. Chapter 3 presents the MCPP, as proposed by the authors, and the methods implemented in order to solve it. Chapter 4 discusses the instances that were developed and data pre-processing steps involved, presents the computational results, and concludes with a what-if analysis of the portfolios achieved through MCPP. Chapter 5 presents the general conclusions of the research.

Chapter 2

Theoretical Framework

The theoretical reference covers the description of the theoretical concepts related to the dataset used in this work and the concepts related to portfolio optimization practices. As a result, this chapter is divided into sections introducing these concepts and their respective definitions.

2.1 Financial Assets and Instruments

As defined by Bodie et al. [2008] there are two types of assets: real assets and financial assets (also known as investment assets). According to Bodie et al. [2008], real assets generate net income for the economy, and some examples are land, properties, machines, and knowledge as well. Financial assets, by the other hand, are rights over the income generated by real assets, defining the allocation of income or wealth of the investors who own them.

Financial assets can be grouped to form an asset class, which is a grouping of investments that exhibit similar characteristics and are subject to the same laws and regulations over how they are created, traded, modified and settled. Historically, the main asset classes have been Real Estate, Fixed Income (also known as Debt), Equities (Stocks), and Derivatives.

Real Estate is a real asset and a type of real property. Real property examples include land, buildings and other improvements, plus the rights of use and enjoyment of that land, its natural resources, and all its improvements.

There are two main strategies to investing in real state: buying and selling, or renting. When buying and selling, the investor believes in an appreciation of the property against its present value, but it may prove wrong, resulting in property depreciation and lost of capital. Furthermore, the *Return On Investment* (ROI) for real estate assets can be a tricky calculation, due to repair/maintenance expenses and methods of figuring leverage — the amount of money borrowed (with interest) to make the initial investment.

Renting properties generates income from the rent payed by the renter to the landlord. But it also comes with dangers, since the property remains vulnerable to depreciation due to external factors, such as an increase in violence on the neighborhood, which can result in lower demand

for tenants. Besides that, the costs associated with the property (taxes, maintenance, etc) are responsibility of the owner, and those can have a higher impact when the property is vacant (thus, not generating income). As stated by Bazin [2017], property purchased as an investment may generate money if rented, but will yield poor and laborious return.

Fixed income, also known as debt securities, are a type of financial assets. Contrary to the public belief, fixed income has this name due to the fact they carry a maturity date (a fixed date to expire), and not because they doesn't vary or pays a fixed interest - there exists pre-fixed and post-fixed debt instruments. Debt holders may receive interest payments (known as coupon) until the maturity date, at which point they receive a lump-sum payment from the issuer.

Debt securities may have different names according to their maturity, collateral and other characteristics, usually being categorized into short-term (maturity in less than one year) or long-term (maturity in more than one year). The Scottish historian and best-selling author Niall Ferguson traces the emergence of fixed income assets back to the 14th century in Italy, when the city-states of Tuscany — Florence, Pisa and Siena, started issuing debt securities to their citizens as a way of finance their armies and paying off the debts acquired in the incessant wars [Ferguson, 2008].

Although quite popular within conservative investors, fixed income investments still carry a variety of risks, namely: interest rate risk, liquidity risk, inflation risk, credit risk, among others. While interest rate risk is mitigated by holding a bond until maturity, and liquidity risk may be avoided by choosing liquid assets (such as Treasury bonds), inflation and credit risk are more difficult to deal with. Inflation risk concerns to the possibility that inflation will rise, thereby lowering the purchasing power of the bond holder, whereas credit risk concerns with the risk of default, which means that the issuer may be unable or unwilling to make further income and/or principal payments.

The short story of friends Sylvia and Mary, told by Max Gunther on his bestseller The Zurich Axioms [Max, 2010], illustrate what can happen. Sylvia, who invested only on fixed income instruments, saw her money's spending power eroded by the unexpected two-digit inflation of the 1970s. Mary, by the other hand, took risks by investing and speculating in stocks and gold, thus being able to get rich. Although this story may lacks evidence, it serves as a reminder that fixed income is not always as safe as promised.

According to Ferguson [2008], after the advent of banking and the birth of the bond market, the next step in the history of economics was the rise of the joint-stock, limited-liability cor-

poration: joint-stock because the company's capital was jointly owned by multiple investors; limited-liability because the liability of the investors was limited to the money they had used to buy a stake in the company, thus protecting them from losing all their wealth if the venture failed. This combination proved very successful along the history, enabling the realization of large and risky projects such as the European overseas exploration during the Age of Discovery, as well as the construction of railways, gas pipelines, refineries, among others.

Equities are a type of financial asset. An equity security (also known as stocks or shares) is a small slice of a company, trust or partnership, representing a share of interest in such entity. The most common form of equity interest is common stock, although preferred equity is also a form of capital stock. The holder of an equity is a shareholder, owning a share, or fractional part of the issuer.

According to Bazin [2017], long term equities are the only investment that falls into the investment category among the alternatives in the financial market, because as a shareholder an investor becomes eligible for the company earnings.

Derivatives are a type of financial asset which derive their value from the value and characteristics of one or more underlining entities such as an equity, index, or interest rate. They can be exchange-traded derivatives and over-the-counter (OTC) derivatives [Heckinger et al., 2014]. Some of the more common derivatives include forwards, futures, options, swaps, and variations of these such as synthetic collateralized debt obligations and credit default swaps.

The first exchange to offer derivatives was the Chicago Board of Trade in 1864, but the high possibility of speculation of this type of instrument caused them to be tightly controlled after the crash of the New York Stock Exchange in 1929, being deregulated only in the 1970s, with the emergence of floating exchange rates [Kishtainy, 2012].

Asset classes are made up of financial instruments which often behave similarly to one another in the marketplace. The *International Accounting Standards* (IAS) [1] 32 and 39 define a financial instrument as any contract that gives rise to a financial asset of one entity and a financial liability or equity instrument of another entity [IAS, 2015].

Financial instruments may be categorized by asset class depending on whether they are equity-based (reflecting ownership of the issuing entity) or debt-based (reflecting a loan the investor has made to the issuing entity). Foreign exchange instruments and transactions are

[1] IAS standards were the first international accounting standards issued by the International Accounting Standards Committee (IASC), an independent international standard-setting body based in London and formed in 1973. IAS was replaced in 2001 by International Financial Reporting Standards (IFRS).

neither debt-based nor equity-based and belong in their own category.

Table 1: Financial instruments classification

Asset class	Instrument type			
	Securities	**Other cash**	**Exchange-traded derivatives**	**OTC derivatives**
Debt (long term) > 1 year	Bonds Debentures	Loans	Bond futures Options on bond futures	Interest rate swaps Interest rate caps and floors Interest rate options Exotic derivatives
Debt (short term) $\leq$ 1 year	Bills	Deposits Certificates of deposit	Short-term interest rate futures	Foward rate agreements
Equity	Stock	N/A	Stock options Equitt futures	Stock options Exotic derivatives
Foreign exchange	N/A	Spot foreign exchange	Currency futures	Foreign exchange options Outright forwards Foreign exchange swaps Currency swaps

Some authors consider financial assets basically a piece of paper [Bodie et al., 2008] — which is correct, technically speaking. But other authors and investors, like the renowned billionaire Warren Buffet, make no fundamental distinction between real assets (for example a company, with it's buildings and machines) from investment instruments. According to Reamer and Downing [2016], this view implies owning a business and owning its shares are basically the same, with the authors going further and stating that investors must act as if he or she has bought the business, not just a piece of paper.

2.2 Market Participants

As defined in Bodie et al. [2008], market participants can be divided into three large groups:

1. Companies are money takers. They raise capital to finance their expansion projects, whether by purchasing equipment, facilities, or hiring personnel. The money generated from their assets are used to pay lenders and it's profits can be distributed with its sociates.

2. Families are physical savers. They buy securities issued by companies or governments that need to raise funds.

3. Governments are borrowers or lenders, depending on the relationship between tax revenue and government spending.

However, as noted in Bodie et al. [2008], corporations and governments do not usually sell their securities directly to individuals. For this purposes exists the financial intermediaries,

such as commercial banks, investment banks, mutual funds and pension funds. According to [Jakab and Kumhof, 2015], banks are not intermediaries but fundamentally money creation institutions, while the other institutions in the category of supposed intermediaries are simply investment funds.

These entities acts as the middleman between the parties involved in a financial transaction, bringing together those economic agents with surplus funds who want to lend (invest) to those with a shortage of funds who want to borrow [Finance Informer]. In doing this, financial intermediaries provide a number of benefits to the average consumer, including safety, liquidity, and economies of scale, alongside the benefits of maturity and risk transformation. They also provides benefits to the companies and governments, since they enable fundraising faster and more efficiently.

Specialist financial intermediaries are ostensibly enjoying a related (cost) advantage in offering financial services, which not only enables them to make profit, but also raises the overall efficiency of the economy. Their existence and services are explained by the "information problems" associated with financial markets [Jakab and Kumhof, 2015].

Financial intermediaries specializes in certain types of services, with Investment banks, mutual funds and stock exchanges focusing on stock markets. Investment banks are responsible for structuring equity interest transactions, usually in the form of Initial Public Offering (IPO) operations — the process of offering shares of a private corporation to the public in a new stock issuance.

Stock exchanges (also known as securities exchange or bourse), are facilities where stockbrokers and traders can buy and sell securities, such as shares of stock, bonds, and other financial instruments. Stock exchanges may also provide facilities for the issue and redemption of such securities and instruments and capital events including the payment of income and dividends.

Securities traded on a stock exchange include stock issued by listed companies, unit trusts, derivatives, pooled investment products and bonds. Stock exchanges often function as "continuous auction" markets with buyers and sellers consummating transactions via open outcry at a central location such as the floor of the exchange or by using an electronic trading platform [Lemke and Lins, 2013].

Most countries have only one stock exchange, e.g., Deutsche Börse in Germany, B3 in Brazil; whereas others have many, like India with two and the US who has a total of five exchanges, with the most prominent being the *New York Stock Exchange* (NYSE) and the *National*

Association of Securities Dealers Automated Quotations (NASDAQ) [Desjardins, 2016].

Another important group is the regulatory institutions. While not directly involved in the transactions, they are responsible for the drafting and ensuring compliance with the rules governing the market. Examples include the *U.S. Securities and Exchange Commission* (SEC) in the United States, *Comissão de Valores Mobiliários* (CVM) in Brazil, among others.

Although the classification presented in Bodie et al. [2008] is widely accepted, it is somewhat generalist, since it does not take into account the way individuals operate in the market. While some authors make no distinction between investing and speculating [Max, 2010], Benjamin Graham [2] was among the first to see a fundamental distinction between the two [Graham, 2009].

In *Security Analysis* [Graham and Dodd, 2008], whose first edition dates back to 1934, Benjamin Graham defines investing as follows: "An investment operation is one which, upon thorough analysis promises safety of principal and an adequate return. Operations not meeting these requirements are speculative."

Following this distinction between investment and speculation, Bazin [2017] states that in the Market there is a hierarchy between five characters — Manipulator, Speculator, Novice Speculator, Institutional Investor, and Individual Investor, as follows:

- The Manipulator is the one who commands the markets. For the size of his money, he can also assume the attitudes of Speculator and Investor, according to his own interests.
- The Speculator may be an Investor but not a Manipulator due to insufficient capital.
- The Novice Speculator is insignificant, due to his inexperience and lack of capital.
- The Institutional Investor (pension funds and insurers) is the one who guarantee the functioning of the markets.
- The Individual Investor does not manipulate or speculate, as a matter of principle. He stands on the sidelines of events, concerned only with the growth of their dividends.

Another group that can be added is that of market gurus, those who develop and sell methods, usually promising quick and easy success in the stock market, thus making a large profit

[2] Benjamin Graham (May 9, 1894 – September 21, 1976) was a British-born American economist, professor and investor. He is widely known as the "father of value investing", and wrote two of the founding texts in neoclassical investing: Security Analysis (1934) with David Dodd, and The Intelligent Investor (1949).

by the selling of books and newsletters — despite the lack of evidence about their methods. As defined by the Russian-American professional trader and teacher of traders Alexander Elder in his best seller *Trading for a Living* [Elder, 1993], market gurus can be classified into the following groups:

- Market cycle gurus: this group promise to predict the ups and downs of the market based on simple theories such as cycles, volumes, Elliott waves, among other methods. They are more prominent in the stock market. Examples: Edson Gould, Joseph Granville.
- Magic method gurus: a "method guru" erupts in the financial scene after developing a new analytical or trading method. They are more prominent in the derivatives market. Examples: Jake Berstein, Peter Steidlmayer.
- Dead gurus: as the name suggests, gurus that already died. Their books are reissued, courses based on their methods are launched, and the legend of their supposed success grows posthumously. Examples: Ralph Nelson Elliott, William Delbert Gann.

Although this work does not focus on any specific group, the authors believes that those who can benefit most are investors in general, both individuals as well as institutional ones, since the model proposed by the MCPP enables the creation of theoretical portfolios to be used as a risk benchmark. However, this thesis warns against three groups: manipulators, speculators and market gurus.

While little can be made against manipulators (this group must be handled by the regulatory institutions and law authorities), the authors believes that a scientific approach combined with empirical knowledge is the proper way to conduct both investments and research in the field, having a lot to offer against speculators and market gurus.

2.3 Market Indexes

A market index (also known as stock market index, investment index or simply index) is traditionally defined as a market-capitalization-weighted average of a specific and relatively static list of securities [Lo, 2016]. In other words, it is a section of the stock market, computed from the prices of selected stocks — typically a weighted average by the market capitalization (the total value of all a company's shares of stock) of the underlying companies. Market indexes

serves as a tool used by investors and financial managers to describe the market, and to compare the return on specific investments, acting as a numerical shorthand for market activity.

Lo [2016] provides a broadening definition of an index using a functional perspective, in which any portfolio strategy that satisfies the following three properties should be considered an index: "(1) it is completely transparent (meaning that every aspect of the index must be public information and verifiable by any interested third party); (2) it is investable (meaning that an investor should be able invest into the portfolio over a period of time and realize the return reported by the index); and (3) it is systematic (meaning it is entirely rules-based and not dependent on any discretion or human judgment)".

Stock market indices may be classified in many ways, but they are usually categorized geographically (e.g., Europe, Asia), by levels of industrialization or income (e.g., Developed Markets, Frontier Markets), or size of the companies (Small Caps, Blue Ships). The major groups are global, regional and national.

Global (also known as world) indexes includes stocks spanning from multiple regions and countries, such as the MSCI World or S&P Global 100. Regional indexes combines stocks from a specific region, such as the FTSE Developed Europe Index or FTSE Developed Asia Pacific Index. National indexes are composed entirely of stocks a given nation, such as the American S&P 500, the Japanese Nikkei 225, the British FTSE 100, the Indian NIFTY 50, and Bovespa in Brazil. There are also indexes based on exchange, such as the NASDAQ-100 or NYSE US 100, or groups of exchanges, such as the Euronext 100 or OMX Nordic 40.

The first market index was initially published on February 16, 1885 by Charles Dow — one of the founders of Dow Jones & Company (founded on November 1882) along with Edward Davis Jones and Charles Milford Bergstresser, the same firm that gave birth to The Wall Street Journal on July 8, 1889. The index was composed of the daily average of twelve stocks Charles had selected, originally consisting of two industrial companies and ten railroads [Kennon, 2019].

In 1896, Charles Dow realized that industrial companies were quickly becoming more important than railroads, so he renamed the index to Dow Jones Rail Average — in the 1970's, the name was updated to the *Dow Jones Transportation Average* (DJTA), to cover the introduction of air freight and other forms of transportation)[Kennon, 2019], and created a new index of stocks consisting of twelve companies called the *Dow Jones Industrial Average* (DJIA). In 1916 the number of companies in the index was increased to twenty, reaching its final number

of thirty companies in 1928 [Kennon, 2019].

Originally the DJIA index was calculated as a simple average of the stocks included in the index, therefore the percentage change in the index would be the percentage change in the average share price. However, over the years the calculation procedure has been changed to adjust the average whenever a share is split or pays a dividend of more than 10%, or when one company is replaced by another. When one of these facts occurs, the divisor used to calculate the average price is adjusted so that the fact does not affect the index [Bodie et al., 2008].

Although the daily performance of the DJIA plays an important part in the financial news, which can partly be explained by the longevity of the index, given that it is based on a small number of companies caution is required to ensure that it is representative of the market in general [Bodie et al., 2008]. Critics also argue against the practice of disregarding the overall size of a firm, focusing instead on a nominal value that can be modified through share repurchases or issuance [Kennon, 2019].

With the aforementioned problems in mind, other market indexes were developed, such as the S&P 500, introduced in 1926 as "Composite Index" and containing only 90 stocks, then expanded to cover 500 stocks in 1957. Since the index is composed by a large number of companies and is adjusted for overall market capitalization, it's a good representative of the US economy, thus being widely user by investors and professional money managers.

According to Bodie et al. [2008], the S&P 500 index return is equal to the rate of return that would be obtained by the investor who holds all the shares of the index portfolio, proportional to the market value of those shares, and disregarding dividend payments. Besides that, investors can easily acquire this market index, either through index funds or *exchange-traded funds* (ETF).

The S&P 500 index is one of the most used market indexes by investors in the US and in the literature. Furthermore, the US stock market is one of the most stable markets in the world, with few periods of downtime and a large historical database [Ferguson, 2008]. Besides that, Investment Funds that track the index have been recommended by renowned investors Warren Buffett and John C. Bogle, for investors with long time horizons [Martin, 2018; Tsang, 2019].

Therefore, given the aforementioned characteristics of the S&P 500 index, and the fact it complies with all the properties defined in Lo [2016], it'll be used throughout this dissertation as a benchmark, and the stocks selected by the proposed optimization model will be retrieved from this index.

2.4 Portfolio Selection

An investment portfolio [3] is a grouping of financial assets such as stocks, bonds, and cash equivalent or money market instruments, as well as their fund counterparts, including mutual, exchange-traded and closed funds, which belongs to an investor. Moreover, non publicly tradable securities like real estate, art, and private investments, can also compose a portfolio. A stock portfolio is a specific type of financial assets grouping, being composed mainly or entirely of stocks.

Portfolios are held directly by investors and/or managed by financial professionals and money managers, and exist due to the finite nature of money and for diversification purposes. With a certain amount of money destined for investment, the portfolio choice will consist of determining which assets will be acquired and which fraction of the total amount will be invested in each one.

Since the late nineteenth century, many economic theorists have presumed investors think and behave rationally when buying and selling stocks, thus modeling humans as agents who are consistently rational and narrowly self-interested, and who usually pursue their subjectively-defined ends optimally. This modelling of human action lead to the creation of the term *homo economicus*, employed for the first time by critics of John Stuart Mill's work on political economy, according to Persky [1995]. The "economic man" is seen as rational in the sense that well-being, as defined by the utility function, is optimized given perceived opportunities.

In reality however, as illustrated by the Economics Nobel Prize winner Robert Shiller, investors do not think or behave rationally [Shiller, 1999]. To the contrary, as discussed in Freitas [2008], investors are driven by greed and fear, speculating stocks between unrealistic highs and lows; being mislead by extremes of emotion, subjective thinking and the whims of the crowd, in a behavioral pattern known as the herd behavior — what the former chairman of the Federal Reserve, Alan Greenspan, called irrational exuberance [Shiller, 2000].

Since the subprime crises in 2008, with the recovery of the global economy, a plethora of investment advisory companies and specialized media vehicles have flourished worldwide, specially in the emerging countries. Nowadays, alongside investment indications, it's possible to get detailed explanations for every event that happens in the financial markets, usually free of charge. The reality however, as discussed by the best seller author Nassim Nicholas Taleb,

[3] A portfolio is a linear combination of assets whose sum of the weights x_i of each asset i, or holdings, is unitary, that is, $\sum x_i = 1$.

indicates that things always become obvious after the fact [Taleb, 2007].

Even worse is the world-wide day-trading boom on recent years. The billionaire investor Mark Cuban compared the current stock market to the dot-com bubble of the late 1990s, citing easy-to-use trading apps, commission-free stock trading, and a lack of other ways to gamble as the reasons behind it [Frankel, 2020]. This boom in trading activity has only skyrocketed during the COVID-19 pandemics — as shown by the growing interest on "day trading" over the first half of 2020; helping push markets to record highs, which lead to speculations whether the next great bubble is inflating in the markets [Fuscaldo, 2020].

Figure 1: Google trends for search term "day trading" in the United States, from 1/9/2019 to 11/9/2020

While there is a great debate among financial professionals whether this increase in the trading activity is increasing or reducing volatility in the markets [Sardana, 2020], empirical evidence suggests that living from trading is almost an impossible mission.

A study commissioned by CVM present strong evidence that it does not make sense, at least economically, to try to live on day-trading: from 19,696 people who started trading between 2013 and 2015 in Brazil; 18,138 (92.1%) withdrew, and of the 1,558 people who persisted for more than 300 trading sessions, 91% had losses and only 13 traders obtained an average daily profit above R$ 300.00 [Chague and Giovannetti, 2020]. This means only 1% of people who engaged with day trading actually made profits from it.

As stated by Chague and Giovannetti [2020], in a financial market occupied by large institutions with high-tech tools, it does not seem reasonable to assume that a person, using the computer in his home, will be able to exercise the activity of day trader in a consistently profitable way. And this perception only reinforces the positions presented in previous sessions of this paper and adopted throughout this research, where we advocate for an empirical, data-driven methodology, with a long-term vision, not speculative, and aiming at diversification.

It is common sense that investors should construct their portfolios in accordance with their

risk tolerance and investing objectives, however asset allocation sits on the cohort of things that are easier on theory than in practice.

2.4.1 The Mean-Variance Model

The mean-variance analysis or model was proposed by Markowitz [1952] [4] [5] and expanded upon in great detail in Markowitz 1959, 1971, 1991. It is a mathematical framework for assembling a portfolio of assets such that the expected return is maximized for a given level of risk.

As presented by Steinbach [2001], from the very beginning Markowitz related his approach to the utility theory of von Neumann and Morgenstern, although he was well aware that "the Rational Man, like the unicorn, does not exist" [Markowitz, 1959, p. 206]. The justification provided for taking these assumptions is that "the 'fun of the game' can be ignored in deciding on a rationale for the selection of a portfolio, especially when this involves the allocation of large amounts of other people's money" [Markowitz, 1959, p. 226].

This implies the model assumes that investors are risk averse, so given a number of portfolios that offer the same expected return, they will prefer the less risky one. Besides that, an investor will take on increased risk only if compensated by higher expected returns. Conversely, an investor who wants higher expected returns must accept more risk.

The exact trade-off will not be the same for all investors, being dependent upon individual risk aversion characteristics. The implication is that a rational investor will not invest in a portfolio if there exists another one with a more favorable risk-expected return profile – i.e., if for that level of risk an alternative portfolio exists that provides better expected returns.

There are various formulations of the mean-variance, and in it's simplest form, it is defined as a single-period mean-variance analysis. Under the model, the reward of a portfolio is the mean of its returns – which is a proportion-weighted combination of the constituent asset's returns, as follows:

$$E(R_p) = \sum_i w_i E(R_i) \tag{2.1}$$

[4] With this work, Harry M. Markowitz was awarded the Nobel Prize in Economics in 1990 "for pioneering work in the economic theory of finance" [Markowitz, 2020], alongside professors William F. Sharpe and Merton H. Miller.

[5] As presented in Chiu and Wong [2014], the great contribution of Markowitz often sees him called the father of modern portfolio theory, however Markowitz [1999] wrote that: "I am often called the father of modern portfolio theory (MPT), but Roy [1952] can claim an equal share of this honor."

where R_p is the return on the portfolio, R_i is the return on asset i and w_i is the weighting of component asset i; and the risk of a portfolio is the variance of the return, as follows:

$$\sigma_p^2 = \sum_i \sum_j w_i w_j \sigma_i \sigma_j \rho_{ij} \tag{2.2}$$

where σ is the (sample) standard deviation of the periodic returns on an asset, and ρ_{ij} is the correlation coefficient between the returns on assets i and j.

The mean-variance model can be solved by means of an optimization problem, either a maximization or minimization one – according to the objectives seek by the investor. As defined in Steinbach [2001], maximizing the expectation of a concave quadratic utility function leads to a formulation like:

$$\max_x \quad \mu\rho(x) - \frac{1}{2}R(x) \tag{2.3}$$

$$\text{s.t.} \quad e^* x = 1 \tag{2.4}$$

where $e \in Rn$ denotes the vector of all ones. The objective models the actual goal of the investor, a trade-off between risk and reward, while the budget equation $e^* x = 1$ simply specifies her initial wealth w_0. Another formulation provided by the same authors, and preferred by them, is to minimize the risk subject to the budget equation and subject to the condition that a certain desired reward ρ can be obtained, which leads to a formulation like:

$$\min_x \quad \frac{1}{2}R(x) \tag{2.5}$$

$$\text{s.t.} \quad e^* x = 1 \tag{2.6}$$

where the investor's goal is split between objective and reward condition. A more detailed formulation of the minimization problem is provided by Freitas [2008], in which the portfolio risk V, for a portfolio return R_d, is given by the following quadratic problem:

$$\min \quad V = \sum_{i=1}^{M} {X_i}^2 v_i + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} X_i X_j \gamma_{ij} \tag{2.7}$$

$$\text{s.t.} \quad \sum_{i=1}^{M} X_i \bar{r}_i = R_d \tag{2.8}$$

$$\sum_{i=1}^{M} X_i = 1 \tag{2.9}$$

$$X_i \geq 0, \quad i = 1, ..., M. \tag{2.10}$$

where M is the size of the portfolio (number of stocks that compose it), $\bar{r}_i$ is the forecast of expected returns for stock i, γ_{ij} is the covariance of the returns of the pair of shares i and j, and v_i is the risk of stock i defined as the variance of its series of returns. Eq. 2.7 is the objective function to be minimized, the risk of the medium-variance portfolio, V; Eq. 2.8 is the constraint that guarantees the desired portfolio return, R_d; Eq.2.9 is the constraint which guarantees the integrity of the portfolio with the total allocation of available resources; and the Eq.2.10 restricts the model to positive participation only.

As demonstrated, the mean-variance model requires forecasts of expected returns, variances and covariance matrix that calculates the covariance between each possible pair of securities within the portfolio based on historical data or through scenario analysis. For n securities, that would require n estimates of expected returns, n estimates of their variances, and a covariance matrix that consisted of $(n^2 - n)/2$ estimates of covariances. The calculations increase rapidly as n increases – the full correlation matrix for the S&P500 requires 124,750 unique, but not independent, forecasts [Rea and Rea, 2014].

Rea and Rea [2014] suggests the problem with trying to use forecasts is that it is notoriously difficult to generate forecasts which yield a consistent correlation matrix without the use of additional specialized software to ensure the forecasts are consistent.

2.4.2 Diversification and Rebalancing

Diversification is a familiar term to investors. In fact, most of them are acquainted with the popular saying: 'Don't put all your eggs in one bucket'. In essence, this is what diversification is all about. It's a strategy used to reduce the total risk of investment portfolios, by not allocating one's wealth into a single asset or class of assets.

Furthermore, diversification is more than just a prescription for holding a large number

of assets. Diversification is also concerned with the relative risk contribution of those assets, which implies the possibility of a sudden loss, and with the relationships between the assets (i.e., their correlation), which determines how the assets will behave in relation to one-another under different market conditions.

Accordingly to von Mises [1949] human action implies a constant search to replace a less satisfactory state of affairs with a more satisfactory one, thus portfolios have an intrinsically dynamic nature. This implies investors are always evaluating their portfolio's composition, which in turn can be the result of intuitive weighting, casual observation of tendencies or deliberate planning.

From time to time, investors may want to realigning the weightings of their portfolio's assets, a practice known as rebalancing. Rebalancing involves periodically buying or selling assets in a portfolio to maintain an original desired level of asset allocation, say for example an asset allocation policy of 30/70 between bonds and stocks.

In this research we propose a computational model for obtaining the minimum correlated portfolio. So given a certain number of stocks to be analyzed and the desired size of the portfolio to be formulated, the output will be the most diverse stocks in terms of correlation. The number of stocks in this portfolio depends entirely on an investors desire, while also considering the 1/N rule (i.e., the portfolio is equally distributed between the stocks that compose it, without individual balancing), in contrast to what the mean-variance framework proposes.

This decision was made due to the fact the 1/N rule proved to outperforms most of other strategies [DeMiguel et al., 2009; Chan et al., 1999; Jagannathan and Ma, 2002], and because even Harry Markowitz, the father of the mean-variance framework, managed his own funds with an application of the 1/N rule [Pflug et al., 2012].

The model proposed on this research doesn't take into account the payment of dividends, due to the difficulties in obtaining this type of data and the complexity it would add to the what-if analysis. However, the authors do believe that investing in stocks aiming at the payment of dividends, alongside the appreciation of share value, is a strategy that provides the greatest potential of long term success and profitability.

Finally, It's important to note that the focus of this research is on the computational model proposed for solving MCPP and thus, we don't recommend using this approach for investing without further empirical evidence on it's efficacy.

2.5 Portfolio Risk and Return

The following sentence is attributed to Benjamin Graham:

"The essence of investment management is the management of risks, not the management of returns. Well-managed portfolios start with this precept."

Risk is usually defined as the possibility that an outcome will not be as expected, implying in the loss of something of value. According to Brooke [2010], Chicago's economist Frank H. Knight held a distinction between risk and uncertainty, where risk refers to outcomes that can be insured against, and uncertainty to outcomes that cannot be insured against. Risk refers to situations in which the outcome of an event is unknown, but the decision maker knows the range of possible outcomes and the probabilities of each, such that anyone with the same information and beliefs would make the same prediction.

Broadly speaking, investors are exposed to two types of risk, systematic and unsystematic. Systematic risk, also known as volatility, market risk, aggregate risk, undiversifiable risk, or simply β, is the risk inherent to the entire market or market segment, and it can affect a large number of assets, not just a particular stock or industry.

Systematic risk is both unpredictable and impossible to avoid completely, being known to affect the market prices of traded financial assets [Gatfaoui, 2007]. It occurs due to macroeconomic factors and cannot be eliminated by diversification, corresponding for most of the risk in a well-diversified portfolio, but it's rewarded — in the long run, investors expect compensation for bearing risk that they cannot diversify away [Maginn et al., 2007].

Unsystematic risk, also known as non-systematic risk, specific risk, diversifiable risk or residual risk, is the type of risk that refers to the uncertainty inherent to a specific company or industry. Examples include a change in management, a product recall, a natural disaster such as a flood, or a regulatory change resulting in slumping sales. It is unlikely that events such as the ones listed above would happen in every firm at the same time, therefore, diversification can greatly reduce unsystematic risk from a portfolio. There is no reward for taking on unneeded unsystematic risk.

One of the most well known phrases in the business world, attributed to Peter Drucker [6], states that "if you can't measure it, you can't improve it". Therefore, in order to do risk

[6] Peter F. Drucker (November 19, 1909 – November 11, 2005) was an Austrian-born American management consultant, educator, and author, whose writings contributed to the philosophical and practical foundations of the modern business corporation. He is revered as the father of modern management.

management, first we need to do risk measurement. According to Holton [2003], in the context of risk measurement, we distinguish between risk measure and risk metric, as follows:

- Risk measure: is the operation that assigns a value to a risk.
- Risk metric: is the attribute of risk that is being measured.

Volatility, credit exposure, delta, beta and duration are examples of risk metrics, while any procedure for calculating these is a risk measure [Holton, 2003]. For example, there are different ways that delta of a portfolio can be calculated: each represents a different risk measure for the same risk metric [Holton, 2003].

According to Holton [2004], risk has two components: exposure and uncertainty. When holding a certain asset, an investor is exposed to the risks inherent to this type of asset (and its market), and uncertain whether this investment will pay off – thus, facing risk. Holton [2003] classifies risk metrics according to what measure they quantify, as follows:

- Exposure: examples of risk metrics that only quantifies exposure are delta, duration and credit exposure.
- Uncertainty: examples of risk metrics that only quantifies uncertainty are probability and standard deviation.
- Uncertainty combined with exposure: examples includes variance and expected credit loss.

Bourgi [2019] classifies risk metrics into two groups: absolute or relative. Absolute risk metrics measure the risk of financial assets in absolute terms, i.e., not in relation to other assets or market returns – examples: Portfolio Standard Deviation, Value-at-Risk (VAR), and Value-at-Risk (VAR). Relative risk metrics measure volatility and the comparable risk of potential investments relative to the broader market – examples: Beta, Sharpe Ratio, Information Ratio, Treynor ratio, and Tracking Error.

Bodie et al. [2008] noted that both causal observation and studies shows that, for investors, risk is as important as expected return. Bodie et al. [2008] also states that when investors thinks about risks, they are interested in the probability of deviation from expected return. This is due to the probabilistic nature of risk, where an investor have the possibility of losing a portion (or even all) of a potential investment.

Risk and return are inextricably intertwined, and therefore, risk is inherent in all financial instruments. Every investment product involves different risks and returns. The risk–return spectrum, also called the risk–return trade-off or risk–reward, is the relationship between the amount of risk undertaken in an investment and the amount of return obtained from that investment. Generally, the more return sought, the more risk that must be undertaken. It's important to note this work don't attempts to estimate returns, focusing only on portfolio risk reduction.

On the next subsections, we will be presenting the metrics to pricing an individual security or portfolio: standard deviation, alpha and beta (the two key coefficients in CAPM), R-squared, and Sharpe ratio.

Return is perhaps the most intuitive portfolio metric, since it's simply the gains or losses one brings in as a result of an investment. The portfolio return is the weighted average of the returns from it's assets. Assume a two-stock portfolio with returns R_A and R_B, and the weights of the two assets w_A and w_B — note that the sum of the weights of the assets in the portfolio should be 1. The returns from the portfolio is given by Equation 2.11, as shown bellow:

$$R_P = w_A R_A + w_B R_B \tag{2.11}$$

Considering our previous example, let's say stock A produced 20% returns and stock B produced 12% returns. As already noted, the weights of the two assets are 60% and 40% respectively. The portfolio returns will be:

$$R_P = 0.60 * 20\% + 0.40 * 12\% = 16.8\%$$

2.5.1 Standard Deviation

The risk of an asset is measured by it's standard deviation, a statistic that is used to quantify the amount of variation or dispersion of a set of values. It is calculated as the square root of variance by determining the variation between each data point relative to the mean, according to the equation bellow.

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}[r(s) - \overline{r}]^2} \tag{2.12}$$

Volatile stocks presents higher standard deviation than more stable stocks, like blue-chip. The downside of this measures, is it calculates all uncertainty as risk, even when it's in the

investor's favor — such as above average returns. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.

In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. For investment purposes, it can be calculated as the mean squared deviations from the investor's expected return, the arithmetic mean, $\bar{r}$ [Bodie et al., 2008]. This can be translated into the following equation:

$$\sigma^2 = \sum p(s)[r(s) - E(r)]^2 \tag{2.13}$$

The risk of a portfolio is measured using the standard deviation of the portfolio. However, the portfolio risk isn't simply an average of the risk of each asset that composes it, since two risky assets can make up a safe collection if their prices tend to move in opposite directions — when one asset presents a drop in price, the other can brings in gains, resulting in low overall risk. For this reason to calculate portfolio risk one must consider the covariance/correlation between the assets. Follows the steps required to calculate the risk of a portfolio:

1. Identify the weight of each asset, which equals asset value divided by portfolio value.
2. Find the standard deviation of the prices of each stock over the selected period.
3. Find the correlation coefficient or covariance between each stock pair.

The portfolio risk can be calculated as follows:

$$\sigma_{\mathrm{P}} = \sqrt{w^2{}_{\mathrm{A}}\sigma^2{}_{\mathrm{A}} + w^2{}_{\mathrm{B}}\sigma^2{}_{\mathrm{B}} + 2w_{\mathrm{A}}w_{\mathrm{B}}\rho_{\mathrm{A,B}}\sigma_{\mathrm{A}}\sigma_{\mathrm{B}}} \tag{2.14}$$

$$\sigma_{\mathrm{P}} = \sqrt{w^2{}_{\mathrm{A}}\sigma^2{}_{\mathrm{A}} + w^2{}_{\mathrm{B}}\sigma^2{}_{\mathrm{B}} + 2w_{\mathrm{A}}w_{\mathrm{B}}Cov_{\mathrm{A,B}}} \tag{2.15}$$

where A and B represents two stocks, and w represents the weight of each asset in the portfolio.

Equations 2.14 and 2.15 shows that the portfolio risk can be calculated using either covariance or correlation. Both variance and correlation, as well as their relationship, are presented in Section 2.6.

Let's calculate the risk of a two-stock portfolio, composed of \$6,000 of stock A and \$4,000 of stock B (the weights of the two assets are 60% and 40% respectively), assuming stock A

have a standard deviation of 10 and stock B of 16, and the correlation between the two assets is -1. The portfolio risk will be calculated as follows:

$$\sigma_P = \sqrt{(0.6^2 * 10^2 + 0.4^2 * 16^2 + 2 * (-1) * 0.6 * 0.4 * 10 * 16)} = 0.4$$

It's worthy noting that, as discussed in [Jan, 2019], standard deviation is a measure of total risk of portfolio including both diversifiable and non-diversifiable risks. Including more than one asset in a portfolio can reduce the diversifiable risk and hence lower the standard deviation.

2.5.2 Alpha and Beta

Alpha (represented by the greek letter α) and Beta (represented by the greek letter β) are two statistical metrics useful for investors to evaluate a security's (either individual stocks or portfolios) performance. Both are risk ratios used in MPT and help to determine the risk/reward profile of investment securities.

As presented in in Feibel [2003], we can obtain both statistics through a linear regression analysis between the returns of a given security and the returns of it's benchmark. The line of best fit, also known as *Security Characteristic Line* (SCL) and shown in Figure 2, represents a linear relation between the return pairs.

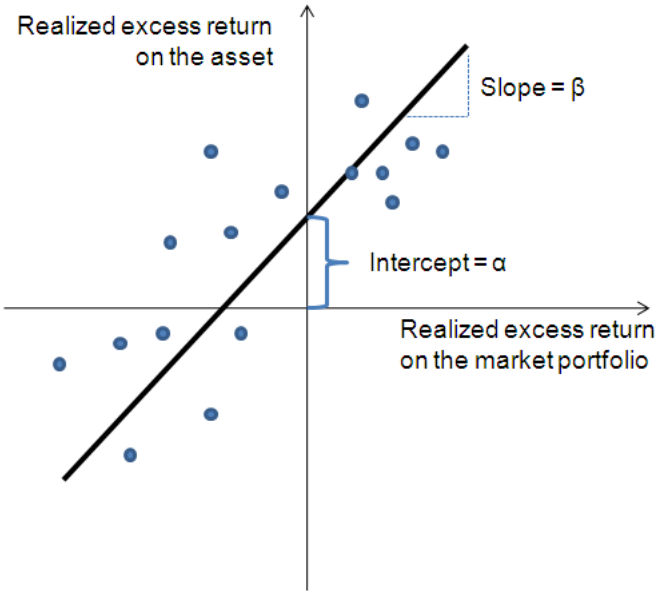

Figure 2: Security characteristic line, source: Wikipedia [b]

Alpha is a measure of the active return (sometimes refered as abnormal return) on an invest-

ment, the performance of that investment compared with a suitable market index. As discussed in Feibel [2003], practitioners use the term “alpha” in several different ways, but here we present the "standard alpha". In this context, alpha is the intercept of the SCL, that is, the coefficient of the constant in a market model regression.

The alpha coefficient is a parameter in the *Single-Index Model* (SIM). SIM is a simple asset pricing model to measure both the risk and the return of a stock, which was developed by Sharpe [1963] and is mathematically expressed as:

$$r_{it} - r_f = \alpha_i + \beta_i(r_{mt} - r_f) + \varepsilon_{it} \tag{2.16}$$

where

- r_{it} is return to stock i in period t.
- r_f is the risk-free rate (i.e. the interest rate on treasury bills).
- r_{mt} is the return to the market portfolio in period t.
- $r_{it} - r_f$ is the excess return on the stock.
- $r_{mt} - r_f$ is the excess return on the market.
- α_i is the stock’s alpha, or abnormal return.
- β_i is the stock’s beta, or responsiveness to the market return.
- $\beta_i(r_{mt} - r_f)$ is the non-systematic risk (aka non-market, diversifiable or idiosyncratic risk).
- ε_{it} are the residual (random) returns, which are assumed independent normally distributed with mean zero and standard deviation σ_i, which mathematically speaking translates to $\varepsilon_{it} \sim N(0, \sigma_i^2)$.

The alpha of a portfolio is the average of the alphas of the individual stocks. For a large portfolio the average will tend to zero, since some stocks will have positive alpha whereas others will have negative alpha. Hence, the alpha for a market index will be zero (considering an efficient market) [Spaulding, 2020].

Likewise, the average of firm-specific risk (aka residual risk) diminishes toward zero as the number of securities in the portfolio is increased. This, of course, is the result of diversification,

which can reduce firm-specific risk, but not market risk, to zero. Hence, the alpha component and the residual risk tends toward zero as the number of securities are increased, which reduces SIM equation to the market return multiplied by the risky portfolio's beta, which is what CAPM predicts [Spaulding, 2020].

Regarding it's numerical interpretations, the alpha coefficient indicates how an investment has performed after accounting for the risk it involved:

- $\alpha_i < 0$: the investment has earned too little for its risk (or, was too risky for the return).
- $\alpha_i = 0$: the investment has earned a return adequate for the risk taken.
- $\alpha_i > 0$: the investment has a return in excess of the reward for the assumed risk.

An alpha of positive 1 (+1.0) means the investment outperformed its benchmark index by 1%. An alpha of negative 1 (-1.0) means the investment underperformed its benchmark index by 1%. An alpha of zero means that the investment earned a return that matched the overall market return as reflected by the selected benchmark index.

The beta coefficient is a measure of the volatility of a security compared to the systematic risk of the entire market. Beta effectively describes the activity of a security's returns as it responds to swings in the market. According to Kenton [2020], in statistical terms beta represents the slope of the line through a regression of data points, and in finance terms, each of these data points represents an individual stock's returns against those of the market as a whole.

By definition, the market (in our case the S&P 500 Index), has a beta of 1.0, and securities are ranked according to how much they deviate from the market. It's important to note that, in order to beta provide useful insights, the market that is used as a benchmark should be related to the asset. As pointed by Kenton [2020], calculating a bond ETF's beta using the S&P 500 as the benchmark would not provide much helpful insight for an investor because bonds and stocks are too dissimilar.

An asset that swings more than the market over time has a beta above 1.0, while an asset that moves less than the market has a beta bellow 1.0. High-beta assets are supposed to be riskier but provide higher return potential, while low-beta assets pose less risk but also lower returns [Mcclure, 2020].

A security's beta is calculated by dividing the product of the covariance of the security's returns and the market's returns by the variance of the market's returns over a specified period. The calculation of beta for an individual stock is as follows:

$$\beta = \frac{cov(R_e, R_m)}{var(R_m)} \tag{2.17}$$

where the dividend is the covariance between R_e, which is the return on an individual stock, and R_m which is the return on the overall market; and the divisor is the variance of the returns on the overall market.

For a portfolio, beta is given by the straight weighted-average of individual beta coefficients. According to Jan [2019], portfolio beta equals the sum of products of individual investment weights and beta coefficient of those investments, whose calculation is given by the following equation:

$$\beta = W_A + \beta_A * W_B + \beta_B + ... + W_N * \beta_N \tag{2.18}$$

Let's calculate the beta of a two-stock portfolio. In respective order, their weights are 30% and 70%, their standard deviations are 2.5% and 3.5%, their betas are 0.9 and 1.1, and their mutual correlation coefficient is 0.5. The portfolio beta will be:

$$\beta = 0.9 * 30\% + 0.6 * 1.1\% = 0.93$$

Regarding the numerical interpretations of portfolio beta, we can break down the types of beta values into three valid groups:

- $\beta_i = 1$: indicates that price activity is strongly correlated with the market, and the asset mirrors the volatility of whatever index used to represent the market. It also indicates the portfolio has systematic risk and probably won't beat the index in terms of risk/return.

- $\beta_i > 1$: indicates that the portfolio is theoretically more volatile than the broader market. For example, if a portfolio beta is 1.2, it is assumed to be 20% more volatile than the market. This suggests holding such a portfolio could be riskier, but may also increase expected return.

- $\beta_i < 1$: indicates that the portfolio is theoretically less volatile than the broader market. This suggests holding such a portfolio could be less risky, but may also decrease expected return.

Negative beta is possible, but highly unlikely for stock portfolios, since it would indicate

an inverse relation to the market. Beta greater than 100 is deemed impossible, as it indicates volatility 100 times greater than the market.

For the most part, stocks of established companies rarely have a beta higher than 4 [Investopedia, 2020]. Regarding portfolios, as discussed by Micklitsch [2018], equity-oriented assets have betas close to +1.0, core fixed income has beta close to 0.0, alternative investments can have lower but still positive betas and outright portfolio hedges, such as S&P 500 puts (type of options), have negative betas.

In the CAPM, beta risk is the only kind of risk for which investors should receive an expected return higher than the risk-free rate of interest [Fama and Books, 1976]. When used within the context of the CAPM, beta becomes a measure of the appropriate expected rate of return, being useful in determining a security's short-term risk, and for analyzing volatility to arrive at equity costs [Kenton, 2020].

However, since beta is calculated using historical data and the volatility of stocks can change significantly over time, it becomes less meaningful for investors looking to predict the future movements of an asset — and therefore, should be used in conjunction with other metrics [Kenton, 2020].

2.5.3 R-squared

In statistics, the coefficient of determination, denoted R^2, r^2 or R-squared, is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is a statistic usually used in the testing of hypothesis, providing a measure of goodness of fit of the regression model [Neilands, 2016].

In investing, R-squared is generally interpreted as the percentage of a security's movements that can be explained by movements in a benchmark index. In order to make sure that a specific security is being compared to the right benchmark, it should have a high R-squared value, in relation to its benchmark. When using beta to determine the degree of systematic risk, a security with a high R-squared value could indicate a more relevant benchmark [Kenton, 2020].

The coefficient of determination R^2 is probably the most famous key figure when it comes to evaluate the fit of an *Ordinary Least Squares* (OLS) estimation [Blog, 2014a]. The purpose of the OLS method is estimating the unknown parameters in a linear regression model, explaining a certain dependent variable y through independent variables x.

The most general definition of the coefficient of determinations is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2.19}$$

where the dividend SS_{res} is the sum of squares of residuals, also called the residual sum of squares (variance explained by the model), and the divisor SS_{tot} is the total sum of squares proportional to the variance of the data (total variance) [Neilands, 2016].

In OLS multiple regression with an estimated intercept term, R^2 equals the square of the Pearson correlation coefficient estimate (defined in Equation 2.28) between the observed y and modeled (predicted) f data values of the dependent variable [Neilands, 2016].

In a OLS simple regression with an intercept term and a single explanator, this is also equal to the squared Pearson correlation coefficient of the dependent variable y and explanatory variable x [Neilands, 2016]. Therefore, the formula gets:

$$R^2 = r^2 = corr(x,y)^2 \tag{2.20}$$

A complete proof of how to derive R^2 from the Squared Pearson Correlation Coefficient between the observed values and the fitted values is provided by Blog [2014b].

R-squared is usually between the range 0 to 1, where:

- R^2 = 0: represents a model that does not explain any of the variation in the response variable around its mean. The mean of the dependent variable predicts the dependent variable as well as the regression model.
- R^2 = 1: represents a model that explains all of the variation in the response variable around its mean. It indicates that the regression predictions perfectly fit the data.

Values of R-squared outside the range 0 to 1 can occur. As observed by Motulsky [2018], its value is never greater than 1, but it can be negative when you fit the wrong model (or wrong constraints) so the SS_{res} is greater than SS_{tot}.

As pointed by Frost [2019], the larger the coefficient of determination, the better the regression model fits your observations. However, R-squared also has limitations, as it doesn't indicate if a regression model provides an adequate fit to your data: a good model can have a low R^2 value, while a biased model can have a high R^2 value [Frost, 2019].

2.5.4 Sharpe and Information Ratios

Ratios are used to make comparisons between two things, in situations where there are differences in the absolute measures being compared [Feibel, 2003]. In finance, the Sharpe ratio was introduce in a series of papers by Sharpe [1966, 1975, 1994], hence the name. It is also known as the Sharpe index (Radcliffe, 1997; Haugen, 2001), the Sharpe measure (Bodie et al., 2008; Elton et al., 2002), and the reward-to-variability ratio — the name that Prof. Sharpe originally intend to use [Bailey and Prado, 2012].

The Sharpe ratio is a measure for the performance of an investment (e.g., a security or portfolio) compared to a risk-free asset, after adjusting for its risk. It is defined as the difference between the returns of the investment and the risk-free return, divided by the standard deviation of the investment (i.e., its volatility). This ratio represents the additional amount of return that an investor receives per unit of increase in risk.

Since its revision by the original author in Sharpe [1994], the ex-ante Sharpe ratio is defined as:

$$S_a = \frac{R_a - R_f}{\sigma_a} \tag{2.21}$$

where R_a is the asset return, R_f is the risk-free return (such as a U.S. Treasury security), and σ_a is the standard deviation of the asset excess return. The dividend term is named *differential return*. In this version, the ratio indicates the expected differential return per unit of risk associated with the differential return [Sharpe, 1994].

The ex-post Sharpe ratio uses the same equation as the one above but with expected returns of the asset and benchmark rather than realized and risk-free returns, as follows:

$$S_a = \frac{E[R_a - R_b]}{\sqrt{\text{var}[R_a - R_b]}} \tag{2.22}$$

where $E[R_a - R_b]$ is the expected value of the excess of the asset return over the benchmark return. In this version, the ratio indicates the historic average differential return per unit of historic variability of the differential return [Sharpe, 1994].

Sharpe ratio is not the only portfolio performance measurement tool available, although most of the alternatives are classified as derivations or special cases, namely: Information ratio, Treynor ratio [Treynor, 1965], and Sortino ratio [Sortino and van der Meer, 1991]. Other ratios

aims to determine the abnormal return of an asset over the theoretical expected return, an example being the Jensen's alpha (also known as Jensen's Performance Index, or ex-post alpha) [Jensen, 1968].

The *Information Ratio* (IR) is often referred to as a variation or generalized version of the Sharpe ratio [Kidd, 2011], the main difference being that the Sharpe ratio uses a risk-free return as benchmark (such as a U.S. Treasury security) whereas the information ratio uses a risky index as benchmark (such as the S&P500).

There are several different methods for calculating IR, but the three more popular definitions are provided in [Blatt, 2004], with the most used one being defined as follows:

$$IR = \frac{E[R_p - R_b]}{\sigma} = \frac{\alpha}{\omega} = \frac{E[R_p - R_b]}{\sqrt{var[R_p - R_b]}} \tag{2.23}$$

where R_p is the portfolio return, R_b is the benchmark return, $\alpha = E[R_p - R_b]$ is the expected value of the active return (not the risk-adjusted excess return or Jensen's alpha), and $\omega = \sigma$ is the annualized standard deviation of excess returns, which is also known as the tracking error [Blatt, 2004].

IR measures the consistency of a portfolio manager's performance versus the benchmark. A higher, positive value indicates the manager has beaten the benchmark on a risk-adjusted basis, that is, without taking excessive risks. Although it is more useful to evaluate active management strategies, it provides useful information that is complementary to the Sharpe ratio (no pun intended).

The Treynor ratio is considered only useful if we assume that the portfolio manager has accounted for unsystematic risk (i.e., company-level risk) [Bourgi, 2019], which is not the case of this research. By the other hand, the Sortino ratio is usually considered a "sharper" ratio [Rollinger and Hoffman, 2015], especially when measuring and comparing the performance of managers whose programs exhibit skew in their return distributions. Recently, the (original) Sharpe ratio has often been challenged with regard to its appropriateness as a fund performance measure during evaluation periods of declining markets [Scholz, 2007].

However, despite its issues and limitations, the broader usage of Sharpe ratio by both the academia and industry in indisputable — perhaps due to its simplified calculations [Rollinger and Hoffman, 2015]. Therefore, the Sharpe ratio was preferably adopted in this research (alongside the information ratio), over the other alternatives mentioned.

2.6 Covariance and Correlation between stocks

For two jointly distributed random variables x and y, with expectations μ_x and μ_y (also known as the mean of x and y) and standard deviations σ_x and σ_y, respectively, then their covariance and correlation are as follows:

$$cov_{x,y} = \sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] \tag{2.24}$$

$$corr_{xy} = \rho_{xy} = \frac{E[(x-\mu_x)(y-\mu_y)]}{(\sigma_x \sigma_y)} \tag{2.25}$$

so that:

$$corr_{x,y} = \frac{cov_{x,y}}{(\sigma_x \sigma_y)} \tag{2.26}$$

where E is the expected value operator.

For multiple random variables, they can be stacked into a random vector whose i^{th} element is the i^{th} random variable. Then the variances and covariances can be placed in a covariance matrix, in which the (i, j) element is the covariance between the i^{th} random variable and the j^{th} one. Likewise, the correlations can be placed in a correlation matrix.

In probability theory and statistics, covariance is a measure of the joint variability of two random variables, or their degree of association [Rice, 2006]. The covariance can be used to determine the direction of a linear relationship between two variables as follows:

- If both variables tend to increase or decrease together, the coefficient is positive.
- If one variable tends to increase as the other decreases, the coefficient is negative.

Finding that two stocks have a high or low covariance might not be a useful metric on its own. Because covariance results are not standardized data, this statistic can only tell how the stocks move together, but it cannot be used to assess the strength of the relationship — for this purpose we need to look at the stock's correlation, which is defined in terms of covariance.

In Statistics, correlation is a statistical method used to assess a possible two-way linear association between two continuous variables [Altman, 1990]. Correlation is measured by a statistic called the correlation coefficient, which assess the strength of the relationship between the relative movements of the two variables. This statistical measure represents how closely two

variables co-vary, ranging from -1 (perfect negative correlation) through 0 (no correlation) to +1 (perfect positive correlation). There are two main types of correlation coefficients: Pearson's product moment correlation coefficient and Spearman's rank correlation coefficient.

According to Rodgers and Nicewander [1988], the Pearson product-moment correlation coefficient (or Pearson correlation coefficient, for short) was the first correlation measure, developed by Karl Pearson from a related idea introduced by Francis Galton in 1885 [Pearson, 1895], from whom Pearson was a disciple. This method measures the strength of a linear association (and the direction of this relation, whether positive or negative) between two variables of metric scale (interval or ratio / ratio).

Usually, the Pearson coefficient is obtained via a Least-Squares fit and a value of 1 represents a perfect positive relation-ship, -1 a perfect negative relationship, and 0 indicates the absence of a relationship between variables. Pearson's product moment correlation coefficient is denoted as ρ for a population parameter and as r (or r^2) for a sample statistic (estimate). Pearson coefficient is given by the following equation:

$$\rho_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \tag{2.27}$$

And the estimate,

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{2.28}$$

As observed in Mukaka [2012], Pearson's r is used when both variables being studied are normally distributed (normally distributed variables have a bell-shaped curve), and since it is affected by extreme values, this method is inappropriate when either or both variables are not normally distributed.

In fact, the usage of Pearson's r correlation implies other assumptions, such as linearity and homoscedasticity. Linearity assumes a straight line relationship between each of the two variables and homoscedasticity assumes that data is equally distributed about the regression line [Witte and Witte, 2016].

Related to the Pearson correlation coefficient, the Spearman rank correlation coefficient (rho), name after Charles Spearman who proposed the method in [Spearman, 1904] influenced by the work of Francis Galton, also measures the relationship between two variables [Lovie and Lovie, 1996]. Spearman's rho can be understood as non-parametric version of the conventional

Pearson correlation, and is given by the following equation:

$$\rho_{x,y} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2} \tag{2.29}$$

where d is the pairwise distances of the ranks of the variables x_i and y_i and n is the number of samples.

Compared to the Pearson correlation coefficient, the Spearman's correlation does not require continuous-level data (interval or ratio) and can be used for variables that are not normal-distributed and have a non-linear relationship, because it uses ranks instead of making assumptions about the distributions of the variables [Dodge, 2008]. However, it's important to note Spearman's correlation assumes that data must be at least ordinal and the scores on one variable must be monotonically related to the other variable (i.e., there is a function between the two ordered sets that preserves or reverses the given order) [Daniel, 1990].

There are many other types of correlation coefficients, including direct descendants of the Pearson correlation coefficient [Witte and Witte, 2016]. As pointed by Rodgers and Nicewander [1988], many "competing" correlation indexes are in fact special cases of Pearson's formula, such as Spearman's rho, the point-biserial correlation, and the phi coefficients, each computable as Pearson's r applied to special types of data [Henrysson, 1971].

While there are many theories in finance that imply monotonic patterns in expected returns, such as the Capital Asset Pricing Model (CAPM), the full set of implications of monotonicity is generally not exploited in empirical work [Patton and Timmermann, 2010]. Furthermore, the Pearson correlation coefficient remains the most widely used measure of relationship [Rodgers and Nicewander, 1988], being cited by Guo et al. [2018] as the dominant tool to measure the relationship between two stocks [Onnela et al., 2004].

Table 2 relates the possible values of correlation coefficient (valid for both Pearson's r and Spearman's rho) with it's interpretation regarding the strength of the linear relationship between the variables.

By providing the degree of association of two variables, correlation is widely used in finance to analyze the behavior between two or more stocks, as a tool for the mitigation of risk. It is in fact a major component of MPT, being used to include diversified assets that can help reduce the overall risk of a portfolio.

Stocks correlations can be calculated from the correlations of the returns of the stocks. To calculate stock returns, one can calculate the returns from daily price data, considering or not

Table 2: Interpretations of correlation values

Value of r	Strength of linear relationship
Exactly –1	Perfect downhill (negative)
(-0.99) — (-0.70)	Strong downhill (negative)
(-0.69) — (-0.40)	Moderate downhill (negative)
(-0.39) — (-0.20)	Weak downhill (negative)
(-0.19) — (-0.01)	Very weak downhill (negative)
0.00	No linear relationship
(+0.01) — (+0.19)	Very weak uphill (positive)
(+0.20) — (+0.39)	Weak uphill (positive)
(+0.40) — (+0.69)	Moderate uphill (positive)
(+0.70) — (+0.99)	Strong uphill (positive)
Exactly +1	Perfect uphill (positive)

the payment of dividends. To obtain the period returns it suffices to calculate the total return for each period and, in case of considering the dividends, treat them as being reinvested into the stocks that issued them. For purposes of simplification, in this research we only consider the daily appreciation of the prices of the stocks, ignoring the payments of dividends.

2.7 Distance between stocks

The MCPP is modeled as a graph (see Chapter 3 for a detailed definition), in which the correlation coefficient between the stocks behaves as a distance metric. According to Rosén [2006]; Keskin et al. [2011], for a distance metric d_{ij} to be valid, it must satisfy the following axioms of Euclidean distance:

(i) $d_{ij} \geq 0$

(ii) $d_{ij} = 0 \Longleftrightarrow i = j$

(iii) $d_{ij} = d_{ji}$

(iv) $d_{ij} \leq d_{ik} + d_{kj}$

Here, d_{ij} expresses the distance between each pair of stocks i and j. However, it is well known that the correlation coefficient does not satisfy all these axioms, as can be observed from Table 2 that correlations lie in the interval $[-1,1]$. Therefore, we need to convert the numerical values to a measure which can be construed to be a distance.

As discussed in Onnela et al. [2004]; Juan Zhan et al. [2015]; Guo et al. [2018], the conversion is motivated by considerations of ultrametricity (also known as ultrametric or ultra-metric), as defined in Mantegna, R. N. [1999] and expanded in Mantegna and Stanley [2000] [7]. Follows the equation for the conversion:

$$d_{ij} = \sqrt{2(1-\rho_{ij})} \tag{2.30}$$

where d_{ij} is the estimated distance and ρ_{ij} is the estimated correlation between stocks i and j. With the chosen transformation, the distance d_{ij} is mapped from $[-1,1]$ to $[2,0]$, with 0 accounting for totally correlated stocks and 2 for totally anti-correlated stocks.

In Equation 2.30, the term $1-\rho_{\text{ij}}$ is known as Pearson's distance, and it has been used in problems involving cluster analysis and data detection for communications [Schouhamer Immink and Weber, 2014]. The reason Equation 2.30 was prefered over Pearson's distance due to a larger usage in the financial literature.

[7] In this book, the authors apply concepts from statistical physics in the description of financial systems.

Chapter 3

Minimum Correlated Portfolio Problem

The Minimum Correlated Portfolio Problem (MCPP) deals with selecting a subset of stocks from some larger collection in such a way that the selected elements exhibit the greatest variety of characteristics. The variety between the stocks is given by their Pearson's correlation coefficient as defined in Chapter 2.

3.1 Formal Definition

Given a matrix M of type $3 \times N$, representing the correlation coefficient between a set S of stocks (e.g., the stocks that compose the S&P500 index), $G = (V, E)$ is the complete undirected graph representing the M matrix, V is the set of vertices (also known as nodes, i.e., stocks), E is the set of edges (with each edge representing the relationship between two stocks) and k the desired size of the stock portfolio to be formulated, the problem of obtaining the Minimum Correlated Portfolio Problem (MCPP) consists of finding the combination that minimizes the sum of the correlations between the assets.

Mathematically speaking the MCPP is defined as follows:

$$MCPP = \{S \subseteq V | \sum_{i \in S} \sum_{\substack{j \in S \\ j \leq i}} \rho(i,j) \leq \sum_{i \in S_2} \sum_{\substack{j \in S_2 \\ j \leq i}} \rho(i,j) \wedge S_2 \subseteq V \wedge |S| = |S_2| = k\} \tag{3.1}$$

In the Equation 3.1, $\rho(i,j)$ is the correlation coefficient between stocks i and j, also known as *edge cost* in the graph searching literature. The purpose of the MCPP is to find the subset S of stocks (i.g., the stock portfolio S) from a set V, composed of k stocks, which minimizes the *total cost* of the solution (the sum of the correlation coefficient between all the stocks that make up the portfolio). That is, the *total cost* of S is less than any other subset of V with the same modulus.

If all values of $\rho(i,j)$ are equal to each other ($\rho(i,j) = \rho(i,t) = \rho(j,t), \forall i,j,t \in V$) the problem resolution would be very trivial, once any subset of V with modulus k would be optimum.

However, in the scenario in which this work is based (real investment stocks) these quadratic correlations may be close to each other, but will probably never be all equal to each other nor equal to zero. Therefore, instances from this problem can only be converted into others for problems that consider the quadratic correction to be not null and distinct.

3.2 Mixed-Integer Nonlinear Programming Formulation

The MCPP presents several difficulties in its solving, but can be modeled as a *mixed-integer nonlinear programming* (MINLP), which uses a binary variable z_i that is equal to 1 when the stock i is present in S subset and is 0 otherwise. It can be defined as follows.

$$\begin{aligned} \min \quad & \sum_{i \in V} \sum_{\substack{j \in V \\ j \leq i}} \rho(i,j) z_i z_j \\ \text{s.t.} \quad & \sum_{i \in V} z_i = k \qquad (3.2) \\ & z \in \{0,1\}^{|V|} \qquad (3.3) \end{aligned}$$

Definition of the variables

$z_i = 1$ if stock i is in the selected portfolio, $z_i = 0$ otherwise.

Definition of the constraints

The size of the portfolio cannot be exceeded:

$$\sum_{i \in V} z_i = k$$

The variables are 0-1:

$$z \in \{0,1\}^{|V|}$$

Constraint 3.2 controls how many stocks can be selected to the subset S. In other hand, Constraint 3.3 defines the binary variable domain. The product of two binary variables is the same applying the junction of two boolean variables, what means both the stocks are present in the subset S.

While MINLP has proven to be a powerful tool for modeling, at the same time it brings several difficulties, since it combines algorithmic design challenges from combinatorial and nonlinear optimization [Sahinidis, 2019].

3.3 Linear Programming Formulation

Due to the difficulties presented by MINLP, an authors proposed Linear Programming Formulation is used instead. This formulation consists in the use of two linearization variables: the variable ρ_{ij} that has the same value as the returned function $\rho(i,j)$ which represents the correlation between the stocks i and j; and also a binary variable y_{ij} which represents the product of the binary variables z_i and z_j, hence it's equal to 1 when both variables are equal to 1, and is 0 otherwise. Mathematically:

$$\begin{aligned}
\min \quad & \sum_{i \in V} \sum_{\substack{j \in V \\ j \leq i}} \rho_{ij} y_{ij} & & \\
\text{s.t.} \quad & y_{ij} \leq z_i, & \forall i,j \in V \,|\, i<j & \quad (3.4)\\
& y_{ij} \leq z_j, & \forall i,j \in V \,|\, i<j & \quad (3.5)\\
& y_{ij} \geq z_i + z_j - 1, & \forall i,j \in V \,|\, i<j & \quad (3.6)\\
& \sum_{i \in V} z_i = k & & \quad (3.7)\\
& z_i \in \{0,1\}, & \forall i \in V & \quad (3.8)\\
& y_{ij} \in \{0,1\}, & \forall i,j \in V \,|\, i<j & \quad (3.9)
\end{aligned}$$

where Constraints 3.4, 3.5 and 3.6 ensures the binary behaviour of the product between the variables z_i and z_j - they follow the rules described in Table 3; the Constraint 3.7 is the same as 3.2. The Constraints 3.8 and 3.9 defines the binary variables domain.

Table 3: Linearization rules

z_i	z_j	y_{ij}
0	0	0
0	1	0
1	0	0
1	1	1

The rules used in the construction of Table 3 are simply the binary multiplication rules, as follows:

- 0 * 0 = 0
- 1 * 0 = 0 * 1 = 0
- 1 * 1 = 1

The following is an evaluation of the rules previously presented:

- $y_{ij} \leq z_i$

  Assuming $z_i = 0$ we have $y_{ij} \leq 0$. No matter the values of z_j the product will be zero, according to Table 3 and in compliance with the first rule..

  Assuming $z_i = 1$ results in $y_{ij} \leq 1$, according to Table 3 and in compliance with the first rule.

- $y_{ij} \leq z_j$

  Assuming $z_i = 0$, it is clear that y_{ij} will always be less than or equal to z_j, according to Table 3 and in compliance with the second rule.

  Assuming $z_i = 1$ results in $y_{ij} \leq z_j$. When $z_j = 0$ then $y_{ij} = 0$, and when $z_j = 1$ then $y_{ij} = 1$, according to Table 3 and in compliance with the second rule.

- $y_{ij} \geq z_i + z_j - 1$

  Assuming $z_i = 0$, $y_{ij} = 0$ we have $0 \geq 0 + z_j - 1$. This results in $z_j \leq 1$. Therefore, z_j assumes binary values belonging to its set $\{0,1\}$, according to Table 3 and in compliance with the third rule.

  Assuming $z_i = 1$, $y_{ij} \leq 1$ then $y_{ij} \geq 1 + z_j - 1$. This results in $y_{ij} \geq z_j$, according to Table 3 and in compliance with the third rule.

Theorem 1. *MCPP is NP-hard*

Proof. To prove that the Minimum Correlated Portfolio Problem is NP-hard by reducing the well known NP-hard heaviest k-subgraph problem (HSP) Brimberg et al. [2009]. Let $G = (V, E)$ an instance of the HSP, an instance $G' = (V', E')$ of the MCPP can be built as given bellow.

- For each vertex $v \in V$, create $v' \in V'$

- For each edge $e = (i, j) \in E$, create $e' = (i', j') \in E'$, such that $\rho(i', j') = w_{e'} = -w_e$
- For each $(i, j) \notin E$, create $e' = (i', j')$, such that $\rho(i', j') = w_{e'} = 0$

First suppose G has a subgraph with k nodes. It is easy to notice that G' also have a subgraph with k nodes. It's trivial to see that given the instance transformation whenever the k-subgraph of G has maximum weight, G' will have a k-subgraph of minimum weight such that for each vertex v in the k-subgraph of G, there is a vertex v' in the k-subgraph of G', such that $v = v'$. The reverse is analogous.

Thereby, once it's possible to convert an instance from MCPP into one of the HSP in polynomial time and vice-versa, we conclude the MCPP belongs to the NP-hard problem class as well. QED

Chapter 4

Computational Experiments

In this section the S&P500 instances creation is explained in great detail as well as the Maximum Diversity literature instances used. Then the computational results for all presented instances are show and discussed and finally a what if analysis is made comparing some of the algorithm found solutions against the index (benchmark) and other randomly generated portfolios over a range of time.

4.1 Instances

As outlined in the problem definition at Section 3.1, MCPP is modeled as a graph in which the correlation coefficient between the stocks behaves as a distance metric. Thus, an instance of the MCPP problem consists of a matrix of type $3 \times N$, in which the first two columns represents two stocks and the third represents the value of the correlation coefficient between these stocks.

According to the literature and as previously pointed out, for a distance metric to be valid, it must satisfy axioms of the Euclidean distance. Therefore, the conversion described on Section 2.7 was applied to the correlation coefficient between the stocks, to normalize the values from $[-1,1]$ to $[2,0]$ range, with 0 accounting for totally correlated stocks and 2 for totally anti-correlated stocks.

The size of the matrix is given by a k-combination of a set S of stocks given by the binomial coefficient, which can be written using factorials as

$$^{n}C_{k} = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4.1}$$

where k $\leq$ n, and k = 2 (i,e, 2-combination). Since the order of selection does not matter, unlike the square correlation matrix presented in Figure 8, the instance of MCPP is a lower triangular matrix (also known as left triangular matrix), having only the values bellow the determinant - which is also discarded, since the correlation between the same variable is 1.

The algorithm developed for the generation of the instance was named *Stock Correlation Matrix* (SCM). It has two modules developed in Python, one aimed at downloading historical

prices of stocks from Yahoo Finance in the format of *comma-separated values* (CSV) files, and the second module being responsible for the generation of the correlation coefficient matrix, calculated from the daily close price of the stocks. The second module of SCM is presented in a pseudo code form in Algorithm 4.1.

Although the correlation of returns are usually used, since the interest is typically in the returns on the portfolio [ldnquant, 2011], in our case we are actually interested in the absolute levels of correlation between the assets. Furthermore, as pointed by Libesa [2018], short-term changes are better interpreted from returns correlations, whilst valorations of long-term evolutions may be improved using prices. Therefore, we opted to use the correlation coefficient between stock prices, instead of stock returns.

Algorithm 4.1: Stock Correlation Matrix (SCM) algorithm

Data: Matrix of stock historical prices S

Result: Correlation coefficient matrix H

```
1  begin
2      prices ← ∅;
3      for stock in S do
4          p ← ClosePrice(stock);
5          prices ← prices ∪ {p};
6      end
7      for price in prices do
8          for i ← 1 in |prices| do
9              for j ← i + 1 in |prices| do
10                 H ← CorrelationCoefficient(stock_i, stock_j);
11             end
12         end
13     end
14 end
```

The program takes as input a matrix of the historical prices of all stocks. From lines 3 to 6, the algorithm performs the reading of the files and stores the closing price of the stocks. Beginning on line 7, the correlations between all the stocks of the current instance are calculated. It's important to note that the correlations are calculated using Equation 2.27 and converted to measures through the use of Equation 2.30.

The data used for the construction of the instance was downloaded from Yahoo Finance

using SCM and ranged from 03 January, 2017 to 30 June, 2017. The first module of SCM, responsible for the data acquisition, obtained a list of 506 companies listed in the S&P 500 index as of late August, 2017 [Wikipedia, a]. Figure 3 shows an example of a historical price file downloaded.

```
Date,Open,High,Low,Close,Adj Close,Volume
2017-01-03,170.779999,171.360001,169.309998,170.600006,170.341583,691300
2017-01-04,170.369995,173.169998,170.369995,172.0,171.739441,641700
2017-01-05,170.869995,173.059998,170.229996,171.880005,171.619629,861000
2017-01-06,171.320007,171.990005,169.300003,169.630005,169.37304699999999,828000
```

Figure 3: Top five rows from Apple stock price

The following stocks had to be removed from this dataset: BHF Brighthouse Financial Inc. (BHF) — since it was first listed on Nasdaq in 2017; QuintilesIMS (Q) and The Lennar Corporation (LEN-B) — due to unavailable data at the time. The reason behind the removal of those stocks was the fact that Pearson's coefficient assumes vectors must have the same size.

Regarding the remaining 503 stocks, it's important to note that none of them presented some day with missing quotation for the period covered. In case of missing values, it would be necessary to use some data imputation technique to allow the calculation of the correlations. One possibility would be to use the median, because as opposed to the average, it is not affected by extreme values [Rice, 2006]. Another possibility would be to use the quotation from a previous day.

One instance was generated by the SCM algorithm, composed of all the stocks from the 503 companies whose quotations were previously downloaded. The instance consists of a pairwise comparison of the correlation of prices of the selected stocks. The first row of each instance informs the number of stocks. From the second row on, the instance contains three columns: the first two columns represents the stocks and the third one contains the correlation coefficient of the prices of each stock. Figure 4 shows the top five rows of the instance.

```
503
0 1 1.1403194261838805
0 2 1.9600797238009329
0 3 0.5042599393827614
0 4 0.5407386318492542
```

Figure 4: Top five rows of the MCPP instance

To make a better study on the proposed algorithms behaviour for the most different scenar-

ios, we have also used some of the Maximum Diversity problem literature instances approached by Martí and Duarte [2010], that are addressed as follows. These instances are equivalent to those produced by the Algorithm 4.1; so they can naturally be used for the experiments.

One instance group type is the GKD: this data set consists of 145 matrices for which the values were calculated as the Euclidean distances from randomly generated points with coordinates in the 0 to 10 range. In detail, both the GKD-a instances introduced by Glover et al. [1998] and those generated by Martí et al. [2010] named GKD-b are composed by a number of coordinates for each point that was generated randomly in the 2 to 21 range. The GKD-c instance group was introduced by Duarte and Marti [2007] with 10 coordinates for each point.

Another group type is SOM: this data set consists of 70 matrices with random numbers between 0 and 9 generated from an integer uniform distribution. Specifically, the SOM-a instances were generated by Martí et al. [2010] with a generator developed by Silva et al. [2004]. The SOM-b instances were generated by Silva et al. [2004] and used in most of the previous papers (see for example Aringhieri et al. [2008]).

The last harder instance group type is MDG: this dataset consists of 100 matrices with real numbers randomly selected between 0 and 10 from a uniform distribution, which is very similar to the S&P500 instance. Duarte and Marti [2007] also generated the MDG-a instance group, lately used by Palubeckis [2007]. The MDG-b group is composed by 40 matrices also created by Duarte and Marti [2007], which were used in Gallego et al. [2009]; Palubeckis [2007]. Finally, the MDG-c dataset with 20 matrices are the biggest instances reported in the literature. These instances are particularly similar to those used by Palubeckis [2007].

4.2 Computational Results

In this section we present the computational result of the proposed algorithm on various groups of instances, both from literature and generated by the authors proposed Algorithm 4.1.

The Table 4 show the results of a single execution of the exact branch-and-bound algorithm against the aforementioned instances. The S&P500 instances requires a percentage of n as the number of stocks to be selected, a.k.a. K. Then, the program needs at maximum three parameters: the instance itself to be calculated, the number of stocks percentage to be chosen (optional) and finally the maximum time to run.

The test machine used was an Ubuntu 16.04 server with the following characteristics: 8 CPU cores, 32 GB of RAM and 2 TB of *hard disk* (HDD). The branch-and-bound algorithm imple-

ments the Linear Programming formulation, described in Section 3.3, and makes usage of the IBM CPLEX Optimization Studio library (https://www.ibm.com/products/ilog-cplex-optimization-studio), being developed in the C++ programming language. The linearization of the problem was required because CPLEX does not support the definition of MINLP problems.

CPLEX is the most widely used commercial optimization tool for solving Mixed-Integer Problems [Lodi et al., 2010]. It is able to solve very large problems using either Primal or Dual variants of the Simplex method, the Barrier interior point method, Convex and non-convex quadratic programming problems (Mixed Integer Quadratic Program - MIQP), and Convex quadratically constrained problems (Mixed Integer Quadratically Constrained Program - MIQCP) - solved via second-order cone programming (SOCP) [Mittelmann, 2010].

Table 4: Summarized results of the branch-and-bound exact method.

Group	Instances	Nodes	Avg Value	Avg Time	Optimal Proved
GKD-a	25	{10,15,30}	5635,8455	0,0465	100,00%
GKD-b	10	{25,50,100,125,150}	27268,8169	445,8562	92,00%
GKD-c	20	{500}	15633,3800	3600,0000	0,00%
SOM-a	10	{25,50,100,125,150}	855,1800	2079,9359	48,00%
SOM-b	4	{100,200,300,400,500}	18106,1000	3497,2016	5,00%
S&P500	1	{503}	960,6096	363,75626	100,00%
		Avg	11409,9886	1664,4660	49,00%

In Table 4, the first column contains the instance group name; it's second and third columns contains the number of instances and their number of nodes, respectively, is such a way the total number of instances in that group is obtained by multiplying the modulus of the set by the number in the second column.

Still on Table 4 we can observe the results of the exact branch-and-bound algorithm. The presented values are averages on these group of instances explained above. All program runs had the maximum running time set to 3600 seconds; all executions of the S&P500 instance were done with the percentage parameters in the set {30,50,75,90,100}, while instances from other groups were executed without this parameter.

On the Table 4 the algorithm found the optimum solutions for all instances in the groups GKD-a and S&P500, composed by instances with a maximum of 100 nodes. However, on groups with more than 100 nodes there was a clear decay on this measure. For example, on the groups GKD-b and SOM-a which contains the same number of instances and distribution on number of nodes, the method clearly failed in one, but not on the other. This was possibly due to

the internal structure of the SOM-a instances, which are mainly composed by natural numbers and various zeros correlation (Section 4.1). Nevertheless, on the same groups the method found the optimum solution 100% times.

On the instance groups GKD-b, GKD-c, SOM-a and SOM-b the optimum was not found 100% of times. The group of big instances GKD-c constains not only instances with 500 nodes, but also most of the correlation values are close to each other. Therefore, most of the solutions in the solution space may have a very close objective function value.

There was three groups with a high number of nodes instances which could not be executed by the branch-and-bound algorithm due to hardware constraints, namely: MDG-a, MDG-b, and MDG-c. The used branch-and-bound framework from IBM allows, for example, the expansion of the searching three on the HDD when there's not enough volatile memory for such operation. However, this same algorithm could not even had its mathematical model loaded in the volatile memory to start the searching tree expansion, and then it could not be executed.

Yet, even in a scenario where the algorithm could be executed for the MDG instance group type, the results would be probably very similar to those presented in the Table 4 for the GKD-c group of instances. This is due the fact all MDG instance groups have a number of nodes greater than or equal to the GKD-c's.

4.3 What-If Analysis

In this section we analysed what would have happened if the portfolios generated by MCCP had been selected in real life. The first step was to generate the MCPP instance through the SCM algorithm, as described on Section 4.1.

With the instance at hand, the branch-and-bound exact algorithm was executed using the following portfolio sizes: 10, 20, 30, 60, 100, and 200. These portfolio sizes were selected based on the discussion presented on Chapter 1. Each resulting portfolio was then respectively named P_010, P_020, P_030, P_060, P_100, and P_200.

Once an index is not an investment asset, the next step was to identify a portfolio to represent it. For representing S&P 500, we opted to use an index-tracking ETF, as it follows the performance of a given index, automatically adding or removing stocks, as well as rebalancing the portfolio as the composition of the index changes. This ETF will act as a benchmark, being compared against the portfolios generated by MCPP.

After a quick analysis on the available options, SPDR S&P 500 trust (NYSE: SPY) was

taken as the winner. It was selected for the following reasons: it was launched in 1993, making it the first ETF in the United States; it is the largest ETF in the world, making it a highly traded security (best for liquidity and volume). All those characteristics makes SPY a good candidate to represent the S&P500 index, although a similar ETF could be used [Chin, 2020].

The next step was to download the quotation for the stocks to be analysed. For this analysis, historical prices of all previously used stocks (present in the MCPP instance), were downloaded ranging from 03 January, 2017 to 03 October, 2020. Here is where things get tricky, since the index changes over time, as well as the companies itself: think of acquisitions, mergers, companies getting delisted (removed from the index or from the stock market i.e., in case of going private), among other financial events.

The reason why financial events such as mergers and acquisitions are so troublesome is due the fact that, most of the time, there is no clear successor for the deceased company(ies). Let's take for example the takeover of Scripps Networks Interactive Inc. by Discovery Communications: Scripps shareholders received about $90 a share — $65.82 a share in cash and 1.0584 a share in Series C common shares of Discovery stock [Discovery Communications]. Observe there is no direct conversion from stock X to Y, and a shareholder could also opt to liquidate his position before or after the completion of the deal.

The financial events mentioned, incurred in missing or incomplete quotation for some stocks [1]. For the sake of simplicity, we opted to remove these companies from the what-if analysis, and listed them on Table 7, thus remaining the 461 companies listed on Table 6. The portfolios P_060, P_100, and P_200 contained some of the stocks removed from the analysis, which are listed on Table 8, so they ended with 59, 94 and 187 stocks, respectively.

Without further ado, it was time to begin the what-if analysis itself. A Python script was developed for loading the portfolios, calculating and plotting both their cumulative returns and risk metrics. Regarding the risk metrics, the standard deviation, alpha, beta, R-squared, Sharpe and information ratios of each portfolio were calculated, according to the definitions provided on Section 2.5. Regarding the Sharpe ratio, for the risk-free asset we considered the 1 Month Treasury Rate as of November 2020, that is $0.09\% \approx 0$ [YCHARTS]. Therefore, we rounded it to zero.

Before discussing the results, we present our hypothesis, which were formulated a priori

[1] Wyndham Worldwide (NYSE: WYN) was renamed to Wyndham Destinations (NYSE: WYND) on June 1, 2018. Since only the name and ticker of the company changed, we could successfully retrieve the quotation for the period.

and based on the literature covered on Chapter 2.

Hypothesis 1. *The volatility of the portfolios will decreased according to their size, as the variances will be smoothed out by a large number of stocks.*

Hypothesis 2. *The portfolios with a lower number of stocks will outperform the larger ones in terms of returns, but not without an increase in risk.*

The portfolios metrics are depicted on Table 5. The comparisons between the portfolios generated by MCPP and the benchmark are shown on Figures 5, 6 and 7. Figure 9 shows a scatter plot of the cumulative returns of MCPP portfolios against the cumulative returns of the benchmark.

Table 5: MCPP Portfolio Statistics

Porfolio	Std Dev	Alpha	Beta	R-squared	Sharpe ratio	IR	Returns
Benchmark	0.2036	0.0	1.0	1.0	0.5439	–	11.08%
P_010	0.2196	0.00011	0.6983	0.4194	0.5280	-0.0035	9.61%
P_020	0.2341	0.00029	1.0754	0.8751	0.8959	0.0627	19.96%
P_030	0.2279	0.00031	1.0665	0.9080	0.9359	0.0781	20.57%
P_060	0.2197	0.00021	1.0349	0.9203	0.8377	0.0585	17.31%
P_100	0.2121	0.00020	1.0044	0.9298	0.8303	0.0562	16.58%
P_200	0.2134	0.00019	1.0173	0.9425	0.8204	0.0602	16.43%

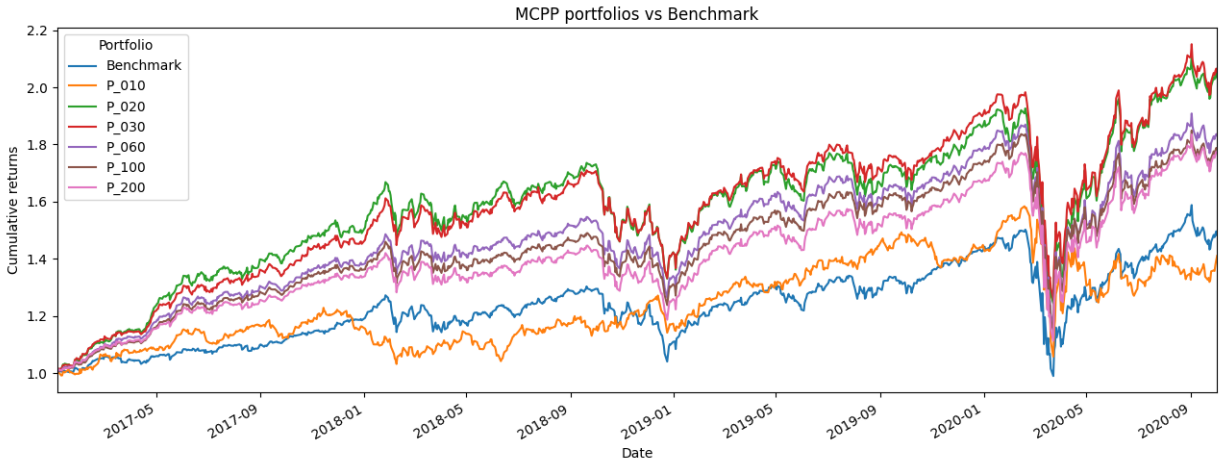

Figure 5: Cumulative returns of MCPP portfolios vs. benchmark, full time period

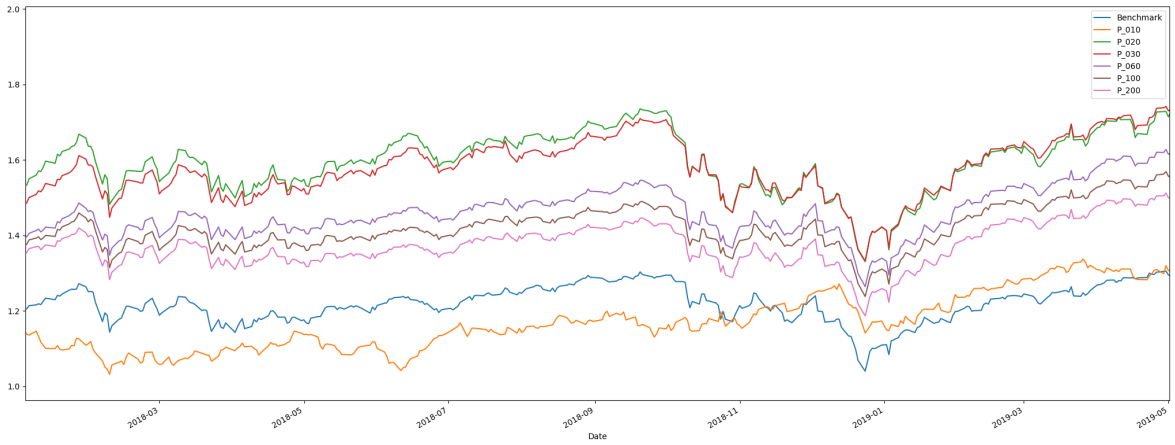

Figure 6: Cumulative returns of MCPP portfolios vs. benchmark, 01/01/2018 to 01/05/2019

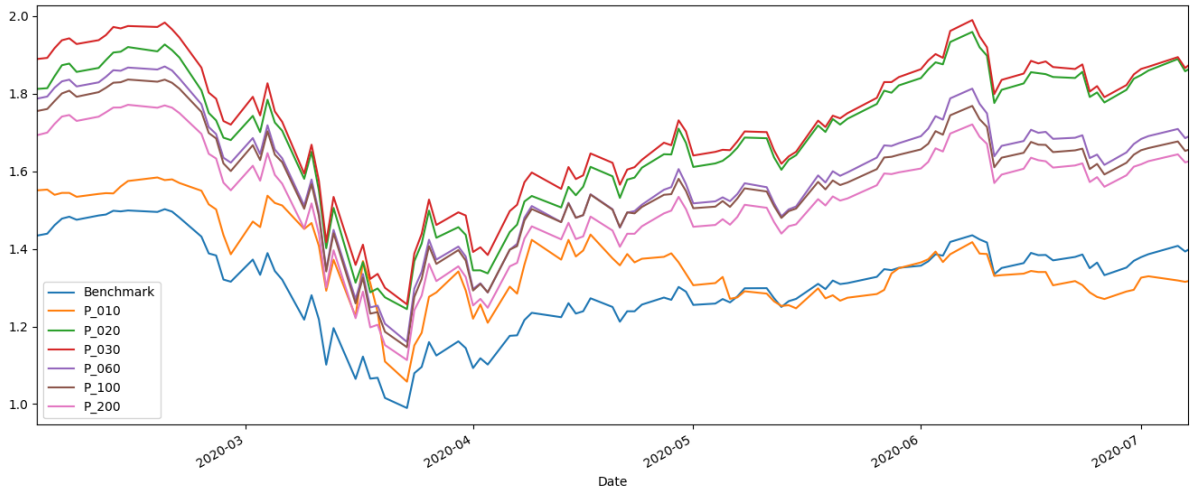

Figure 7: Cumulative returns of MCPP portfolios vs. benchmark, 23/01/2020 to 07/05/2020

First things first, we will provide an explanation for the metrics obtained to the benchmark. As expected, the standard deviation of the benchmark was smaller than any portfolio, since it is a proxy for the market. As presented in Subsection 2.5.2, an alpha of zero means the investment has earned a return adequate for the risk taken, while a beta of 1 indicates that price activity is strongly correlated with the market. This was expected, since the benchmark closely tracks the S&P500 index, which also explains the value for R-squared.

The low value of Sharpe ratio can be an indicative of the highly volatile moments the markets has experienced throughout the period [2] [3]. The IR does not apply, as the S&P500 is the

[2] In 2018, the year begun with markets concerned about the prospect of protectionist trade policies by USA President Donald Trump, and closed with a down on S&P 500 index, due to the trade war between USA and China.

[3] In 2020, markets crashed due to the Coronavirus pandemics. It began on 20 February 2020 and ended on 7 April, being the fastest fall in global stock markets in financial history and the most devastating crash since the Wall Street Crash of 1929.

benchmark, and the return was annualized by the number of trading days (252) [Wikipedia, 2020].

Regargind our hypothesis, the volatility of the portfolios didn't decreased, as given by their standard deviation. From table 5 we can observe that while portfolios P_20, P_30, and P_060 behaved as expected (that is, as we increase the size of stocks, the standard deviation decreased), P_200 presented a higher standard deviation than P_100. Therefore, Hypothesis 1 was refuted.

Although Hypothesis 1 was refuted, with the exception of P_010, we can see that as the number of stocks increase, the alpha, beta and annualized returns of the portfolios decrease. This gets evident on Figure 9, where as the portfolios grows in size, its returns better fit the returns of the benchmark. However, we can also observe the presence of outliers, even in the larger portfolios.

In contrary to our belief, P_10 (the smallest portfolio) underperformed the benchmark and the other portfolios both by risk and return (as clearly demonstrated by the erratic behavior of its cumulative returns, low value of alpha, and negative IR). It's worth noting P_10 presented a low value of R-squared, which indicates little of its movements can be explained by the index. This evidence suggests that 10 is an insufficient number of stocks for an investor to enjoy the benefits of diversification, given the portfolio presented a high level of idiosyncratic risk.

Therefore, according to the performance presented by P_010 and P_020, Hypothesis 2 was also refuted. From Hypothesis 2, we would expect to see better figures for portfolio P_010, however it underperformed P_020 in both volatility (risk) and returns. The same happens if we compare portfolios P_020 and P_030. By the other hand, we can observe an increase of alpha and returns as we get to P_030. This evidence reinforces the belief in the existence of an efficient frontier, as proposed by Markowitz [1952].

Portfolios P_20 and P_30 both overperformed the benchmark and their counterparts in terms of returns, as indicated by their high values of alpha, IR and annualized returns. Furthermore, P_20 and P_30 exhibited high values of standard deviation, beta and R-squared, indicating they are more volatile (riskier) than the index. This evidence suggests that an investor who is looking for higher returns and is able to withstand greater portfolio volatility, should aim at holding 20 to 30 stocks — tests can be performed to determine the optimal size.

Whilst no portfolio presented a standard deviation lower than the benchmark, most of them presented better figures of Sharpe Ratio. This indicates that, although slightly more volatile, the portfolios presented a better risk-adjusted performance. The higher volatility of the portfolios

is evident on Figures 6 and 7, where we can speculate higher drawdown [4] figures, however we didn't used this metric due to its complications (namely, the requirement to create test scenarios).

An important observation can be made by analysing the metrics of P_030. While the standard deviation of this portfolio was higher than others, its high values of Sharpe and information ratios proves this was in fact the best risk-adjusted portfolio of all. Finally, we conclude the optimization model proposed was able to generate better risk-adjusted portfolios than the benchmark.

[4] A drawdown is a peak-to-trough decline during a specific period for an investment. A drawdown is usually quoted as the percentage between the peak and the subsequent trough.

Chapter 5

Conclusions

The selection of investments is a fundamental problem in the area of finance and of great relevance both in academia and industry. Although the groundbreaking works of Harry Markowitz on the Modern Portfolio Theory (MPT) have revolutionized the economic theory of finance, for many years, its principles for portfolio selection were poorly implemented due to the excessive number of estimates needed to calculate portfolio risk.

In this work, we proposed and evaluated a novel approach to investment portfolio diversification, named Minimum Correlated Portfolio Problem (MCPP), which seeks to reduce the risk of a diversified portfolio by minimizing the sum of correlations of the stocks that make it. This approach builds upon the concepts of MPT, which suggests the reduction of correlation between assets in a portfolio as a good way to reduce its diversifiable risk. However, in opposite to the ideas of weight optimization of MPT, we adopted the 1/N rule (also known as naive diversification strategy).

The MCPP was mathematically defined, and proof of NP-completeness was given (that is, it belongs to the NP-hard problem class). Furthermore, we proposed and implemented an exact branch-and-bound method which uses a linear model proposed by the authors.

An instance for MCPP was defined an generated with 503 stocks that composed the S&P500 index. Besides this instance, we tested the proposed resolution method against well known instances of the Maximum Diversity problem, namely: GKD, SOM and MDG.

Regarding the operational aspects of this method, we adopted a non-speculative approach, focusing on holding the assets selected by the model for longer periods of time, and using a lump-sum investing strategy. Furthermore, we ignored associated costs and simplified the calculation of returns (ignoring the payment of dividends), in order to simplify the simulations.

Investment simulations with the proposed selection model were able to generate portfolios with a better risk-adjusted profile than the benchmark, although with a slightly higher volatility. Although they do not represent a final word on the effectiveness of the method, the risk metrics showed a good result for most of the portfolios generated by the optimizer.

Finally, our experimental results refuted the formulated hypotheses. Instead, evidence sug-

gested the existence of an efficient frontier, as defined by Markowitz [1952]. According to our analysis, an investor looking for a better risk-adjusted portfolio should hold from 20 to 30 stocks. Portfolios with lower or higher number of stocks presented lower standard deviation, but they also presented inferior values of alpha, beta, Sharpe and information ratios, as well as lower return figures.

Future work opportunities include:

- The investigation of the method's effectiveness, through the assessment of other risk measures, involving the creation of scenarios, such as: Value-at-Risk (VaR), drawdown, among others.

- Evaluation of the selection model under other stock markets, involving several periods of time and market volatility, and considering variables that were omitted, such as acquisition cost, payment of dividends, and changes in the composition of the index.

- The comparison of the performance of the proposed portfolio selection model with other portfolio selection models, such as the medium-variance model, and including other investment strategies, such as trading systems; and, also, with other stocks other than that restricted to the market index.

- Study of the application of other selection techniques for the pre-selection of stocks for the proposed model. Assess whether companies in perennial segments, for example, could generate less volatile portfolios, but maintaining the optimized risk-adjusted profile.

- The refinement of the proposed portfolio selection model to incorporate weight optimization, as proposed by the MPT, and produce investment strategies more adherent to real investment problems.

Bibliography

(2015). *IAS 32 Financial Instruments: Presentation*, chapter 22, pages 273–282. John Wiley Sons, Ltd.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. Chapman Hall/CRC Texts in Statistical Science (Book 12). Chapman and Hall/CRC.

Aringhieri, R., Cordone, R., and Melzani, Y. (2008). Tabu search versus grasp for the maximum diversity problem. *4OR*, 6(1):45–60.

Bai, Y. and Green, C. J. (2010). International diversification strategies: Re-visited from the risk perspective. *The Journal of Banking and Finance*, 34:236–245.

Bailey, D. H. and Prado, M. M. L. D. (2012). The sharpe ratio indifference curve.

Bazin, D. (2017). *Faça Fortuna com Ações – Antes que seja tarde*. Editora Cla Cultural Ltda.

Benjelloun, H. (2010). Evans and archer - forty years later. 7:98–104.

Black, F. and Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43.

Blatt, S. L. (2004). An in-depth look at the information ratio. Master's thesis, Worcester Polytechnic Institute. Last accessed on 20 November 2020.

Blog, E. T. (2014a). The coefficient of determination or r2. https://economictheoryblog.com/2014/11/05/the-coefficient-of-determination-latex-r2/. Last accessed 15 November 2020.

Blog, E. T. (2014b). Relationship between coefficient of determination squared pearson correlation coefficient. https://economictheoryblog.com/2014/11/05/proof/. Last accessed 15 November 2020.

Bodie, Z., Kane, A., and Marcus, A. J. (2008). *Investments, 8th Edition*. McGraw-Hill/Irwin.

Bourgi, S. (2019). Know the different ways to evaluate portfolio risk. Last accessed on 20 November 2020.

Brimberg, J., Mladenović, N., Urošević, D., and Ngai, E. (2009). Variable neighborhood search for the heaviest k-subgraph. *Computers & Operations Research*, 36(11):2885–2891.

Brooke, G. T. F. (2010). Uncertainty, profit and entrepreneurial action: Frank knight's contribution reconsidered. *Journal of the History of Economic Thought*, 32(2):221–235.

Chague, F. and Giovannetti, B. (2020). Day-trading stocks for a living? *Brazilian Review of Finance*, 18(3):1–4.

Chan, L. K., Karceski, J., and Lakonishok, J. (1999). On portfolio optimization: Forecasting covariances and choosing the risk model. Working Paper 7039, National Bureau of Economic Research.

Chin, M. (2020). Spy vs. voo: Is there any difference? [Online; Last accessed 02 November 2020].

Chiu, M. C. and Wong, H. Y. (2014). Mean–variance portfolio selection with correlation risk. *Journal of Computational and Applied Mathematics*, 263:432 – 444.

Chong, J. and Phillips, G. M. (2013). Portfolio size: Revisited. *The Journal of Wealth Management*, 15:49–60.

Choueifaty, Y. and Coignard, Y. (2008). Toward maximum diversification. *The Journal of Portfolio Management*, 35(1):40–51.

Daniel, W. (1990). *Applied nonparametric statistics*. The Duxbury advanced series in statistics and decision sciences. PWS-Kent Publ.

DeMiguel, V., Garlappi, L., and Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953.

Desjardins, J. (2016). All of the world's stock exchanges by size. http://money.visualcapitalist.com/all-of-the-worlds-stock-exchanges-by-size/. Last accessed 01 November 2020.

Dimitrovsky, H. Z. and Silberman, L. H. (2018). Talmud and midrash (judaism): The making of the talmuds: 3rd-6th century.

Discovery Communications. Discovery communications completes acquisition of scripps networks interactive; changes company name to discovery, inc. https://corporate.discovery.com/discovery-newsroom/discovery-communications-completes-acquisition-of-scripps-networks-interactive-changes-company-name-to-discovery-inc/. Last accessed 03 November 2020.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. The Concise Encyclopedia of Statistics. Springer New York.

Domain, D., Louton, D., and Racine, M. (2007). Diversification in portfolios of individual stocks: 100 stocks are not enough. *Financial Review*, 42(04):557–570.

Duarte, A. and Marti, R. (2007). Tabu search and grasp for the maximum diversity problem. 178:71–84.

Ehling, P. and Ramos, S. B. (2006). Geographic versus industry diversification: Constraints matter. *Journal of Empirical Finance*, 13(4):396 – 416. Special Issue: International Finance.

Elder, A. (1993). *Trading for a Living: Psychology, Trading Tactics, Money Management*. Wiley Finance. Wiley.

Elton, E., Gruber, M., Brown, S., and Goetzmann, W. (2002). *Modern Portfolio Theory and Investment Analysis*. Wiley.

Evans, J. L. and Archer, S. H. (1968). Diversification and the reduction of dispersion: An empirical analysis. *Journal of Finance*, 23(5):761–767.

Fama, E. F. and Books, B. (1976). *Foundations Of Finance*. Basic Books.

Feibel, B. J. (2003). *Investment Performance Measurement*. Frank J. Fabozzi series. J. Wiley.

Ferguson, N. (2008). *The Ascent of Money: A Financial History of the World*. A Penguin book : Business/History. Penguin Press.

Finance Informer. The role of financial intermediaries. https://archive.is/20150416230828/http://www.financeinformer.com.au/the-role-of-financial-intermediaries. Last accessed 01 November 2020.

Fisher, L. and Lorie, J. H. (1970). Some studies of variability of returns on investments in common stocks. *The Journal of Business*, 43(2):99–134.

Fletcher, J. (2011). Do optimal diversification strategies outperform the 1/n strategy in u.k. stock returns? *International Review of Financial Analysis*, 20(5):375 – 385.

Frankel, M. (2020). The day-trading boom and why it's so concerning. The Mootley Fool. Last accessed 16 November 2020.

Freitas, F. D. (2008). *Modelo de seleção de carteiras baseado em erros de predição*. PhD thesis, Universidade Federal do Espírito Santo. Monografia (Doutorado em Engenharia Elétrica), UFES (Universidade Federal do Espírito Santo), Espírito Santo, Brazil.

French, C. (2003). The treynor capital asset pricing model. *Journal of Investment Management*, 1(2):60–72.

Frost, J. (2019). *Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models*. James D. Frost.

Fuscaldo, D. (2020). Robinhood's day trading surge will end badly for investors, money managers warn. Forbes. Last accessed 16 November 2020.

Gallego, M., Duarte, A., Laguna, M., and Martí, R. (2009). Hybrid heuristics for the maximum diversity problem. *Computational Optimization and Applications*, 44(3):411.

Gartner, I. R. (2012). Differentiated risk models in portfolio optimization: a comparative analysis of the degree of diversification and performance in the São Paulo Stock Exchange (BOVESPA). *Pesquisa Operacional*, 32:271 – 292.

Gatfaoui, H. (2007). *How Does Systematic Risk Impact Stocks? A Study of the French Financial Market*, pages 183–213. Palgrave Macmillan UK, London.

Glover, F., Kuo, C.-C., and Dhir, K. S. (1998). Heuristic algorithms for the maximum diversity problem. *Journal of information and Optimization Sciences*, 19(1):109–132.

Graham, B. (2009). *The Intelligent Investor, Rev. Ed*. HarperCollins e-books.

Graham, B. and Dodd, D. (2008). *Security Analysis: Sixth Edition, Foreword by Warren Buffett*. Axis 360. McGraw-Hill Education.

Grubel, H. G. (1968). Internationally diversified portfolios: Welfare gains and capital flows. *The American Economic Review*, 58(5):1299–1314.

Guesmi, K., Saadi, S., Abid, I., and Ftiti, Z. (2018). Portfolio diversification with virtual currency: Evidence from bitcoin. *International Review of Financial Analysis*.

Guo, X., Zhang, H., and Tian, T. (2018). Development of stock correlation networks using mutual information and financial big data. *PLOS ONE*, 13(4):1–16.

Haugen, R. (2001). *Modern Investment Theory*. Prentice Hall finance series : Portfolio analysis. Prentice Hall International.

Heckinger, R., President, V., Ruffini, I., Markets, F., Wells, K., and President, V. (2014). 27 understanding derivatives — markets and infrastructure federal reserve bank of chicago, © 2014 over-the-counter (otc) derivatives.

Henrysson, S. (1971). Gathering, analyzing, and using data on test item. In Thorndike, R. L., editor, *Educational measurement*, chapter 1, pages 130–159. Washington, American Council on Education, Oxford.

Hicks, C. (2017). Is false diversification lurking in your portfolio? https://money.usnews.com/investing/buy-and-hold-strategy/articles/2017-12-05/is-false-diversification-lurking-in-your-portfolio.

Holton, G. A. (2003). *Value-at-risk: Theory and Practice*. Academic Press advanced finance series. Academic Press.

Holton, G. A. (2004). Defining risk. *Financial Analysts Journal*, 60(6):19–25.

Investopedia (2020). Using beta to understand a stock's risk. https://www.investopedia.com/investing/beta-gauging-price-fluctuations/. Last accessed 15 November 2020.

Jagannathan, R. and Ma, T. (2002). Risk reduction in large portfolios: Why imposing the wrong constraints helps. Working Paper 8922, National Bureau of Economic Research.

Jakab, Z. and Kumhof, M. (2015). Banks are not intermediaries of loanable funds – and why this matters. Working Paper 529, Bank of England.

Jan, O. (2019). Portfolio beta. https://xplaind.com/749317/portfolio-beta. Last accessed 15 November 2020.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2):389–416.

Jorion, P. (1985). International portfolio diversification with estimation risk. *The Journal of Business*, 58(3):259–278.

Juan Zhan, H., Rea, W., and Rea, A. (2015). An application of correlation clustering to portfolio diversification. *ArXiv e-prints*.

Kennon, J. (2019). Understanding the dow jones industrial average (djia). https://www.thebalance.com/understanding-the-dow-jones-industrial-average-djia-357912.

Kenton, W. (2020). Beta. https://www.investopedia.com/terms/b/beta.asp. Last accessed 15 November 2020.

Keskin, M., Deviren, B., and Kocakaplan, Y. (2011). Topology of the correlation networks among major currencies using hierarchical structure methods. *Physica A: Statistical Mechanics and its Applications*, 390(4):719 – 730.

Kidd, D. (2011). The sharpe ratio and the information ratio. Last accessed on 20 November 2020.

Kishtainy, N. (2012). *The Economics Book*. Big Ideas. Dorling Kindersley.

ldnquant (2011). Correlation between prices or returns? Last accessed 14 November 2020.

Lemke, T. P. and Lins, G. T. (2013). *Soft dollars and other trading activities*. West.

Levy, H. and Sarnat, M. (1970). International diversification of investment portfolios. *The American Economic Review*, 60(4):668–675.

Libesa (2018). Correlation with prices or returns: that is the question. https://quantdare.com/correlation-prices-returns/. Last accessed 14 November 2020.

Lintner, J. (1965a). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4):587–615.

Lintner, J. (1965b). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1):13–37.

Lo, A. W. (2016). What is an index? *The Journal of Portfolio Management*, 42(2):21–36.

Lodi, A., Milano, M., and Toth, P. (2010). *Integration of AI and OR Techniques in Constraint Programming for Combinatorial Optimization Problems: 7th International Conference, CPAIOR 2010, Bologna, Italy, June 14-18, 2010, Proceedings*. LNCS sublibrary: Theoretical computer science and general issues. Springer.

Lovie, P. and Lovie, A. D. (1996). Charles edward spearman, f.r.s. (1863-1945). *Notes and Records of the Royal Society of London*, 50(1):75–88.

Maginn, J., Tuttle, D., McLeavey, D., and Pinto, J. (2007). *Managing Investment Portfolios: A Dynamic Process*. CFA Institute Investment Series. Wiley.

Mantegna, R. N. and Stanley, H. E. (2000). *An Introduction to Econophysics: Correlations and Complexity in Finance*. Cambridge University Press, New York, NY, USA.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *Eur. Phys. J. B*, 11(1):193–197.

Markowitz, H. M. (1952). Portfolio selection. *The Journal of Finance*, 7:77–91.

Markowitz, H. M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. Yale University Press.

Markowitz, H. M. (1971). *Portfolio Selection: Efficient Diversification of Investments*. Cowles Foundation Monograph: No. 16. Yale University Press.

Markowitz, H. M. (1991). Portfolio selection. *Journal of Finance*, 46:469–477.

Markowitz, H. M. (1999). The early history of portfolio theory: 1600–1960. *Financial Analysts Journal*, 55(4):5–16.

Markowitz, H. M. (2020). The sveriges riksbank prize in economic sciences in memory of alfred nobel 1990. https://www.nobelprize.org/prizes/economic-sciences/1990/summary/. Last accessed 18 November 2020.

Martin, E. (2018). Warren buffett and tony robbins agree on the best way to invest your money. https://www.cnbc.com/2018/06/19/warren-buffett-and-tony-robbins-agree-invest-in-index-funds.html.

Martí, G. and Duarte (2010). Maximum diversity problem. Last accessed September 16, 2020.

Martí, R., Gallego, M., and Duarte, A. (2010). A branch and bound algorithm for the maximum diversity problem. *European Journal of Operational Research*, 200(1):36 – 44.

Max, G. (2010). *The Zurich Axioms: The rules of risk and reward used by generations of Swiss bankers*. Harriman House.

Mcclure, B. (2020). What beta means when considering a stock's risk. https://www.investopedia.com/investing/beta-know-risk. Last accessed 15 November 2020.

Micklitsch, J. (2018). Using beta to understand a stock's risk. https://ancora.net/portfolio-beta-number-care/. Last accessed 15 November 2020.

Mittelmann, H. D. (2010). Recent benchmarks of optimization software. In *22nd European Conference on Operational Research, EURO XXII, Prague, Czech Republic*. Dept of Math and Stats Arizona State University.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica*, 34(4):768–783.

Motulsky, H. (2018). Can r^2 be greater than 1? Cross Validated. url:https://stats.stackexchange.com/q/335498 (version: 2018-03-19).

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med J*, 24(3):69–71.

Neilands, S. A. G. B. S. T. B. (2016). *Primer of Applied Regression Analysis of Variance*. McGraw-Hill Education / Medical, 3rd edition edition.

Onnela, J.-P., Kaski, K., and Kertész, J. (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2):353–362.

Palubeckis, G. (2007). Iterated tabu search for the maximum diversity problem. *Applied Mathematics and Computation*, 189(1):371–383.

Patton, A. J. and Timmermann, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the capm, and portfolio sorts. *Journal of Financial Economics*, 98(3):605 – 625.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. volume 58, page 240–242. Taylor & Francis.

Persky, J. (1995). Retrospectives: The ethology of homo economicus. *The Journal of Economic Perspectives*, 9(2):221–231.

Pflug, G. C., Pichler, A., and Wozabal, D. (2012). The 1/n investment strategy is optimal under high model ambiguity. *Journal of Banking Finance*, 36(2):410 – 417.

Radcliffe, R. (1997). *Investment: Concepts, Analysis, Strategy*. Addison-Wesley series in economics. Addison-Wesley.

Rea, A. and Rea, W. (2014). Visualization of a stock market correlation matrix. *Physica A: Statistical Mechanics and its Applications*, 400:109 – 123.

Reamer, N. and Downing, J. (2016). *Investment: A History*. Columbia University Press.

Rice, J. (2006). *Mathematical Statistics and Data Analysis*. Number p. 3 in Advanced series. Cengage Learning.

Rodgers, J. L. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Rollinger, T. N. and Hoffman, S. T. (2015). Sortino: A sharper ratio. Working paper, Red Rock Capital.

Rosén, F. (2006). Correlation based clustering of the stockholm stock exchange.

Roy, A. D. (1952). Safety first and the holding of assets. *Econometrica*, 20(3):431–449.

Sahinidis, N. V. (2019). Mixed-integer nonlinear programming 2018. *Optimization and Engineering*, 20(2):301–306.

Sardana, S. (2020). The day-trading boom is a 'welcome phenomenon,' and has actually helped to reduce market volatility, a veteran wall street trader says. Forbes. Last accessed 16 November 2020.

Scholz, H. (2007). Refinements to the sharpe ratio: Comparing alternatives for bear markets. *Journal of Asset Management*, 7:347–357.

Schouhamer Immink, K. and Weber, J. (2014). Minimum pearson distance detection for multi-level channels with gain and/or offset mismatch. *Information Theory, IEEE Transactions on*, 60:5966–5974.

Sharpe, W. F. (1963). A simplified model for portfolio analysis. *Management Science*, 9:277–293.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, 39(1):119–138.

Sharpe, W. F. (1975). Adjusting for risk in portfolio performance measurement. *The Journal of Portfolio Management*, 1(2):29–34.

Sharpe, W. F. (1994). The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58.

Shiller, R. (1999). Human behavior and the efficiency of the financial system. In Taylor, J. B. and Woodford, M., editors, *Handbook of Macroeconomics*, volume 1, Part C, chapter 20, pages 1305–1340. Elsevier, 1 edition.

Shiller, R. (2000). *Irrational exuberance*. Princeton University Press, Princeton, NJ.

Silva, G. C., Ochi, L. S., and Martins, S. L. (2004). Experimental comparison of greedy randomized adaptive search procedures for the maximum diversity problem. In *International Workshop on Experimental and Efficient Algorithms*, pages 498–512. Springer.

Sortino, F. A. and van der Meer, R. (1991). Downside risk. *The Journal of Portfolio Management*, 17(4):27–31.

Spaulding, W. (2020). *The Pauper'$ Money Book: Increase Your Wealth by Saving More, Investing More, and Earning More*. Independently Published.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Statman, M. (1987). How many stocks make a diversified portfolio? *Journal of Financial and Quantitative Analysis*, 22(03):353–363.

Statman, M. (2000). The diversification puzzle. *Financial Analysts Journal*, 60(04):44–53.

Steinbach, M. C. (2001). Markowitz revisited: Mean-variance models in financial portfolio analysis. *SIAM Review*, 43(1):31–85.

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

Treynor, J. L. (1961). Market value, time, and risk.

Treynor, J. L. (1962). Toward a theory of market value of risky assets. In *Asset Pricing and Portfolio Performance: Models, Strategy and Performance Metrics*. Risk Books.

Treynor, J. L. (1965). How to rate management of investment funds. *Harvard Business Review*, 43(1):63–75.

Tsang, A. (2019). 5 pieces of advice from john bogle. https://www.nytimes.com/2019/01/17/business/mutfund/john-bogle-vanguard-investment-advice.html.

von Mises, L. (1949). *Human Action. A Treatise on Economics*. Yale University Press.

Wikipedia. List of sp 500 companies. Online; Last accessed 02 November 2020.

Wikipedia. Security characteristic line. Last accessed on 18 November 2020.

Wikipedia (2020). Trading day — Wikipedia, the free encyclopedia. Last accessed on 29 September 2012.

Witte, R. and Witte, J. (2016). *Statistics*. Wiley.

YCHARTS. 1 month treasury rate. Online; Last accessed 02 November 2020.

Appendices

Appendix A

Figures

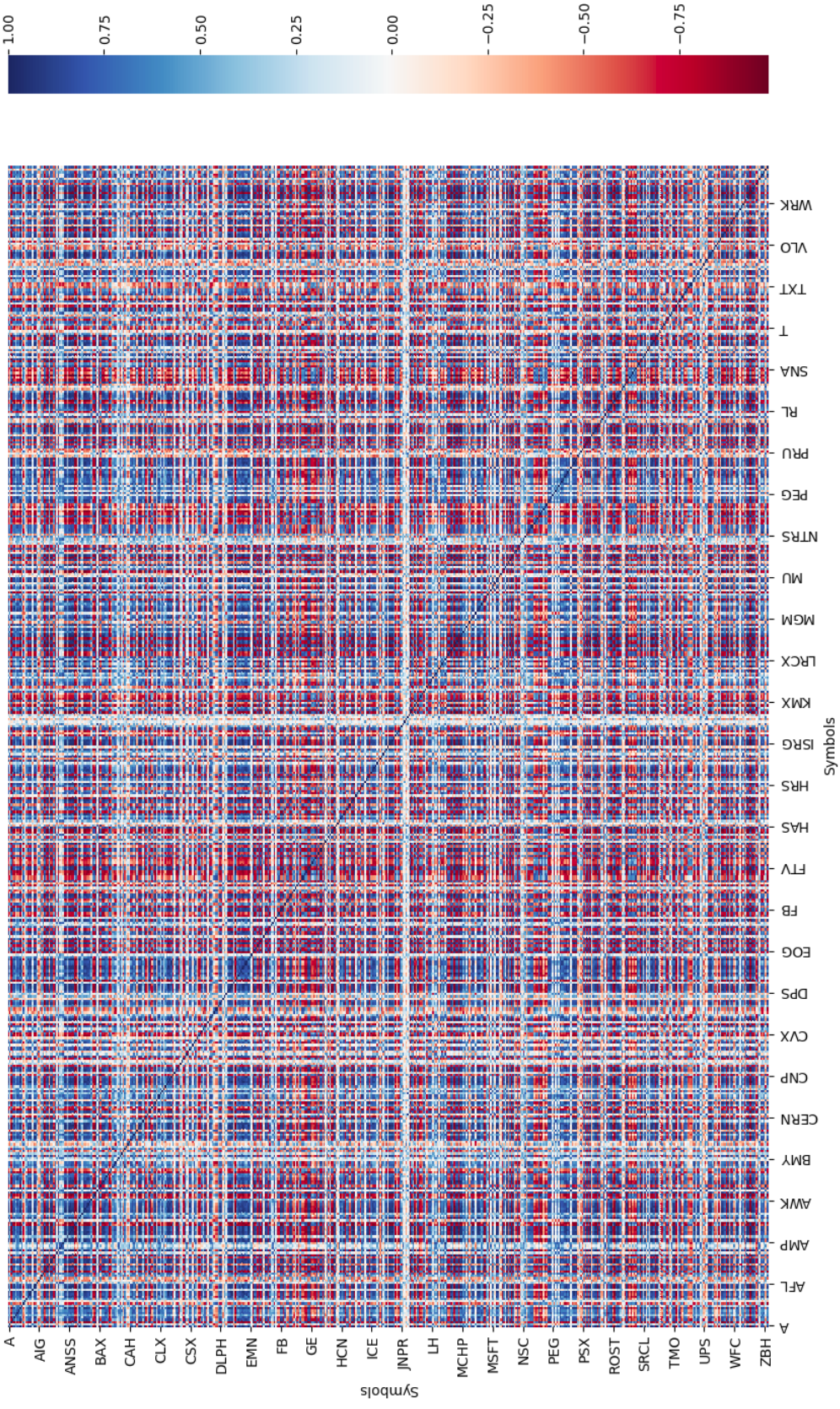

Figure 8: Graphic of the stocks correlation matrix

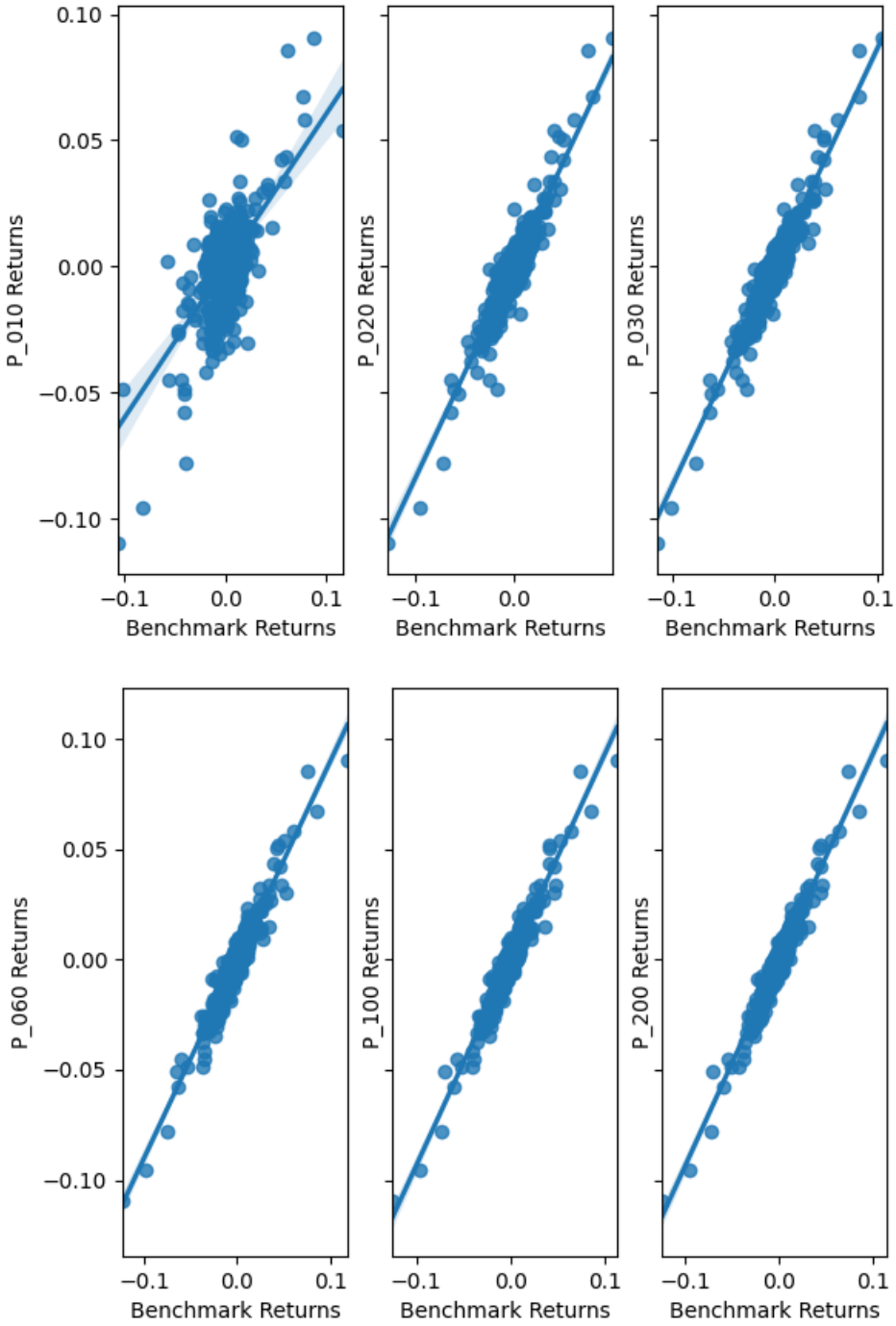

Figure 9: Scatter plots of the cumulative returns of MCPP portfolios vs. benchmark

Appendix B

Tables

Table 6: List of S&P 500 stocks used on the what-if analysis.

Ticker	Company
A	Agilent
AAL	American Airlines Group
AAP	Advance Auto Parts
AAPL	Apple Inc.
ABBV	AbbVie
ABC	AmerisourceBergen Corp
ABT	Abbott Laboratories
ACN	Accenture plc
ADBE	Adobe Systems Inc
ADI	Analog Devices, Inc.
ADM	Archer-Daniels-Midland Co
ADP	Automatic Data Processing
ADS	Alliance Data Systems
ADSK	Autodesk Inc
AEE	Ameren Corp
AEP	American Electric Power
AES	AES Corp
AFL	AFLAC Inc
AIG	American International Group, Inc.
AIV	Apartment Investment & Mgmt
AIZ	Assurant Inc
AJG	Arthur J. Gallagher & Co.
AKAM	Akamai Technologies Inc

Table 6: List of S&P 500 stocks used on the what-if analysis.

ALB	Albemarle Corp
ALGN	Align Technology
ALK	Alaska Air Group Inc
ALL	Allstate Corp
ALLE	Allegion
ALXN	Alexion Pharmaceuticals
AMAT	Applied Materials Inc
AMD	Advanced Micro Devices Inc
AME	AMETEK Inc
AMG	Affiliated Managers Group Inc
AMGN	Amgen Inc
AMP	Ameriprise Financial
AMT	American Tower Corp A
AMZN	Amazon.com Inc
ANSS	ANSYS
ANTM	Anthem Inc.
AON	Aon plc
AOS	A.O. Smith Corp
APA	Apache Corporation
APD	Air Products & Chemicals Inc
APH	Amphenol Corp
ARE	Alexandria Real Estate Equities
ARNC	Arconic Inc
ATVI	Activision Blizzard
AVB	AvalonBay Communities, Inc.
AVGO	Broadcom
AVY	Avery Dennison Corp
AWK	American Water Works Company Inc
AXP	American Express Co
AYI	Acuity Brands Inc
AZO	AutoZone Inc

Table 6: List of S&P 500 stocks used on the what-if analysis.

BA	Boeing Company
BAC	Bank of America Corp
BAX	Baxter International Inc.
BBY	Best Buy Co. Inc.
BDX	Becton Dickinson
BEN	Franklin Resources
BF-B	Brown-Forman Corporation
BIIB	BIOGEN IDEC Inc.
BK	The Bank of New York Mellon Corp.
BLK	BlackRock
BLL	Ball Corp
BMY	Bristol-Myers Squibb
BRK-B	Berkshire Hathaway
BSX	Boston Scientific
BWA	BorgWarner
BXP	Boston Properties
C	Citigroup Inc.
CAG	ConAgra Foods Inc.
CAH	Cardinal Health Inc.
CAT	Caterpillar Inc.
CB	Chubb Limited
CBOE	Cboe Global Markets
CCI	Crown Castle International Corp.
CCL	Carnival Corp.
CDNS	Cadence Design Systems
CERN	Cerner
CF	CF Industries Holdings Inc
CFG	Citizens Financial Group
CHD	Church & Dwight
CHK	Chesapeake Energy
CHRW	C. H. Robinson Worldwide

Table 6: List of S&P 500 stocks used on the what-if analysis.

CHTR	Charter Communications
CI	CIGNA Corp.
CINF	Cincinnati Financial
CL	Colgate-Palmolive
CLX	The Clorox Company
CMA	Comerica Inc.
CMCSA	Comcast A Corp
CME	CME Group Inc.
CMG	Chipotle Mexican Grill
CMI	Cummins Inc.
CMS	CMS Energy
CNC	Centene Corporation
CNP	CenterPoint Energy
COF	Capital One Financial
COG	Cabot Oil & Gas
COL	Rockwell Collins
COO	The Cooper Companies
COP	ConocoPhillips
COST	Costco Co.
COTY	Coty, Inc
CPB	Campbell Soup
CRM	Salesforce.com
CSCO	Cisco Systems
CSX	CSX Corp.
CTAS	Cintas Corporation
CTL	CenturyLink Inc
CTSH	Cognizant Technology Solutions
CTXS	Citrix Systems
CVS	CVS Health
CVX	Chevron Corp.
CXO	Concho Resources

Table 6: List of S&P 500 stocks used on the what-if analysis.

D	Dominion Resources
DAL	Delta Air Lines
DE	Deere & Co.
DFS	Discover Financial Services
DG	Dollar General
DGX	Quest Diagnostics
DHI	D. R. Horton
DHR	Danaher Corp.
DIS	The Walt Disney Company
DISCA	Discovery Communications-A
DISCK	Discovery Communications-C
DISH	Dish Network
DLR	Digital Realty Trust
DLTR	Dollar Tree
DOV	Dover Corp.
DRE	Duke Realty Corp
DRI	Darden Restaurants
DTE	DTE Energy Co.
DUK	Duke Energy
DVA	DaVita Inc.
DVN	Devon Energy Corp.
DXC	DXC Technology
EA	Electronic Arts
EBAY	eBay Inc.
ECL	Ecolab Inc.
ED	Consolidated Edison
EFX	Equifax Inc.
EIX	Edison Intĺ
EL	Estee Lauder Cos.
EMN	Eastman Chemical
EMR	Emerson Electric Company

Table 6: List of S&P 500 stocks used on the what-if analysis.

EOG	EOG Resources
EQIX	Equinix
EQR	Equity Residential
EQT	EQT Corporation
ES	Eversource Energy
ESS	Essex Property Trust, Inc.
ETFC	E*Trade
ETN	Eaton Corporation
ETR	Entergy Corp.
EW	Edwards Lifesciences
EXC	Exelon Corp.
EXPD	Expeditors Intĺ
EXPE	Expedia Inc.
EXR	Extra Space Storage
F	Ford Motor
FAST	Fastenal Co
FB	Facebook
FBHS	Fortune Brands Home & Security
FCX	Freeport-McMoran Cp & Gld
FDX	FedEx Corporation
FE	FirstEnergy Corp
FFIV	F5 Networks
FIS	Fidelity National Information Services
FISV	Fiserv Inc
FITB	Fifth Third Bancorp
FL	Foot Locker Inc
FLIR	FLIR Systems
FLR	Fluor Corp.
FLS	Flowserve Corporation
FMC	FMC Corporation
FRT	Federal Realty Investment Trust

Table 6: List of S&P 500 stocks used on the what-if analysis.

FTI	FMC Technologies Inc.
FTV	Fortive Corp
GD	General Dynamics
GE	General Electric
GILD	Gilead Sciences
GIS	General Mills
GLW	Corning Inc.
GM	General Motors
GOOG	Alphabet Inc Class C
GOOGL	Alphabet Inc Class A
GPC	Genuine Parts
GPN	Global Payments Inc
GPS	Gap (The)
GRMN	Garmin Ltd.
GS	Goldman Sachs Group
GT	Goodyear Tire & Rubber
GWW	Grainger (W.W.) Inc.
HAL	Halliburton Co.
HAS	Hasbro Inc.
HBAN	Huntington Bancshares
HBI	Hanesbrands Inc
HCA	HCA Holdings
HD	Home Depot
HES	Hess Corporation
HIG	Hartford Financial Svc.Gp.
HLT	Hilton Worldwide Holdings Inc
HOG	Harley-Davidson
HOLX	Hologic
HON	Honeywell Intĺ Inc.
HP	Helmerich & Payne
HPE	Hewlett Packard Enterprise

Table 6: List of S&P 500 stocks used on the what-if analysis.

HPQ	HP Inc.
HRB	Block H&R
HRL	Hormel Foods Corp.
HSIC	Henry Schein
HST	Host Hotels & Resorts
HSY	The Hershey Company
HUM	Humana Inc.
IBM	International Business Machines
ICE	Intercontinental Exchange
IDXX	IDEXX Laboratories
IFF	Intl Flavors & Fragrances
ILMN	Illumina Inc
INCY	Incyte
INFO	IHS Markit Ltd.
INTC	Intel Corp.
INTU	Intuit Inc.
IP	International Paper
IPG	Interpublic Group
IR	Ingersoll-Rand PLC
IRM	Iron Mountain Incorporated
ISRG	Intuitive Surgical Inc.
IT	Gartner Inc
ITW	Illinois Tool Works
IVZ	Invesco Ltd.
JBHT	J. B. Hunt Transport Services
JCI	Johnson Controls International
JNJ	Johnson & Johnson
JNPR	Juniper Networks
JPM	JPMorgan Chase & Co.
JWN	Nordstrom
K	Kellogg Co.

Table 6: List of S&P 500 stocks used on the what-if analysis.

KEY	KeyCorp
KHC	Kraft Heinz Co
KIM	Kimco Realty
KLAC	KLA-Tencor Corp.
KMB	Kimberly-Clark
KMI	Kinder Morgan
KMX	Carmax Inc
KO	Coca Cola Company
KR	Kroger Co.
KSS	Kohlś Corp.
KSU	Kansas City Southern
L	Loews Corp.
LB	L Brands Inc.
LEG	Leggett & Platt
LEN	Lennar Corp.
LH	Laboratory Corp. of America Holding
LKQ	LKQ Corporation
LLY	Lilly (Eli) & Co.
LMT	Lockheed Martin Corp.
LNC	Lincoln National
LNT	Alliant Energy Corp
LOW	Loweś Cos.
LRCX	Lam Research
LUV	Southwest Airlines
LYB	LyondellBasell
M	Macyś Inc.
MA	Mastercard Inc.
MAA	Mid-America Apartments
MAC	Macerich
MAR	Marriott Intĺ.
MAS	Masco Corp.

Table 6: List of S&P 500 stocks used on the what-if analysis.

MAT	Mattel Inc.
MCD	McDonaldś Corp.
MCHP	Microchip Technology
MCK	McKesson Corp.
MCO	Moodyś Corp
MDLZ	Mondelez International
MDT	Medtronic plc
MET	MetLife Inc.
MGM	MGM Resorts International
MHK	Mohawk Industries
MKC	McCormick & Co.
MLM	Martin Marietta Materials
MMC	Marsh & McLennan
MMM	3M Company
MNST	Monster Beverage
MO	Altria Group Inc
MOS	The Mosaic Company
MPC	Marathon Petroleum
MRK	Merck & Co.
MRO	Marathon Oil Corp.
MS	Morgan Stanley
MSFT	Microsoft Corp.
MSI	Motorola Solutions Inc.
MTB	M&T Bank Corp.
MTD	Mettler Toledo
MU	Micron Technology
MYL	Mylan N.V.
NAVI	Navient
NBL	Noble Energy Inc
NCLH	Norwegian Cruise Line Holdings
NDAQ	NASDAQ OMX Group

Table 6: List of S&P 500 stocks used on the what-if analysis.

NEE	NextEra Energy
NEM	Newmont Mining Corp. (Hldg. Co.)
NFLX	Netflix Inc.
NI	NiSource Inc.
NKE	Nike
NLSN	Nielsen Holdings
NOC	Northrop Grumman Corp.
NOV	National Oilwell Varco Inc.
NRG	NRG Energy
NSC	Norfolk Southern Corp.
NTAP	NetApp
NTRS	Northern Trust Corp.
NUE	Nucor Corp.
NVDA	Nvidia Corporation
NWL	Newell Brands
NWS	News Corp. Class B
NWSA	News Corp. Class A
O	Realty Income Corporation
OKE	ONEOK
OMC	Omnicom Group
ORCL	Oracle Corp.
ORLY	OŔeilly Automotive
OXY	Occidental Petroleum
PAYX	Paychex Inc.
PBCT	Peopleś United Financial
PCAR	PACCAR Inc.
PCG	PG&E Corp.
PDCO	Patterson Companies
PEG	Public Serv. Enterprise Inc.
PEP	PepsiCo Inc.
PFE	Pfizer Inc.

Table 6: List of S&P 500 stocks used on the what-if analysis.

PFG	Principal Financial Group
PG	Procter & Gamble
PGR	Progressive Corp.
PH	Parker-Hannifin
PHM	Pulte Homes Inc.
PKG	Packaging Corporation of America
PKI	PerkinElmer
PLD	Prologis
PM	Philip Morris International
PNC	PNC Financial Services
PNR	Pentair Ltd.
PNW	Pinnacle West Capital
PPG	PPG Industries
PPL	PPL Corp.
PRGO	Perrigo
PRU	Prudential Financial
PSA	Public Storage
PSX	Phillips 66
PVH	PVH Corp.
PWR	Quanta Services Inc.
PXD	Pioneer Natural Resources
PYPL	PayPal
QCOM	QUALCOMM Inc.
QRVO	Qorvo
RCL	Royal Caribbean Cruises Ltd
RE	Everest Re Group Ltd.
REG	Regency Centers Corporation
REGN	Regeneron
RF	Regions Financial Corp.
RHI	Robert Half International
RJF	Raymond James Financial Inc.

Table 6: List of S&P 500 stocks used on the what-if analysis.

RL	Polo Ralph Lauren Corp.
RMD	ResMed
ROK	Rockwell Automation Inc.
ROP	Roper Industries
ROST	Ross Stores
RRC	Range Resources Corp.
RSG	Republic Services Inc
SBAC	SBA Communications
SBUX	Starbucks Corp.
SCHW	Charles Schwab Corporation
SEE	Sealed Air
SHW	Sherwin-Williams
SIG	Signet Jewelers
SJM	JM Smucker
SLB	Schlumberger Ltd.
SLG	SL Green Realty
SNA	Snap-On Inc.
SNPS	Synopsys Inc.
SO	Southern Co.
SPG	Simon Property Group Inc
SPGI	S&P Global, Inc.
SRCL	Stericycle Inc
SRE	Sempra Energy
STI	SunTrust Banks
STT	State Street Corp.
STX	Seagate Technology
STZ	Constellation Brands
SWK	Stanley Black & Decker
SWKS	Skyworks Solutions
SYF	Synchrony Financial
SYK	Stryker Corp.

Table 6: List of S&P 500 stocks used on the what-if analysis.

SYY	Sysco Corp.
T	AT&T Inc
TAP	Molson Coors Brewing Company
TDG	TransDigm Group
TEL	TE Connectivity Ltd.
TGT	Target Corp.
TIF	Tiffany & Co.
TJX	TJX Companies Inc.
TMO	Thermo Fisher Scientific
TPR	Tapestry, Inc.
TRIP	TripAdvisor
TROW	T. Rowe Price Group
TRV	The Travelers Companies Inc.
TSCO	Tractor Supply Company
TSN	Tyson Foods
TXN	Texas Instruments
TXT	Textron Inc.
UA	Under Armour
UAA	Under Armour
UAL	United Continental Holdings
UDR	UDR Inc
UHS	Universal Health Services, Inc.
ULTA	Ulta Salon Cosmetics & Fragrance Inc
UNH	United Health Group Inc.
UNM	Unum Group
UNP	Union Pacific
UPS	United Parcel Service
URI	United Rentals, Inc.
USB	U.S. Bancorp
V	Visa Inc.
VAR	Varian Medical Systems

Table 6: List of S&P 500 stocks used on the what-if analysis.

VFC	V.F. Corp.
VLO	Valero Energy
VMC	Vulcan Materials
VNO	Vornado Realty Trust
VRSK	Verisk Analytics
VRSN	Verisign Inc.
VRTX	Vertex Pharmaceuticals Inc
VTR	Ventas Inc
VZ	Verizon Communications
WAT	Waters Corporation
WBA	Walgreens Boots Alliance
WDC	Western Digital
WEC	Wec Energy Group Inc
WFC	Wells Fargo
WHR	Whirlpool Corp.
WLTW	Willis Towers Watson
WM	Waste Management Inc.
WMB	Williams Cos.
WMT	Wal-Mart Stores
WRK	WestRock Company
WU	Western Union Co
WY	Weyerhaeuser Corp.
WYND	Wyndham Destinations
WYNN	Wynn Resorts Ltd
XEC	Cimarex Energy
XEL	Xcel Energy Inc
XLNX	Xilinx Inc
XOM	Exxon Mobil Corp.
XRAY	Dentsply Sirona
XRX	Xerox Corp.
XYL	Xylem Inc.

Table 6: List of S&P 500 stocks used on the what-if analysis.

YUM	Yum! Brands Inc
ZBH	Zimmer Biomet Holdings
ZION	Zions Bancorp
ZTS	Zoetis

Table 7: List of S&P 500 stocks removed from the what-if analysis.

Ticker	Description
AET	Aetna Inc. was acquired by CVS Health on November 28, 2018.
AGN	Allergan (NYSE: AGN) was acquired by AbbVie (NYSE: ABBV)
ANDV	Andeavor was acquired by Marathon Petroleum Corp. on 3 October, 2018.
APC	Anadarko Petroleum (NYSE: APC) was bought by Occidental Petroleum (NYSE: OXY).
BBT	BB&T Corp. (NYSE: BBT) and Atlanta's SunTrust Banks Inc (NYSE:STI) merged on December 6, 2019.
BCR	C.R.Bard, Inc. was acquired by Becton Dickinson, fusion was completed on December 29, 2017.
BHGE	The company divested from General Eletric in 2019, becaming Baker Hughes Company (NYSE: BKR).
CA	CA Technologies was acquired by Broadcom on November 5, 2018.
CBG	CBRE Group, Inc. updated ticker on Monday, March 19, 2018.
CBS	Viacom and CBS Corp. merged as ViacomCBS.
CELG	Celgene Corp. was acquired by Bristol-Myers Squibb.
CSRA	CSRA was acquired by General Dynamics on April 02, 2018.
DLPH	Delphi Technologies was acquired by BorgWarner Inc. on October 2, 2020.
DPS	Dr Pepper Snapple Group was acquired by Keurig Green Mountain, stock changed from NYSE to NASDAQ.
DWDP	DowDuPont changed name to DuPont on June 1, 2019 and updated ticker (NYSE: DD) on June 3, 2019.
ESRX	Express Scripts Holding Company was acquired by Cigna on December 20, 2018.
EVHC	Envision Healthcare was acquired by Kohlberg Kravis Roberts on October 11, 2018.
FOX	21st Century Fox was demerged into 21CF and Fox. Walt Disney Company acquired 21FC on March 20, 2019.
FOXA	21st Century Fox was demerged into 21CF and Fox. Walt Disney Company acquired 21FC on March 20, 2019.
GGP	General Growth Properties Inc. acquired by Brookfield Property Partners on August 28, 2018.
HCN	Welltower Inc. changed its ticker symbol (NYSE: WELL) on February 28, 2018.
HCP	Health Care Property Investors, Inc. changed name to Healthpeak Properties, Inc. on October 2019.
HRS	Harris Corporation and L3 Technologies merged on June 29, 2019.
JEC	Jacobs Engineering Group updated ticker symbol on Dec. 10, 2019.
KORS	Michael Kors Holdings Limited was renamed to Capri Holdings on January 2, 2019.
LLL	Harris Corporation and L3 Technologies merged on June 29, 2019.
LUK	Leucadia National Corporation changed name to Jefferies Financial Group Inc. (NYSE: JEF) on 23 May, 2018.
MON	Monsanto Company was acquried by Bayer on June 7, 2018.
NFX	Newfield Exploration Company was acquired by Ovintiv Inc. (formerly Encana Corporation) on February 2019.
PCLN	The Priceline Group Inc. was renamed to Booking Holdings on February 27, 2018.
PX	Praxair Inc. and Linde AG merged to form Linde PLC.
RHT	Red Hat Inc. was acquired by IBM on July 9, 2019.
RTN	Raytheon Company merged with United Technologies and updated ticker to RTX on April 3, 2020.
SCG	SCANA Corp. was acquired by Dominion Energy on January 2019. By March 2019, the SCANA name were retired.
SNI	Scripps Networks Interactive, Inc. was acquired by Discovery Communications on March 6, 2018.
STI	BB&T Corp. (NYSE: BBT) and Atlanta's SunTrust Banks Inc (NYSE: STI) merged on December 6, 2019.
SYMC	Symantec Corporation was acquired by Broadcom on November 4, 2019 and renamed to NortonLifeLock Inc.
TMK	Torchmark Corporation changed name to Globe Life Inc. on August 8, 2019 and updated ticker on August 9, 2019.
TSS	Total System Services was acquired by Global Payments on September, 2019.
TWX	Time Warner was acquired by AT&T.
UTX	Raytheon Company merged with United Technologies and updated ticker to RTX on April 3, 2020.
VIAB	Viacom and CBS Corp. merged as ViacomCBS.
XL	XL Group was acquired by Axa in September 2018 and was delisted.
BHF	BHF Brighthouse Financial Inc. (BHF) wasn't present in the MCCP instances.
Q	QuintilesIMS (Q) wasn't present in the MCCP instances.
LEN-B	Lennar Corporation (LEN-B) wasn't present in the MCCP instances.

Table 8: List of S&P 500 stocks removed from MCPP portfolios

Ticker	MCPP_1	MCPP_2	MCPP_3	MCPP_4	MCPP_6	MCPP_7
AET				X	X	X
BCR					X	X
DLPH					X	X
HCN						X
MON						X
PCLN						X
PX						X
RHT						X
RTN					X	X
TSS						X
TWX						X
UTX					X	X
XL					X	X