POLITECNICO DI TORINO

Master degree course in
Mathematical Engineering

Master Degree Thesis

A decision support system for planning stock portfolios based on a two-stage itemset-based approach

Supervisors
Prof. Luca Cagliero, Supervisor
Dott. Jacopo Fior, Co-Supervisor

Candidate
Daniele Giovanni Gioia

October 2020

Abstract

Financial Technologies and Intelligent system have become established solutions to support decision making in stock market investment. Hitherto, fundamentals and technical analysis have been milestones for stock price forecasting. Hence they are often taken into account to refine the decision, but it is difficult to assert an analytical combination that is always right for every market or intent. Simultaneously, risk management could avoid torpedos or decisions that do not reflect the risk aversion of the manager, thus specific risk measures are usually consulted to reach robustness and reliability of a financial decision. An effective financial decision support tool should be able to combine the most advanced technological solutions with the experience of financial investors, represented into the knowledge of indicators, ratios and self-judgment in risk aversion. This study aims to propose a decision support system for long-term stock portfolio generation based on the above-mentioned guidelines by a two-stage approach. Starting from an automatic itemset-based mining algorithm that extracts candidate sets of stocks considering the historical performances, suitable portfolios are generated and the user knowledge is put into practice by a second step. In particular, a generalization of the well-known Markowitz mean-variance model, that uses whole portfolios proposals rather than single stocks, optimizes through constraints in diversification, risk measures and financial indicators to fulfil different perspectives.

Acknowledgements

I would like to extend my sincere thanks to prof. Luca Cagliero that has given me his experience. Thanks to this opportunity I was able to ask myself the right questions and to build the appropriate tools for seeking the correct answers. I'm deeply indebted to the PhD student Jacopo Fior for continuous support. I had confrontations in almost every critical step and I thank him for helping me to fill my computational gaps.

During my academic journey in Turin and Eindhoven, I have met colleagues that have become friends, professors that have represented guides and housemates that have changed my life forever.

If I would try to list all of them I am sure I will forget someone, moreover I could not report all the experiences that have defined who I am now. If you are reading this, you have likely been part of my journey and I would like to thank you for enriching my life with yours.

I am grateful to my friends who have been close and yet so far away in Enna. Socials can connect what roads divide.

Last but not least I thank my parents, my brother, my aunt and my uncle that blindly believe in me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Data mining approaches are rapidly spreading in almost every technological field. Financial management and portfolios planning has not been immune, but, on the contrary, it is often the first environments where technologies are suddenly applied. A lot of financial markets exist, such as Derivatives markets, Cryptocurrency or Commodity markets. Perhaps, the most known is the stocks market. Regarding this latter world, also who is not interested certainly have come across an online trading pop-up rather than have heard investment suggestion from websites like Yahoo! or Reddit. Online tools span among different level of provided features, indeed some of them have been built for pundits, whereas others are not so reliable. Anyway, what is clear is that a financial newspaper seems to be overcome or, better, not sufficient anymore to deeply catch what market wants to say. Financial Technology (FinTech) is nowadays an emerging industry that aims to improve activity in finance by technology, hence also by automated systems or decision support systems. Companies, academia and professional users continuously develop and research both new tools that can invest with no human interaction (e.g. Mining Twits) and also decision support systems to help financial decision-making activities. These mechanism usually make use of potentially large databases that involves an abundance of sources.

The data related to the stocks market are usually divided as Technical and Fundamental. This classification separates what is driven by the price of the stocks in the market from what is related to the company itself. There is no guarantee that if a company has, for example, a low revenue this will be reflected in the market. The usual technical data are Prices and Trading Volume, it follows that in a daily mindset the Open price, Closing price, Maximum and Minimum are useful statistics, indeed these values build a ’candlestick’. Regarding fundamentals, this term may be associated with an incredible quantity of different data source such as newspapers rankings, on-line forums gossip, management declarations, ... Among these different sources, the most authoritative are the company financial statements that are often divided into the Balance sheet, Cash flow and Income statement. These latter documents, together with candlesticks, will be taken into account in this study.

Different investment mindsets lead to different financial technologies and, usually, the cornerstone that diversifies these possibilities is the trading frequency.

I will introduce a decision support system assuming a buy-and-hold approach, hence portfolios are bought in a certain moment and then the saver holds for several months or even years the purchased stocks.
Completely different is the intraday market, where day traders looking to make multiple trades throughout a single trading session, paying attention to enter or exit signals.

Countless studies have analyzed the stock-market using both technical and fundamental data, or only one of them, through various data-mining and machine learning techniques. Coherently with the machine learning algorithm classification, supervised and unsupervised approaches have been proposed. Some supervised examples are regression analysis, Support vector machine or decision trees, where the prediction intent is typically in the foreground. On the other hand, unsupervised methods, such as clusters or associations rules, could better make results readable and interpretable for the user, easier to suggest decision conscientiously.
Among unsupervised machine learning possibilities, typically, algorithms used in financial technology are related to clustering or neural networks. What the study involves is a new scheme based on means of weighted itemsets that uses a variant of the frequent itemset mining.

Choosing a portfolio to invest in the stocks market in a buy-and-hold approach is far from easy and Decision Support Systems help to make a sane choice that weight up correctly the available data.
Since historical stock price series do not necessarily reflect future price trends, complementary information must be considered to drive portfolio generation. The goal of this work is to combine performance indicators with technical and fundamental analyses.
This thesis work focuses on the development of OPTIMDISPLAN (OPTIMized DIversified Stock Portfolio plANner), a Decision Support System that gives the possibility to tune to what extent various fundamental and technical indicators are reliable according to the user, set the risk aversion of the investment and guarantee a payoff increment that does not nullify a pure machine learning primary step to suggest the best stock portfolio with regards to historical price performances.
This study aims to agglomerate different knowledge from business economics, machine learning, risk management and integer optimization to provide a complete decision support system designed to allow tens of combinations that reflect the user market perceptions.
The portfolio generation process relies on a two-step process: first, a set of candidate stock portfolios is generated using an unsupervised, itemset-based model built on top of historical stock prices and a sector-based stock categorization. Secondly, an optimization strategy is used to identify the most appropriate portfolio according to a set of analyst-provided constraints related to stock diversification, fundamentals, and the underlying price trends.
Although a standard setting is supplied and explained together with the software, one of the main well-known problems is related to the huge alternatives that are presented in the literature for indicators and ratios, both in technical and fundamental analyses. How to use them is still nowadays heavily debated and, more specifically, their effectiveness seems to be strongly related to the analyzed markets. Nevertheless, every expert manager always takes into account these indicators, weighting, usually according to a personal taste, which of them is more reliable and which less. From a risk management point of view, it is normal

to see risk measures adapted to choose a safe investment or at least to satisfy the expert risk perception. Experts daily arguing on which ratio or value is better than another, hence I will clarify why I decided to rely on some options rather than others with regards to the desired application. The final panorama of given options assumes 6 fundamental indicators and a customizable moving average. Concerning the risk of investment, its abstraction is undoubtedly a hard task that involves both the capacity of the user to understand his risk aversion and the measures used to quantify it. The key idea is generate first the most promising stock combinations and then optimize a convex combination of performance indicators at the portfolio level. This approach is new because, to the best of our knowledge, previous two-step approaches focus on optimizing the selection of single stocks rather than those of candidate portfolios.

Likewise to an owner's manual, what is the effect produced by some choices is deeply discussed, moreover the optimization value of the approach is statistically investigated and its significance proved. It is also remarkable to note that OPTIMDISPLAN computes numerous measures inspired by European regulations, expert financial environments and academic literature, trying to dispel any doubts that could come up only looking at the provided equity lines or volatility plots. Moreover, when formal statistics are required, the supplied values will play a main role.

The performance of the OPTIMDISPLAN system was validated on real stock data ranging from 2008 to 2020 and the experiments were conducted on the NASDAQ-100 (National Association of Securities Dealers Automated Quotation) stock exchange. The goal of the empirical validation was to test payoff, Value-At-risk, drawdown, and volatility of the proposed strategy compared to the baseline version of the DISPLAN system (based on a pure data-mining approach) and a set of renowned real hedge funds operating on the same market. The results confirmed the strength of the proposed approach even in recent periods, when the outbreak of COVID-19 pandemic has produced significant market oscillations. Specifically, the optimization step has shown to significantly improve portfolio returns and to limit portfolio volatility.

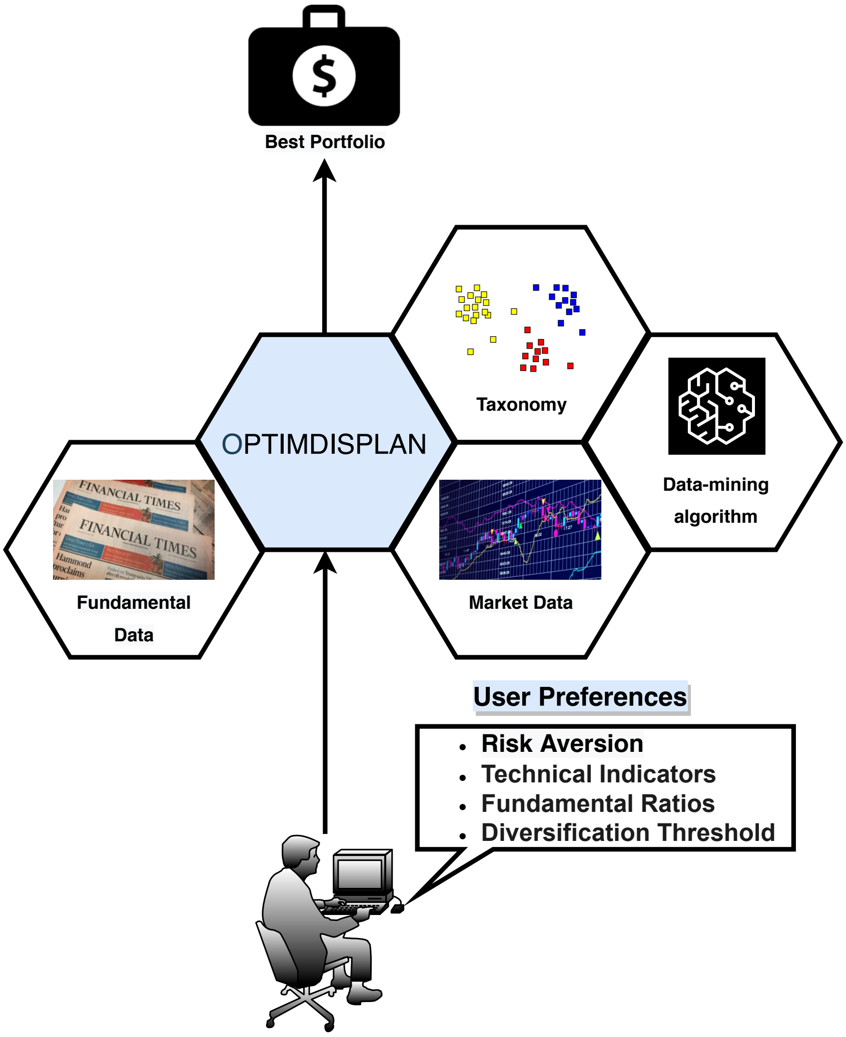


Figure 1.1: OPTIMDISPLAN scheme

Chapter 2

Related works

The study of financial markets is carried on by academical researcher and professional users extensively. Making use of new technology, ideas and years of economical knowledge, everyone wishes to maximize profits and forecast the market future behaviour.

2.1 The state of the art

To make an opening rut, become essential to address this work among different investment typologies. After this clarification, pro and cons of the state of the art that regard OPTIMDISPLAN will be discussed.

2.1.1 The market and the investment philosophy

The main differentiation in investment systems is the trading frequency, indeed it is usually identified the dichotomy of short-term and long-term operations.
Several financial approaches like Althelaya et al. [5] or Kamble [25] may take into account both the paradigms, anyway financial technologies tools are usually focused on a single one. In particular, OPTIMDISPLAN, relying on the fact that the stocks price normally increase (Chen et al. [14]), aims to maximize the profit buying the stocks at a given price and holding them for a long period.

The stock market behaviour is not the same across the world. Trivial examples are emerging markets that have some visible peculiarities (Serra [40]), but differences are more general, indeed all the stocks collection have idiosyncratic features. For example, according to Dellas and Hess [17] some differences could be related to the country financial development.
Thanks to systematic studies such as Nti et al. [35] or Bustos and Pomares-Quimbaya [12], it appears that some of the most common stock markets used to study predictions and forecast movement are from Japan, China and the USA. OPTIMDISPLAN can be placed in this latter market together with other studies that are concentrated in NASDAQ like Ansari et al. [7], Fior et al. [20] or Baralis et al. [8].

2.1.2 The tool features

Once identified where OPTIMDISPLAN acts, it useful to show what its idea is.
Decision support systems, especially in FinTech, are widely spread tools, however different data sources and approaches create a large family full of possibilities. Some examples use the wisdom of financial communities (Gottschlich and Hinz [22]) or text-mining techniques (Al-augby et al. [4]), but fundamental and technical analysis used by our software results as a more conventional approach (Nti et al. [35], Williams and Turton [46]).

One of the innovative features of OPTIMDISPLAN lies in the two-stage approach. Other studies that merge technical and financial analysis like Lam [28] usually use them together, whereas we are going to exploit first a data-mining algorithm and secondly adapt and enhance results according to the user.
The two-step philosophy is a hybrid innovation that has been also recently developed by Yang et al. [48] in the FinTech environment. In this article, it is proposed a first step of stock prediction that, differently from OPTIMDISPLAN, uses extreme learning machine (ELM) techniques and focuses on a fast computing speed in a one by one stock analysis and selection.
In the second step, a stock scoring is realized according to various fundamental factors and a final selection of an equally weighted stocks portfolio is made by quantiles of the distribution of the scores. On the contrary, one of the OPTIMDISPLAN main features is to focus on the whole itemsets in the second step and to use both technicals and fundamentals in a risk-oriented optimization that follow the mean-risk idea of Markowitz [29], where the fundamental scoring is part of one constraint, but subject to a more general optimization procedure.

The portfolio optimization stage of OPTIMDISPLAN, as previously said, it is inspired by Markowitz [29], but with regards to the whole itemset extracted by a primary step rather than single stocks.
The majority of the studies that regard portfolio optimization, approximately follow our mindset of generalization to some extent. Part of the most recent are summarized in Milhomem and Dantas [32] and they usually act on the risk measures (Uryasev [45]) or on the implementation of different heuristic methods rather than exact ones such as dynamic optimization, hybrid algorithms or Bayesian statistics with the addition of some realistic constraints. Nevertheless, as far as we know, no proposals deal with a mean-variance model generalization that deals with portfolios in a top-down approach rather than a bottom-up one that starts from single stocks.
It is remarkable that other fusions of optimization and machine-learning approach are known with the purpose of considers a pre-extraction of a stock subset thanks to a machine learning-based algorithm that is successively computed by optimization models stock by stock to choose the weights of the portfolio. An example is Paiva et al. [36], where the support vector machine (SVM) plays the first step role and the mean-variance model is involved in the second one.

Focusing on the machine learning techniques, the automatic support of financial decision is well explored in literature and, for example, some of these proposals are Chen et al.

[15] where rules on stock market are generated through genetic network programming or Kantardzic et al. [26] that detects remarkable market changes that drives the portfolio thanks to time-series analysis and diversification.
In order to choose a suitable algorithm, it must be observed that the aim of this work is not to forecast the trends of the markets and to predict the signals like Tilakaratne et al. [44] where the intramarket influence is exploited to manage signals, hence unsupervised approaches are preferred.
Thanks to the unsupervised characteristics, it is easier to find meaningful and readable automated results that can address the user investment.
Clustering-based approaches are widely used and some example are Aguiar and Sales [3] that uses overreaction and underreaction market based cluster to build portfolios or Kedia et al. [27] where a k-means technique is exploited to generate an efficient stock selection. Other unsupervised options are, for instance, time series analysis like Fu-lai Chung et al. [21] that are newsworthy solutions, however, what OPTIMDISPLAN will exploit is an innovative weighted itemset mining algorithm presented by Baralis et al. [8] that can be expected to yield good performances and perfect matching with the software requisites. Moreover, it has been chosen for further application yet (Fior et al. [20]).
Since the Data mining algorithm alone is originally called DISPLAN(DIversified Stock Portfolio plANner), I am going to identify this subsequent work as OPTIMDISPLAN, even if it seems similar, the name is completely detached from the pure machine learning algorithm point of view.

2.2 In front and behind the scenes

The pursuit of the parameters that should be implemented in a financial DSS is not a trivial task. numerously well-known studies have been selected to justify the presence of fundamentals, the way how they have been selected among thousands and the relative thresholds (Beneish et al. [10],Fama [18],Piotroski [37],Mohanram [33]).
Also for technicals, several insights result questionably, starting from the effectiveness itself (Menkhoff [31],Hsu et al. [24]). Once clarified this aspect, one of the chosen guidelines is usually provided by the worldwide famous J.J.Murphy book (Murphy [34]).

With regards to the risk management, a trade-off between the widely used portfolio diversification (Markowitz [30]) and the investment results is a discussed matter (Qi et al. [38]) in literature. This and several risk measures are considered in OPTIMIDSIPLAN. Some of the most meaningful alternatives as standard deviations and Value at Risk following Anderson [6] are good examples.

Lastly, tests and performances required further investigations on what makes sense to be compared and analyzed. On this purpose, the study follows the measures proposed by some of the big sharks in the financial world as BlackRock, JPM, Morgan Stanley and Pioneer, involving famous financial measures like the Sharpe ratio (Sharpe [41]) and conforming the software result with the European regulations in long term financial products market (securities and markets authority [39]).

Chapter 3

DISPLAN

This chapter aims to explain which strategy resides in the data-mining algorithm of DISPLAN (DIversified Stock Portfolio plANner) (Baralis et al. [8]) and, by consequence, the first hided step of OPTIMDISPLAN.
A coarse explanation of the chosen tool is supplied to better analyze both what should be the main features of a general data-mining algorithm interfaced with OPTIMDISPLAN and how the computation according to this financial automation affects the decision support system results.

The decision support system requires a first data-mining approach that provides as output a list of ranked portfolios taking the market candlesticks data as input. The second stage is going to exploit these results.
As aforementioned, DISPLAN is based on a weighted itemset algorithm that is briefly depicted in Figure 3.1. This automated portfolio generator becomes the first necessary block of the whole OPTIMDISPLAN, but it must be remarked that it originally managed a diversification too. Since OPTIMDISPLAN will involve this characteristic yet, the automation has been modified and partially reducted.

Roughly speaking, what is expected as the (middle overall) final result of the implemented data-mining algorithm, it is a list of ranked portfolios that, in this case, are processed by means of weighted frequent itemsets that are developed, coherently with the point of view of the entire study, according to a buy-and-hold strategy. Markets stocks are supposed to be bought at a given price to be held for a long time.
Weighted itemset mining discovers patterns from the analyzed data. These pattern are what we call itemsets and they consists of stocks sets of arbitrary size. It follows that each itemsets represent a portfolio possibility.
Historical closing daily prices and diversification play the main roles in choosing the best possibility for long-term investments. Thus, DISPLAN originally started from the weighted itemset mining on the daily closing prices to identify the best historical performing portfolios and then applied diversification thresholds according to a given taxonomy, however, this latter part (pointed up in dashed red in Figure 3.1) has been moved in OPTIMDISPLAN to be better personalizable with the other parameters that will be introduced later.
Regarding the weighting process, it is assigned with regards to the stock relative return compared with a reference price.

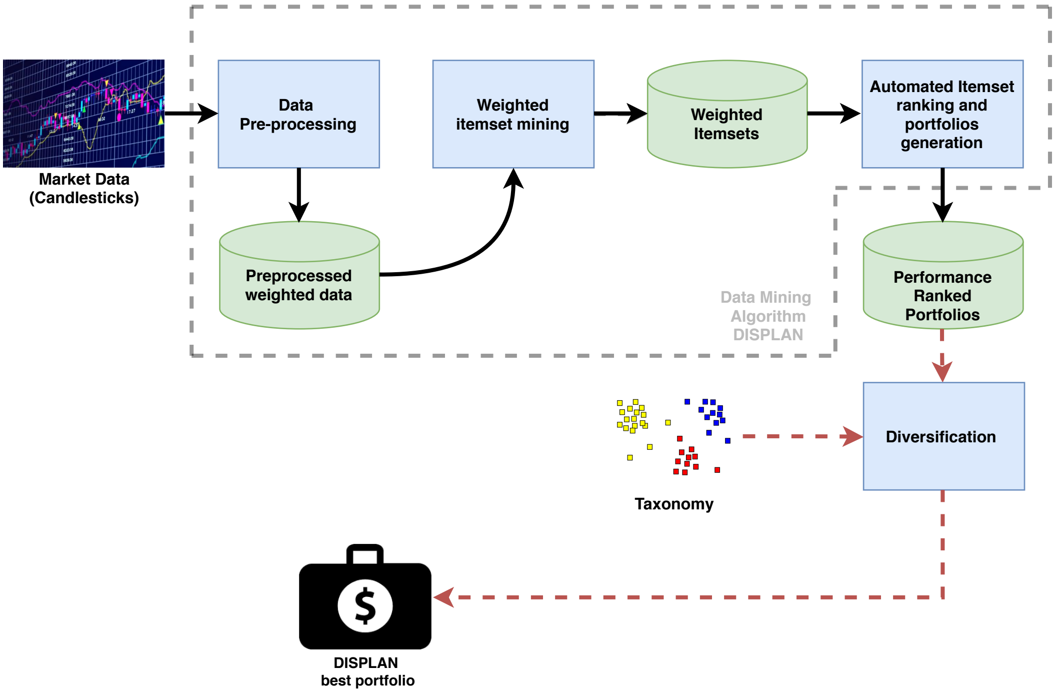

Market Data (Candlesticks)
Data Pre-processing
Weighted itemset mining
Weighted Itemsets
Automated Itemset ranking and portfolios generation
Preprocessed weighted data
Data Mining Algorithm DISPLAN
Performance Ranked Portfolios
Taxonomy
Diversification
DISPLAN best portfolio

Figure 3.1: The DISPLAN framework

3.1 Data processing and Weighted itemset mining

Considering a (training) time interval and a set of stocks (e.g. stocks in a specific collections such as NASDAQ), some particular kind of historical data are processed for each firm to train the algorithm and to successively extract itemsets. Specifically, DISPLAN is general enough to be based on any of the value of financial candlesticks, but because of his common use as a market factor, the Close daily prices have been chosen as principal data.

Weighted stock dataset Let T be the previously mentioned time interval. A discretization is made according to the chosen daily timestamps as $t_k \in t_1, \ldots, t_n$.
Let p_k^j be the price of the stock s_j and r_k^j the relative return of stock s_j defined as

$$r_k^j = \frac{p_k^j - P_j}{P_j} \times 100 \tag{3.1}$$

where P_j is the stock reference price or, in other words, the price at the starting point of the training.
It is defined weighted item i_k the pair $< s_j, r_k^j >$ and a set of weighted items $\{< s_j, r_k^j >\}$ is called weighted transaction tr_k.
The relative return 3.1 in transaction tr_k indicates the percentage profit/loss of an investor who bought the stocks at the reference timestamp P_j.

Weighted itemsets mining Frequent itemsets mining is a well known data-mining approach (Agrawal et al. [2]) to discover useful patterns in the analyzed data.

In our case items are stocks and their sets are the portfolios options.
Since generating all the possible portfolios is computationally infeasible and typically redundant, this kind of algorithm is often driven by a support threshold that represents the frequency of occurrences that a specific portfolio (itemsets) must exceed in the source dataset to be taken into account.
It is first required the introduction of a generalization for the frequent itemsets (Agrawal et al. [2]) algorithm that involves the weight given by the corresponding daily relative returns within each transaction (Cagliero and Garza [13]). On this purpose, it is defined weighted support of an itemset (w-support) the average of its matching weights for all the dataset transactions combining returns by the minimum function on each market-open day.
Let us clarify the above explanation through a simple example.

Time stamp	Weighted stock transaction
t_1	<A,5%>,<B,5%>,<C,-1%>,<D,7%>,<E,5%>
t_2	<A,2%>,<B,6%>,<C,0%>,<D,2%>,<E,2%>
t_3	<A,4%>,<B,5%>,<C,-2%>,<D,4%>,<E,5%>
t_4	<A,4%>,<B,2.5%>,<C,-4%>,<D,10%>,<E,4%>
t_5	<A,1%>,<B,4%>,<C,-2%>,<D,7%>,<E,1%>
t_6	<A,-1%>,<B,6%>,<C,0%>,<D,1%>,<E,-1%>

Table 3.1: Example of weighted stock transactions

The matching weights of itemsets {a,e} in Table 3.1 are: 5 in t_1, 2 in t_2, 4 in t_3, 4 in t_4, 1 in t_5 and -1 in t_6, in fact these are the minimum values of the combined stocks returns. Concerning the weighted support, it will be: $\frac{5+2+4+4+1-1}{6} = 2.5$, hence the average daily profit, on the considered time frame, of the least performing stock with respect to the portfolio is 2.5%.
Following the prior logic, The frequent itemsets extracted from the data source in Table 3.1 are provided in Table 3.2, where a 2.5% threshold is assumed.

Itemsets	w-support
{a,d,e}	2.50
{b,d}	2.75
{a,e}	2.50
{d}	4.83
{b}	4.75
{e}	2.66
{a}	2.50

Table 3.2: Extracted frequent itemsets

3.1.1 Execution time limitations

The itemsets mining process complexity is striclty related to the size $\#S$ of the set of stocks. In particular, fixed a maximum portfolio size l, this procedure has complexity

$$\#S^l \tag{3.2}$$

From a practical point of view, considering an academic hardware system, it will be quite challenging to operate with index as Russell2000 or even S&P500, it follows that our preference for the real analysis went to NASDAQ both for the abovementioned prolematic and for several technical results (Hsu et al. [24]) that will be later further investigated. Future experiments that dela with heuristics could overcome this software limitation.

3.2 Diversification and Taxonomy

One of the most common strategies to reduce the risk of an investment is diversification (Markowitz [30]), however, over-diversification could incur marginal returns.
How portfolios are diversified depends on a given Taxonomy that associates a cluster to each of the stocks, thereupon this concept could be generalized according to a different way to cluster firms.
In the original DISPLAN (Baralis et al. [8]) work, the Yahoo! Finance sector classification was used, but here I am going to reckon on the GICS ([43]) (Global Industry Classification Standard), produced by S&P Global together with MSCI. This choice permits to eventually generalize the sector taxonomy for industries, or even sub-industries. For the sake of completeness, a different way to identify stocks clusters will be discussed in A, following an innovative time series correlation study developed by Fior et al. [20].
Let us now examine in depth what the diversification effect is.
Let I be an arbitrary itemset and T a taxonomy. Let delineate $\#sec(I)$ the number of different clusters present into the portfolio according to a given taxonomy. We define the diversification level as:

$$div(I,T) = \frac{\#sec(I)}{|I|} \tag{3.3}$$

where $|I|$ is the cardinality of the itemset.
DISPLAN will extract itemsets that represents portfolios with:

- w-support equal or higher a selected threshold
- diversification level equal or higher a selected threshold

Coming back to the exampled in Table 3.2, it is possible to assume a Taxonomy according to Table 3.3. If a minimum diversfication threshold of *mindiv*=0.6 is fixed, the suitable portfolios are marked in red in Table 3.4.

Technology	Financial	Real Estate
a,b,e	d	c

Table 3.3: Example Taxonomy

Itemsets	w-support	diversification
{a,d,e}	2.50	0.66
{b,d}	2.75	1
{a,e}	2.50	0.5
{d}	4.83	0
{b}	4.75	0
{e}	2.66	0
{a}	2.50	0

Table 3.4: Extracted frequent itemsets

It is a remarkable fact that a portfolio composed by a single stock has not a diversification level of 100%, but, since it is neither diversified according to taxonomy nor to stocks at all, it will be considered as a 0 diversified one[1].

The portfolio generator will rank itemsets first in order of decreasing length and, secondly, with respect to the w-support value. Of course, this setting satisfies all the threshold requirements in minimum support and diversification and it will choose the portfolio that allows spreading the bet over the largest number of different assets.The example in Table 3.4 would give {a,d,e} as best option.

It should not be overlooked that it is assumed that the investor will uniformly bet over the selected stocks.

3.3 Summary of the Empirical results

A well-structured investigation on the DISPLAN performance is presented in his original article (Baralis et al. [8]). What I am going to report here is the general setting, that will inspire the OPTIMDISPLAN tests, and some insight concerning which have been idealized as default parameters for minimum diversification, minimum support and training period.

Market conditions The stocks sets that have been selected by financial indexes take into account the complexity problem of this algorithm approach 3.1.1, hence the tests are focused on NASDAQ-100 and DowJones-30. They are used also as the main benchmark to compare the performances of the chosen options.

Regarding the time horizon, two opposite situations are deeply analyzed:

- A favourable market condition in 2013 and 2014
- An unfavourable market condition during the economic crisis in 2008 and 2009.

In both the circumstances, the DISPLAN equity lines perform better than the benchmark, even if a fair amount of volatility is perceptible across results. Moreover, also against other

[1] Despite a portfolio with a single stock is far from the goal of OPTIMDISPLAN, when the diversification will be moved on the complete decision support system, it will allow a 0 diversification threshold.

kinds of equity long-term funds or sector-based stocks collections, results are surprisingly up to the mark.
In the next chapter 4, I will show in details some quantitative results and equity lines by using DISPLAN results as OPTIMDISPLAN benchmark themselves, there, this assertions will become touchable.

Default options Hitertow, the effects of the parameters of the system have not been well explained. Ideally, all of them can be chosen by the user, anyway, it is useful to understand how they affect results and which could be sane thresholds. In the next sections, an answer will be given for OPTIMDISPLAN. Concerning DISPLAN, I am not going to looking in details all the experiments proposed, but it must be remarked that:

- Based on the performed analyses, it has been identified a standard configuration that allows to achieve the best trade-off between profit and diversification on the analyzed dataset. Specifically the parameters are:
    - min diversification = 0.7
    - min w-support = 8%
- Regarding the learning period, its size can heavily affect the generated portfolio behaviour. Considering medium-size periods of 6 months appeared to be the best trade-off between model generality and accuracy.
  Since this choice appears to be tough and related to technical insight, I will consider it as the most permanent when the user interface will be presented.

Chapter 4

OPTIMDISPLAN

The decision support systems aim to supports business decision-making activities.
The OPTIMDISPLAN role is placed in financial activities as private investors or financial pundits. It provides as output its best portfolios in a buy-and-hold investment mindset according to some input parameters that can take into account fundamental ratios, technical analysis and historical performances (1.1).
Nowadays, the manual visualization of all the data linked to a possible investment is not easily affordable. Though some user could have complete confidence in an automated system that use only historical market data rather than companies insights, others may argue that often an overall proposals view, according to their peculiar measures of risk, company insights and historical performances in their union, could strongly help to facilitate and optimize a choice.

These features are provided in OPTIMDISPLAN, furthermore it owns a standard configuration that is going to be better explained subsequently. This naive solution is proposed to build a profitable solution also for a not expert user, but it is possible to tune all the software ingredients to best satisfy the environment knowledge and intuitions.
Exploiting a data mining algorithm that produces a ranked list of profitable portfolios according to the historical performances of the stocks by the input candlesticks data as the first stage, the decision support system divides the second stage into several parts resumed in Figure 4.1.
It is identifiable

- An input taxonomy that is used to reduce the risk through diversification. This taxonomy can reflect the company sector, industry, sub-industry or other advanced clusterizations to better analyze the firm properties.
- The fundamentals role that consists in a personalized set of ratios and indicator that are usually extracted from the balance sheet, the cash flow or the income statement.
- The technical analysis block and the final optimization that combine risk and payoffs to better manage unexpected outcomes.

The fundamental inputs are various financial statements of the top firms that are rearranged to compute some well-known quantities that try to represent the company stability, investments and management. Here, only the top firms are taken into account because these quantities are usually not easy comparable if the companies do not share some characteristic that aggregates them.

Concerning the technical analysis, the possible families of computable indicators per portfolio are huge, hence OPTIMDISPLAN relies on a few of them with regards to the investment mindset. Anyway, the computation requires the same data used by the data-mining algorithm, thus market data with the daily granularity are sufficient.

Fundamentals, technical indicators and diversification together filter out the ranked best portfolios that do not satisfy the requirement on their customizable thresholds and optimization re-orders the remained options to further minimize the historical risk behaviour of the suggested options, avoiding an exposed pure pay-off approach.

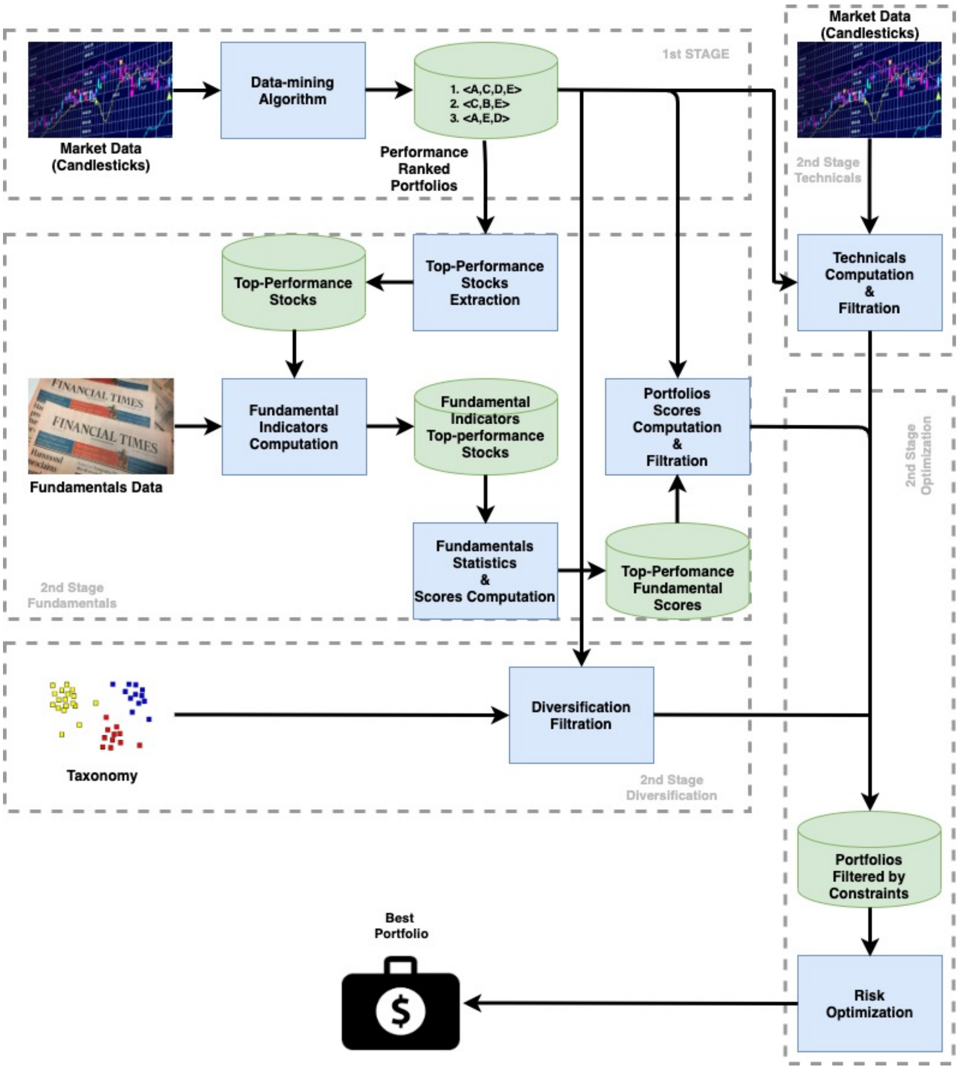

1st STAGE
Market Data (Candlesticks)
Data-mining Algorithm
1. <A,C,D,E>
2. <C,B,E>
3. <A,E,D>
Performance Ranked Portfolios
Market Data (Candlesticks)
2nd Stage Technicals
Technicals Computation & Filtration
Top-Performance Stocks
Top-Performance Stocks Extraction
Fundamental Indicators Computation
Fundamental Indicators Top-performance Stocks
Portfolios Scores Computation & Filtration
Fundamentals Data
Fundamentals Statistics & Scores Computation
Top-Perfomance Fundamental Scores
2nd Stage Fundamentals
2nd Stage Optimization
Taxonomy
Diversification Filtration
2nd Stage Diversification
Portfolios Filtered by Constraints
Best Portfolio
Risk Optimization

Figure 4.1: The OPTIMDISPLAN framework

4.1 The model skeleton

Let us define the following indexes:

$$j : \{\text{index of the itemsets extracted by DISPLAN}\} \tag{4.1}$$

The related stocks portfolios will be pointed out as:

$$S_j : \{\text{Set of stocks in itemset } j\} \tag{4.2}$$

And an integer variable $\delta_j \in \{0,1\}$ will be associated with each portfolio.
Starting from the objective function, two different measures must be taken into account for each portfolio set of stocks. More specifically, the first one $\psi(S_j)$ will be related to the pay-off, but it is remarkable that, rather than pure pay-off, this part depends on DISPLAN without the diversification, thus it will allow for the portfolio length and the w-support. On the contrary, the second measure $\phi(S_j)$ is centred on the risk and several possibilities like Value At Risk or standard deviation could be considered, but we deeply discuss it later.
On the other hand, constraints will be divided in:

- Fundamentals requirement
- Technical requirement
- Diversification requirement
- Single portfolio choice

Even if other personalizable parameters that act like filters, such as the minimum support, are hidden in DISPLAN.
Willing to provide a general explanation of how these constraints should be built, it is possible to introduce for each itemset extracted by DISPLAN:

- A threshold in fundamentals **threshold**$(S_1, \ldots, S_n)$ that aims to transform what the financial statements supplies for each firm in a more general and comparable view, taking into account a global computation of several ratios, using scores that will be indicated as **Fund**(S_j).
- A technical signal **Tech**(S_j) that summarize the trend of the portfolio
- A diversification threshold **Div**(S_j) that depends on a given taxonomy

Summing up, a first model could be sketched as:

$$\begin{aligned} \min_{x} \quad & \delta_j\{\lambda\psi(S_j) + (1-\lambda)\phi(S_j)\} \\ \textbf{s.t.} \quad & \text{Fund}(S_j) \geq \text{threshold}(S_1, \ldots, S_n) \quad \forall j. \\ & \text{Tech}(S_j) = \text{TRUE} \quad \forall j. \\ & \text{Div}(S_j) \geq \mathit{mindiv} \quad \forall j. \\ & \sum_j \delta_j = 1 \end{aligned} \tag{4.3}$$

Looking at the model 4.3, some questions suddenly come up. Let us better investigate what it is the meaning and the logic of this formulation.

Risk aversion The first parameters that have been not explained yet is λ. This value aims to abstract the risk aversion of the user and it is supposed to vary continuously across 0 and 1 to make a convex combination of pay-off and risk.
It follows that when $\lambda = 0$ the selected risk measure will be not taken into account and OPTIMDISPLAN will only filter DISPLAN result through the chosen constraints. On the contrary, a full risk aversion turns out to be driven by the pay-off anyway. This is because the itemsets across we are investigating in this second step are filtered according to the minimum support, that is fundamental following the idea of the frequent itemset algorithm (Agrawal et al. [2]).

Fundamental threshold Dependencies with regards to all the portfolios have been highlighted in the fundamentals constraint. This is due to the inapplicability of a myopic approach that does not consider what the indicators panorama is.
Differently from technical indicators, that may provide entry and exit points, trends and insight just looking at the stock price pattern, when fundamentals are analyzed, relativity becomes the main actor. Moreover, the comparisons should consider a similar kind of players. As a matter of fact, it is possible to figure out the power of a fundamental value only if it is known where it should range.
The majority of the studies, that try to use a Fundamental-based market strategy, never think about top values, rather they usually are focused on identifying:

- which firm has not a solid foundation to manage eventual market stress (Piotroski [37],Mohanram [33])
- who are misrepresenting results to be more endearing to investors (Beneish [9])
- who is doomed to be a torpedo [1] asset (Beneish et al. [10])

Technical constraint Moving on technicals, it is noticeable that the constraint returns a boolean. This setting has been chosen because what will be analyzed is the trend.
Since it has been set a buy and hold approach, entry and exit points are not coherent because it is supposed that the wish of the user is not to wait for a particular point as the intra-day market but to buy in a specific moment to build his investment that will end after a fixed horizon.
About the vastity of possible indicators, the chosen strategy follows firstly the results of which indicators are hopefully effective in our market (Hsu et al. [24]) and, secondly, a simplicity that avoids an amiss discussion to figure out to what extent complexity in technicals could repay the effort.

4.2 Fundamentals requirement

Fundamental analysis is usually done spanning in clustered contexts and is widely used, for long-term insights, by fund managers in pretty much every market (Menkhoff [31]). This

[1] It is defined *Torpedo* a stock that is fast decreasing in price.

approach is due to a comparability of the values that are computed. It is enough to think about how different ROA or ROE could be when different market sectors are considered. Numerous ways can be taken to diversify firms, some examples are the GICS sectors (or industries) ([43]) or the Book-To-Market value (FAMA and FRENCH [19]). The solution here adopted will take advantage of the structure that has been step-wise built. Pointedly, following the idea of Beneish et al. [10], the cluster is made by the extreme performer, or, looking at DISPLAN, by who is top-ranked by the frequent itemset algorithm. In Figure 4.1, the 2nd stage in Fundamentals well remarks this approach.

4.2.1 Indicators

It is now possible to move on which indicators will be used. I will list and analyze each of them, but the tool aims to furnish a plethora of justified choices, then is a duty of the user trust or not all of them. It must be said that the standard configuration of OPTIMDISPLAN is meant to use the full package.

Over a validation period of 12-months, Beneish et al. [10] shows that extreme losers and extreme winners share many common traits and these common traits make it difficult to isolate torpedo stocks from rockets [2], hence he exploits the two-stage approach that it is here modified according to DISPLAN.

Once clustered the firms, he shows that SGI, GMG, R&D, CHGEPS, ACCRUAL and CAPX values are statistically significant in forecasting the performances.

SGI The acronym SGI stands for rate of sales growth over the past year, and, assuming t the current time, it is analytically computed as:

$$\text{SGI} = \frac{\text{Sales}_t}{\text{Sales}_{t-1}} \tag{4.4}$$

where it can be considered the time interval $[t-1, t]$ either as annual or quarterly. This indicator is mentioned in Beneish [9], where is shown that earnings manipulator tends to have an higher SGI.

GMG The GMG is related to the gross margin and is defined as follows:

$$\text{GMG} = \Delta\text{Sales} - \Delta\text{Gross Margin} \tag{4.5}$$

Where

$$\Delta X = \frac{X_t - {}^{(X_{t-1} + X_{t-2})}\!/_{2}}{{}^{(X_{t-1} + X_{t-2})}\!/_{2}} \tag{4.6}$$

According to Abarbanell and Bushee [1] it is possible to detect an indication of deteriorating earning quality.

[2] In the financial slang, a rocket stock is who makes really positive earnings

CHGEPS Bernard and Thomas [11] exhibits that the measurement of the earning surprise from the most recent time batch can predict future returns. It is evaluated as:

$$\text{CHGEPS} = \frac{\text{EPS}_{\hat{t}} - \text{EPS}_{\hat{t}-1}}{\text{Price}_{\hat{t}-1}} \tag{4.7}$$

Where the EPS is the earnings per share, Price is the closing value of the stock when the EPS is calculated and $\hat{t}$ has been differently demarcated because the time interspace must here be yearly.

ACCRUAL together with R&D this is a popular indicator that is possible to find in Piotroski [37] and it his effect is explained in Sloan [42], where it is highlighted that firms with more positive accruals earn higher subsequent returns. It is calculated as the total accruals scaled by average total assets, thus:

$$\text{ACCRUAL} = \frac{\text{Total accruals}}{\text{Average total asset}} \tag{4.8}$$

CAPX This variable, introduced by Beneish et al. [10], is measured as total capital expenditures divided by average total assets and results to be higher in upcoming extreme winners.

R&D Lastly, research and development, that will be deflated by total assets, is a well-known indicator of how much a company is investing. Unfortunately, this data are often not made public, hence it will be set 0 if missing.

The tool Sketching the final fundamental tool feature, it is possible to imagine something like:

- ☑ SGI
- ☑ GMG
- ☒ CHGEPS
- ☒ ACCRUAL
- ☑ CAPX
- ☑ R&D

Let us visualize an output provided by OPTIMDISPLAN:

```
         CHGEPS       SGI       GMG   ACCRUAL      CAPX        RD
WYNN   0.011639  1.088598  0.046090  0.028425  0.122557  0.000000
NFLX   0.008086  1.028516  0.021031       NaN  0.100269  0.028681
ISRG   0.012754  1.207401 -0.059721  0.125356  0.073838  0.015503
NVDA   0.006494  1.192829 -0.028454  0.158954  0.053650  0.051659
AMZN  -0.001355  1.738811  0.245700  0.108712  0.021434  0.034233
```

```
GOOGL        NaN  1.140695  0.014823  0.085354  0.034553  0.024897
HSIC    0.006696  1.140740  0.039051  0.213733  0.024749  0.000000
PEP    -0.005034  1.213843  0.034285  0.126747  0.051837  0.000000
VRTX   -0.015665  1.243283 -0.022436  0.052072 -0.238970  0.191761
EBAY   -0.030492  1.154236  0.000287  0.031274  0.041062  0.011019
AAPL    0.054054  1.545440  0.021371  0.064549  0.115250  0.008189
CTXS    0.002572  1.141822  0.002251  0.089108  0.048374  0.024248
CSCO    0.004298  1.012827 -0.009653  0.061362  0.022620  0.021400
MNST    0.047714  0.997686  0.022722  0.149645  0.173754  0.000000
ILMN   -0.022029  1.154805 -0.025808       NaN       NaN  0.000000
ATVI   -0.062885  4.665626 -1.492966  0.269439 -0.008553  0.047645
INCY   -0.007828  1.457698       NaN  0.005626 -0.106846  0.118384
BIIB    0.001663  1.131861 -0.010250  0.045504  0.071462  0.026573
ALXN    0.062812  1.531298 -0.002182  0.138409 -0.029875  0.046785
GOOG    0.000571  1.140695  0.014823  0.085354  0.034553  0.024897
WDC     0.041991  1.248018 -0.536224  0.236087  0.006739  0.026522
EA           NaN  2.348438 -0.137846  0.128782 -0.054771  0.049806
IDXX    0.002294  1.067938 -0.000264  0.154354  0.036924  0.023881
SWKS    0.007740  1.105427  0.003382  0.141492  0.042470  0.028948
```

The massive presence of 0s in R&D is evident, however, the software has to deal with others missings identified as `NaN`. This is due to the quality of the fundamentals data coming from Yahoo! Finance. The actuated philosophy will be to penalize the uncertainty, thus the value is considered as the lowest with respect to the individual indicator distribution. By consequence, the specific firm will not be placed among the top ranked for that peculiar value. It is clear that bias is generated, but, luckily, these phenomena will happen quite rarely and it is supposed to have a better data quality if commercial implemented.

It is even remarkable that no normalization is needed.

4.2.2 Fundamental Score

Once computed the fundamental indicators for the extreme firms, it is necessary to find a way to gather them per portfolio and successively have a confrontation.

The adopted solution involved a scoring assignment that follows the lines of Mohanram [33] and Piotroski [37]. More specifically, the explicit portfolio dependence remarked into the model 4.3 plays now his role.

Primarily, overall statistics among the fundamentals are calculated and, after that, it is assigned a score for each stock that spans from 0 to the number of activated indicators. This value is build adding 1 point each time one of the stock ratios is in the 'right part' respect to the median.

Let me better explain this concept starting from the 'right part' assertion. What is meant is that not all the indicators are good when the value is high (or low), but some as SGI are preferable to be low, whereas R&D or CHGEPS should be big. Secondly, the median statistic is used because it takes an assumed value and gives to the user a summary of what the general behaviour is.

When each stock has a score, they are added together in each portfolio and averaged by his length.

For the sake of clarity, it is now reported the stocks score computation of the above mentioned example.

	SCORE
WYNN	3
NFLX	4
ISRG	4
NVDA	5
AMZN	1
GOOGL	1
HSIC	3
PEP	2
VRTX	2
EBAY	2
AAPL	2
CTXS	2
CSCO	3
MNST	4
ILMN	1
ATVI	3
INCY	1
BIIB	4
ALXN	4
GOOG	1
WDC	4
EA	3
IDXX	3
SWKS	5

The constraint

Coming again to the model, what I shall now define is the fundamental threshold.
Reporting the setting,

$$\text{Fund}(S_j) \geq \text{threshold}(S_1, \ldots, S_n) \quad \forall j \tag{4.9}$$

the threshold will vary with regards to a user-selected parameter.
In particular, a portfolio is filtered out if it belongs to the negative tail of the scores statistics and how big this tail should be can be decided by the client via quantiles.
The standard setting places *fund_quant*=0.2 drawing inspiration from Mohanram [33]. It follows that only the upper 80th percentile will be taken into account.

4.3 Technical requirement

As previously explained, entry and exit points are not coherent with a buy and hold approach, hence OPTIMDISPLAN will focus on the trend analysis thanks to one of the most famous families of pattern tool, the moving averages.
All the tools used by chartist have the purpose to identify and measure the trend of the market and it is possible to divide trends in up, down and sideways. It must be said that the sideways trend is referred to as a trendless situation and most technical tools perform poorly with these conditions.
According to Murphy [34], the moving average is a trend following device and its purpose is to signal that a new trend has begun or an old one either has ended or has been reversed. Anyway, this tool must not be misunderstood and interpreted as a predictor, it inherits a time lag in the information, thus the trend is shown only after it starts.
The types used by most technicals analyst are the Simple Moving Average (SMA) and the Exponentially Moving Average (EMA). They are both available in OPTIMDISPLAN as a possible constraint, therefore I am going to better explain their features.

SMA The simple moving average is basically an arithmetic mean:

$$\mathrm{SMA}(n) = \frac{p_{t-n} + \cdots + p_t}{n} \tag{4.10}$$

where p_i is the closing price of the market day i, t is the current day, and n is the covered period taken into account.

EMA It is possible to think about moving averages as weighted means. In agreement with this, the SMA will be a constant weighted approach, whereas the EMA gives greater weight to the most recent day's price as $w_i = (1-\alpha)^i$.
α is the smoothing factor that in its simplest version is:

$$\alpha = \frac{2}{n+1} \quad n \geq 1 \tag{4.11}$$

With this value, α it is created an EMA whose weights have the same center of mass of the equivalent n-day SMA.
Analytically the EMA will be:

$$\mathrm{EMA}(n) = \alpha \left[p_t + (1-\alpha)p_{t-1} + (1-\alpha)^2 p_{t-2} + (1-\alpha)^3 p_{t-3} + \cdots \right] \tag{4.12}$$

It is clear that these quantity are suitable for a plethora of generalization, but often simplicity is the key.

4.3.1 The use of One moving average

To generate specific market signals, crossover methods with double or triple moving averages are used, whereas, for trend analysis, only one of them is sufficient (Murphy [34]). When the closing price moves above the MA a positive trend is captured and when it goes below a negative one.

In Figure 4.2, both the moving average options are represented and are applied to an OPTIMDISPLAN simulated portfolio. It is visible that the green curve that represent the exponential moving average depicts heavier the last price movements when, in this example, a little negative trend begins.

The covered time period If the average is too sensitive (it covers a too short period), some of the short term random movement could activate fake signals of the trend, but, if it is too little, the price movements will not be captured at all. It turns primary to choose the right horizon and, on this purpose, stock traders rely heavily on a 50 day (or 10 weeks) moving average (Murphy [34]), thus this has been set as the standard OPTIMDISPLAN value. In point of fact, the user can choose both the kind of MA (SMA or EMA) and the time horizon to taste.

The constraint The model 4.3 allows a boolean value for the technicals, this is the trend variable that asserts a positive tendency (TRUE) or a negative one (FALSE) following the user experience and filtering out the not promising portfolios. The moving average is applied to a portfolio simulation during the training period as if it was built at its begin.

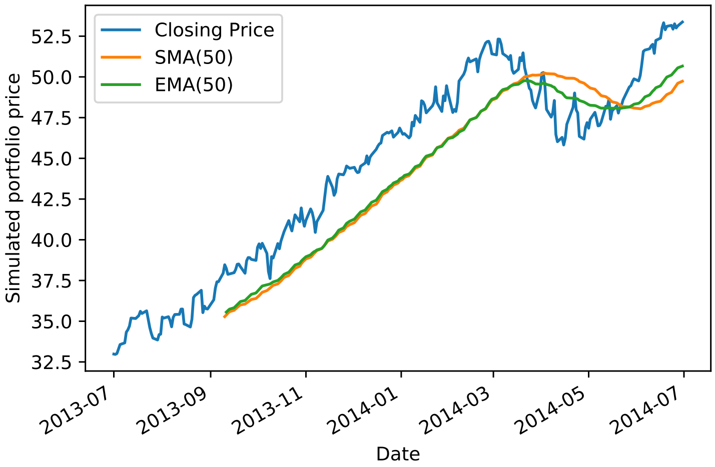


Figure 4.2: Different moving averages in OPTIMDISPLAN

4.4 Measures of the objective function

At this point the remained portfolios satisfy all the required constraints and it follows an example of what OPTIMDISPLAN returns.

```
                    Itemsets  Support Value (%)  diversification  len  trend  fund_score  notTorpedo
PEP_EA_AAPL_ILMN_AMZN_ALXN_ISRG              8.5             True    7   True    2.428571        True
 PEP_EA_WDC_AAPL_AMZN_ALXN_ISRG              8.4             True    7   True    2.857143        True
 PEP_EA_WDC_ILMN_AMZN_ALXN_ISRG              8.4             True    7   True    2.714286        True
     PEP_EA_AAPL_AMZN_IDXX_ALXN              8.6             True    6   True    2.500000        True
     PEP_EA_AAPL_AMZN_IDXX_ISRG              8.6             True    6   True    2.500000        True
```

The objective function in 4.3 considers two different measures that have been combined in a convex way through a risk aversion parameter.
Let me properly discuss them.

4.4.1 The risk measure

Following the steps of Anderson [6], measuring risk is far from easy. Often the best starting point for the estimation of risk is to consider what has happened in the past even if is obvious that the world has changed. It would be foolish to do not pay attention to what has been observed so far.
Lots of risk measures could be provided to the user (e.g. Expected Shortfall, Economic Capital, ...), but the choice fell to two simple possibilities for the client

- The volatility (standard deviation)
- The Value at Risk ($V@R_\alpha$)

It goes without saying that these are neither the best nor the most complex, anyway they are easily understandable and can be used by users with different mathematical backgrounds. Moreover, it is not proven that increasing their complexity a better result is reached.

Value at Risk The standard settings of OPTIMDISPLAN use the V@R as the risk measure. Differently from the standard deviation it considers only the bad tail of the distribution represented by the price value of the stock, hence it does not penalize exceptional earnings. However, it must be said that this is not a certification that future variation will involve only positive jumps, in fact, the instability of a firm in a positive way does not exclude the same behaviour in a negative one.
Looking at this value deeply, with regards to the historical Loss distribution of a stock, it is defined as:

$$\mathrm{V@R}_\alpha = \inf(x : \mathbb{P}(\mathrm{Loss} \geq x) \leq \alpha) \tag{4.13}$$

It is now possible to make different choices to estimate this value. The easier way to do it is through a pure data-driven approach that involves the empirical quantiles, but it could be possible to refine the approximation thanks to a more analytical tool like the extreme theory value (EVT), however, this improvement it is beyond the OPTIMDISPLAN scope.
Let me now try to better explain what is meant as Loss distribution and how OPTIMDISPLAN is going to evaluate the mentioned value.
I am going to take into account the daily variations, these values are computed as the percent change with regards to the previous day. In Figure 4.3, the percent changes time series for the previously shown portfolio(Figure 4.2) is represented and the $\mathrm{V@R}_{.95}$ is selected by the 95th percentile and highlighted in red.

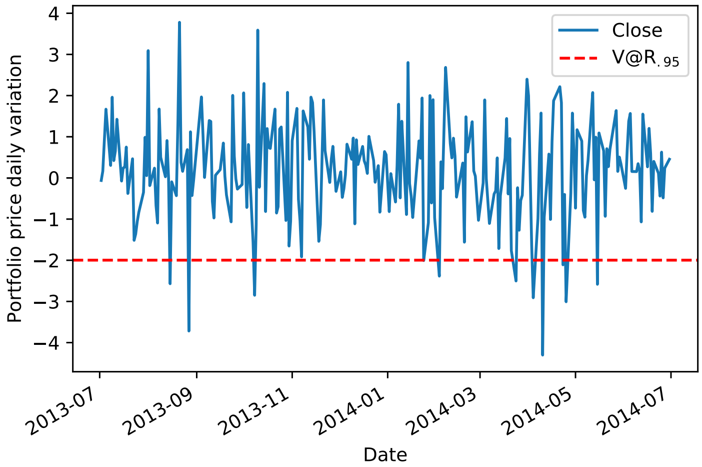


Figure 4.3: An example of the V@R computation

Lastly, I would like to say something about α. This value is critical in V@R, and, although it can be modified by the user, some typical choices are .95 or .99. The standard setting will be $\alpha = .95$ in order to reach a good trade-off between the data availability and the distribution tail cut-off point.

Volatility The second possibility that OPTIMDISPLAN provide is the volatility, that is estimated as the variance on the daily price variations (dV). Analytically:

$$\text{volatility} = \sqrt{\frac{\sum_{t=1}^{n}(\mathrm{dV}_t - \overline{\mathrm{dV}})^2}{n-1}} \tag{4.14}$$

Where $\overline{\mathrm{dV}}$ is the mean empirical estimation of the daily varriation, and n the total number of market days. In Figure 4.4 is represented the same portfolio of Figure 4.3 with the empirical expected value ($\overline{\mathrm{dV}}$) of the percent change in green ± volatility(σ)

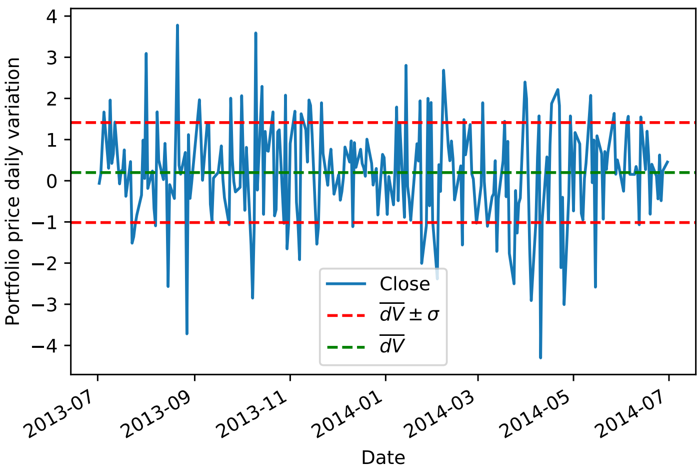


Figure 4.4: An example of the volatility computation

4.4.2 The pay-off measure and the complete objective function

This second part of the objective function is strictly related to DISPLAN. It returns the ordinal position of the analyzed portfolio according to the data-mining rules after that all the constraints have filtered out the unwanted itemsets.

It follows that also the risk measure must be interpreted in an ordinal way rather than in a cardinal one, thus the final risk computation will act ranking the portfolios.

Finally, the risk aversion term will play his role merging risk and pay-off rankings according to the user taste.

A further aspect which deserves attention is the minimization of the objective function. This is because the lower is the portfolio overall rank, the higher it is classified.

4.5 The investment philosophy

Once the stocks set is selected by the software, it must be chosen how to invest.

The stocks collection such as S&P500 and NASDAQ are modified by capitalization weights, but, for the sake of simplicity and to be coherent with the DISPLAN approach, it is supposed to buy the stocks with a uniform assets expenditure. It is likely that future works will revise this approach.

Chapter 5

Test & Performances

The structure of this chapter is divided into several main points.
First of all, it is necessary to introduce what will be used to compare different itemsets to benchmarks, to themselves or to external portfolios, hence the experiments settings are exposed 5.1 and the OPTIMDISPLAN developed measures are explained 5.2.
Secondly, a confrontation with DISPLAN results must be provided to justify this new decision support tool 5.3, furthermore some well-known companies historical results are shown critically 5.4.
After that, the statistical significance of the proposed decision support system will be investigated in 5.5.
It can be useful to show what are the consequence of playing with the parameters that the user can modify and prove that the optimization plays a touchable role, hence we are going to discuss these aspects in 5.6 and 5.7 respectively.
Lastly, some other tests are exposed with regards to the current COVID-19 pandemic 5.8.

5.1 Software and Hardware Design

Mainly, the performed experiments are based on data acquired through Yahoo! Finance `https://finance.yahoo.com` [1].The historical prices structure is composed of the close prices adjusted for splits that will be our analysis linchpin, the adjusted close price adjusted for both dividends and splits and the other typical candlestick values such as open, high and low.
About fundamentals, they are divided into 3 different files:

- Income Statement
- Balance Sheet
- Cash Flow

Each of them contains several breakdowns that I will not list in details. However, those who are involved in OPTIMDISPLAN are schematized hereafter:

[1] Last Access: Mar 2020, COVID historical updated data Aug 2020 (not fundamentals refresh)

- Income Statement
  - Total Revenue
  - Gross Profit
  - Research Development
  - Net Income available to common shareholders
- Balance Sheet
  - Net Receivables
  - Total Assets
- Cash Flow
  - Net cash used for investing activities

Other historical data concerning the risk free treasury bonds (essentials for the Sharpe ratio computation in formula 5.4) come from `https://www.treasury.gov`[2].
The workstation is a hexa-core 2.67 GHz Intel Xeon with 32GB of RAM, running Ubuntu Linux 18.04.4 LTS.
Regarding the used programming languages, the weighted frequent itemsets algorithm is developed in C/C++, the secondary operations that deal with supports, portfolios length and taxonomy are written in Java, the OPTIMDISPLAN core and everything that works with data management to show performances exploit Python. These languages are linked by several script Bash. Lastly, statistical tests are performed by R.

5.2 The performance measures

Looking at the literature or at the leader of large equity portfolios managers, clear guidelines to introduce performances are not well-defined. However, it is possible to identify several points that are pretty much everywhere provided, even if some particular rules have been established for a couple of indicators (securities and markets authority [39]).
Naturally, the primary way to figure out the portfolio behaviour is through plots such as raw time series or volatility plot, but a more quantitative analysis is needed to show analytical results.

5.2.1 Payouts and Losses

The most detected indicators for an investor are surely the final and max payout and the max loss.
Regarding these values, OPTIMDISPLAN automatically supplies:

- Max Payout

[2] Last Access: Apr 2020

- Max Loss
- Payoff yearly
- Payoff after the first month

but the Compound Annual Growth Rates (CAGRs) are usually taken into account in a business environment to compare the historical returns of stocks with bonds or with a savings account. CAGRs are geometric progressing ratios that assume a constant return over a fixed number of periods. In our case, it is assumed the profits would have been reinvested at the end of n months to achieve the final payout. Specifically, it is separately computed for $n = 3,2,1$ (Year, Semester, Trimester) and it is analytically identified as:

$$\text{CAGR}(n) = \left[\sqrt[n]{\frac{\text{Year end Price}}{\text{Initial Price}}} - 1 \right] \cdot 100 \tag{5.1}$$

An example of the above mentioned results is in Table 5.1

	OPTIMDISPLAN	Benchmark	DISPLAN
maxPayout	66.45	36.17	55.45
maxLoss	-0.06	0.0	-1.45
payoffYear	59.75	35.82	51.40
payoffMonth	8.37	7.30	0.76
CAGR_Y	59.75	35.82	51.40
CAGR_S	26.39	16.54	23.04
CAGR_T	12.42	7.95	10.92

Table 5.1: An example of the payoffs measures

5.2.2 Risk and others main indicators

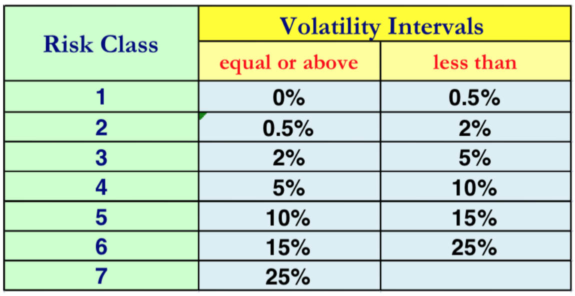
It stands to reasons that the volatility will be given to the user, but some adjustment must be done. In particular, it will be scaled due to the general methodology provided by the securities and markets authority [39].
Let us primarily introduce the Synthetic Risk and Reward Indicator (SRRI).
This value is based on the volatility of the portfolio and it is generated by the historical data of it. More specifically it is computed with a annualized weekly based volatility that, assuming m =number of weeks per period and $T = m \times$ number of periods, depends on the following expression:

$$\text{annualized volatility} = \sigma_m = \sqrt{\frac{m}{T-1}\sum_{t=1}^{T}(r_t - \bar{r})^2} \tag{5.2}$$

Where r_t is the portfolio's return measured over T not overlapping periods of the duration of $\frac{1}{m}$ years and

$$\bar{r} = \frac{1}{T}\sum_{t=1}^{T} r_t \tag{5.3}$$

For example, if 5 years are taken into account, $m = 52$ and $T = 260$. This value is then translated in a class risk in Figure 5.1 and this final result is usually provided in the KIID[3]. Coherently, the same path is followed for the daily verison of the vlatility provided in the final fact-sheet.



Risk Class	Volatility Intervals	
	equal or above	less than
1	0%	0.5%
2	0.5%	2%
3	2%	5%
4	5%	10%
5	10%	15%
6	15%	25%
7	25%	

Figure 5.1: European Regulation SRRI

The second risk measure will be the $V@R_\alpha$ defined in 4.13 , it is considered with $\alpha = .95$, but, reproducing the volatility approach, both daily and weekly is calculated. Summing up what has been done so far, an OPTIMDISPLAN example of these common risk measures is sketched in Table 5.2.

	OPTIMDISPLAN	Benchmark	DISPLAN
volatilityDB	19.85	13.09	26.30
SRRI	6	-	6
$V@R_{.95}^{\text{LossDaily}}$	1.73	1.28	2.59
$V@R_{.95}^{\text{LossWeekly}}$	3.87	2.06	4.20

Table 5.2: An example of the risk measures

Sharpe Ratio One of the most used ratio to understand the profitability of a long term portfolio or fund has been created by Sharpe (Sharpe [41]) that won the 1990 Nobel Prize in Economic Sciences. It characterizes how well the return of an asset compensates the investor for the risk taken and, specifically, it adjusts the risk comparing a risk-free return to the asset return. Usually, a treasury bill is supposed to be a risk-free investment[4].
In formula it is :

$$S = \frac{\mathbb{E}[r - r_f]}{\sqrt{\text{Var}(r)}} \tag{5.4}$$

where r is the portfolio's return distribution, whereas r_f is the risk-free one.

[3] "The KIID is a two-page 'fact-sheet' style document which includes the critical information about a fund. The document aims to help investors understand the nature and key risks of the fund in order to make a more informed investment decision."(2020 BlackRock©, Inc.)

[4] OPTIMDISPLAN will use the U.S.A treasury bills returns (T-Bills)

Active Returns Taking a cue from the Sharpe ratio, if the risk-free return r_f is substituted with the benchmark return r_b, it is possible to define the *simple active return*:

$$\text{Active return} = \mathbb{E}[r - r_b] \tag{5.5}$$

It follows that it is possible to divide by their difference volatility, that is identified as *Tracking Error* (TE):

$$\sigma_{\Delta r} = \text{TE} = \sqrt{\text{Var}(r - r_b)} \tag{5.6}$$

This value shows how close is the behaviour of the portfolio with regard to the benchmark, thus low average returns with an high TE means that something is wrong with the investment.
Computing the previously announced ratio we build the *Information Ratio* (IR)

$$IR = \frac{\mathbb{E}[r - r_b]}{\text{TE}} \tag{5.7}$$

That says us how good are is the expected portfolio return if it is scaled by the volatility.

Beta The user would likely like to investigate how the portfolio changes when the benchmark varies and an indicator that is given in pretty much every KIID is the *Beta* (β). The mathematical interpretation is based on this linear regression:

$$r_t = \alpha + \beta r_{b,t} + \epsilon_t \approx \alpha + \beta r_{b,t} \tag{5.8}$$

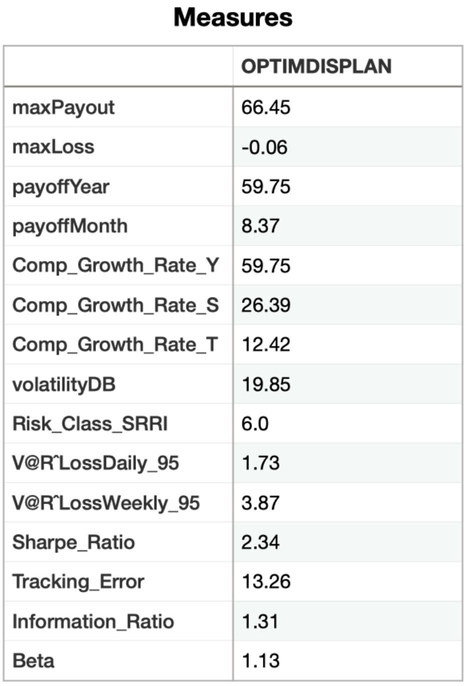
where ϵ_t is the Gaussian assumed error of the linear regression and t is the daily time indicator.
The rough key to understanding this value is that for each 1% earned by the benchmark, the portfolio will earn a β%.
The common way to compute it exploits the empirical covariance as:

$$\beta = \frac{\text{Cov}(r, r_b)}{\text{Var}(r_b)} \tag{5.9}$$

A full example of the measures supplied by OPTIMDISPLAN is shown in Figure 5.2

5.3 DISPLAN comparisons & Benchmark

To analyze the OPTIMDISPLAN performance is necessary to choose a benchmark. Since the used stock set will be mainly NASDAQ-100, this will be adopted as the primary touchstone.
Returning to DISPLAN (Baralis et al. [8]), its results are deeply investigated in unfavourable market conditions and then in favourable one. Thus I am going to show the OPTIMDISPLAN results in the same periods.
It must be remarked that the training time horizon will coherently follow the data-mining settings previously exposed 3, whereas, concerning the validation, it will be considered one year.
This latter choice is linked to the fact that usually a portfolio is updated periodically, hence



Measures

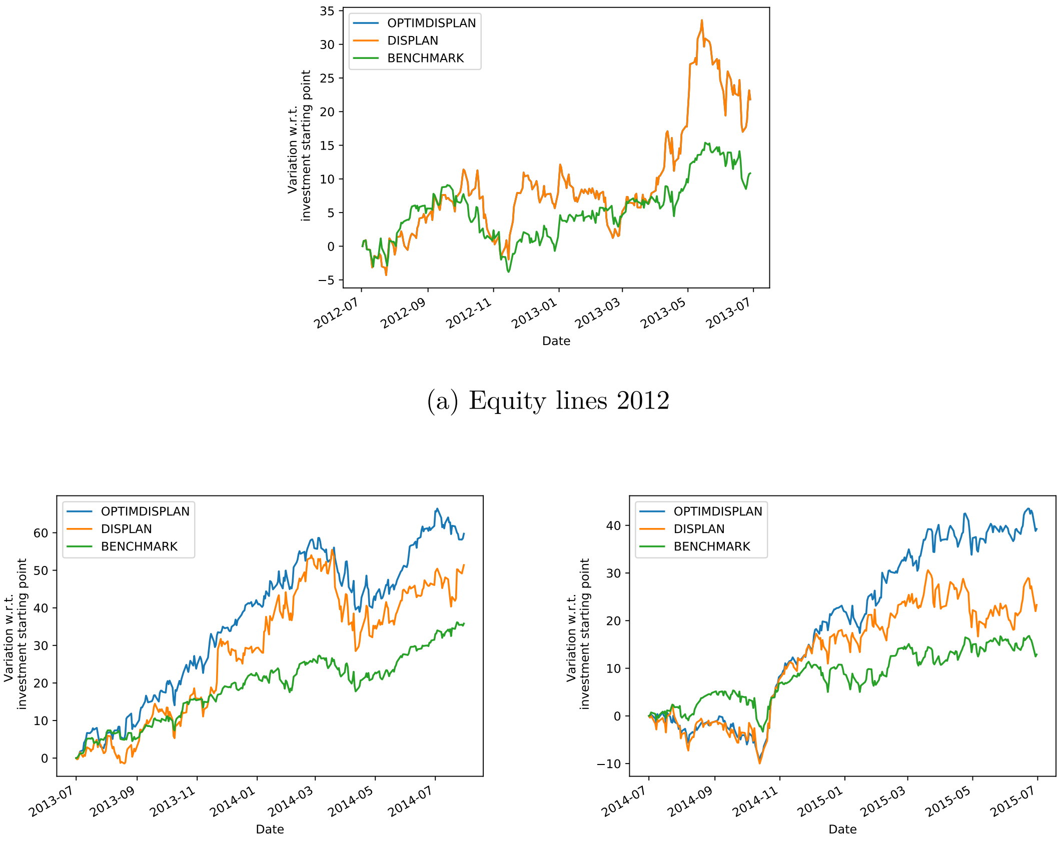
	OPTIMDISPLAN
maxPayout	66.45
maxLoss	-0.06
payoffYear	59.75
payoffMonth	8.37
Comp_Growth_Rate_Y	59.75
Comp_Growth_Rate_S	26.39
Comp_Growth_Rate_T	12.42
volatilityDB	19.85
Risk_Class_SRRI	6.0
V@RˆLossDaily_95	1.73
V@RˆLossWeekly_95	3.87
Sharpe_Ratio	2.34
Tracking_Error	13.26
Information_Ratio	1.31
Beta	1.13

Figure 5.2: A full example of the OPTIMDISPLAN validation measures

it makes sense to validate in one year and re-train the algorithm during its in order to reinvest with a fresh point of view.
The general methods that these tests will follow are based on a triple juxtaposition (OPTIMDISPLAN, DISPLAN, Benchmark) both in performances and graphically equity lines.

5.3.1 Favorable market conditions

One of the most favourable scenarios where the tool can be tested is the 2013. The OPTIMDISPLAN standard configuration with a risk aversion of 0.5 and the published parameters for DISPLAN are used. For the sake of completeness, I am going to supply also the previous and successively results of 2012 and 2014.
In Figure 5.3 are represented the average percentage variation above discussed. What we notice is that in 2012 the OPTIMDISPLAN choice coincides with the DISPLAN portfolio, but the performances w.r.t the benchmark are heavily better.
Concerning 2013 and 2014, OPTIMDISPLAN outperforms both DISPLAN and benchmark, but the main point is that in 2013 the volatility is visibly minor.
Quantitively, wanting to better understand to what extent the performances have been modified, the volatility plot for both the options are shown in Figure 5.4. The values assumed by OPTIMDISPLAN are comparable with the benchmark that, being a well-established collection, has relatively low volatility.
The Tables 5.1 and 5.2, previously used as examples, report a numerical interpretation of what is visible in the plots.
The variation of the standard parameters of OPTIMDISPLAN will generate different



(a) Equity lines 2012

(b) Equity lines 2013

(c) Equity lines 2014

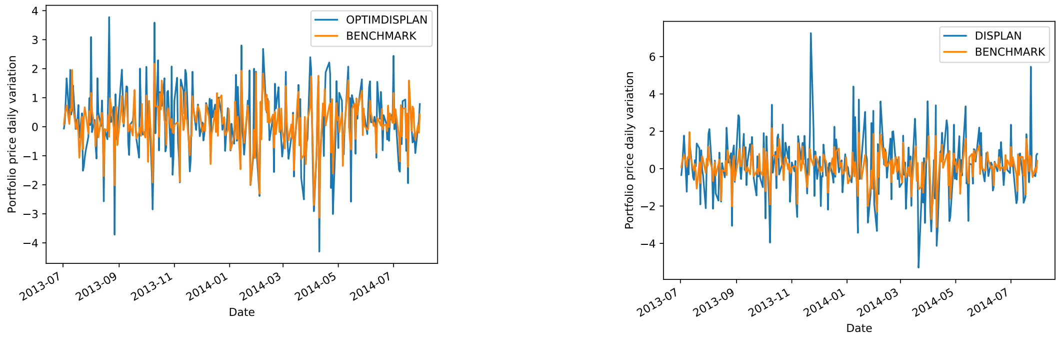
Figure 5.3: Average percentage variation w.r.t. 2012-2013-2014. Comparison with DISPLAN and NASDAQ-100

portfolios, but I will better investigate the effect in the following sections.

5.3.2 Unfavorable market conditions

Let us now proceed with the global financial crisis of 2008. The first attempt will take into account a risk aversion of $\lambda = 0.5$ as previously done. Plotting the average percentage variations in Figure 5.5, although good results are reached for 2009 with the same portfolio of DISPLAN, it is not possible to visibly assert the same for 2008 with regards to the basic software, even if the benchmark is often outperformed by both.
It is anyway remarkable that, as expected, both the volatility and the V@R are lower for the OPTIMDISPLAN tool (Table 5.3) and, regarding the volatility, it is also lower than the benchmark itself.

As I am successively going to show, large equity portfolios based on a buy and hold strategy have heavily suffered the 2008 crisis in the asset management, hence these values



(a) OPTIMDISPLAN portoflio volatility plot 2013

(b) DISPLAN portoflio volatility plot 2013

Figure 5.4: Volatility plots year 2013

	OPTIMDISPLAN	Benchmark	DISPLAN
volatilityDB	43.20	44.94	47.59
SRRI	7	-	7
$\text{V@R}_{.95}^{\text{LossDaily}}$	4.30	4.59	4.73
$\text{V@R}_{.95}^{\text{LossWeekly}}$	10.06	8.23	10.48

Table 5.3: Risk measures of 2008 validation

are far from slightly improvement of the benchmark.

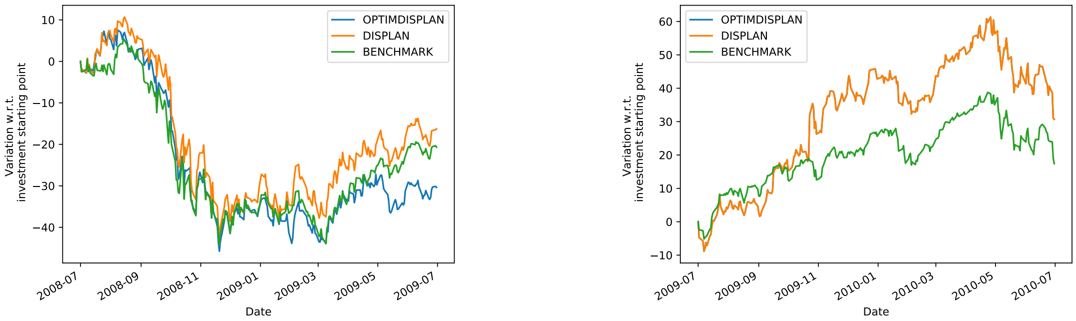
When the tool is exploited, the user is supposed to be well conscious that the 2008 market behaviour was drastically negative so far. As an example, the risk aversion could be set higher (Let us suppose $\lambda = 0.8$) and, due to the global negative correlation of the trend, a further idea could be to avoid technicals indicators. With these options, the results (Figure 5.6) are enhanced and, also in 2009, OPTIMDISPLAN has a better performance than before.

5.4 Performance comparison with established U.S. hedge funds

The financial domain where a buy and hold strategy in collections like NASDAQ or S&P500 holds is the funds market or the large equity cap portfolios.

The number of possible examples is incredibly huge, but for the sake of clarity, it makes no sense to analyze as many of them as possible, hence, I have selected a blend made of three of the most famous, powerful and reliable financial companies[5]:

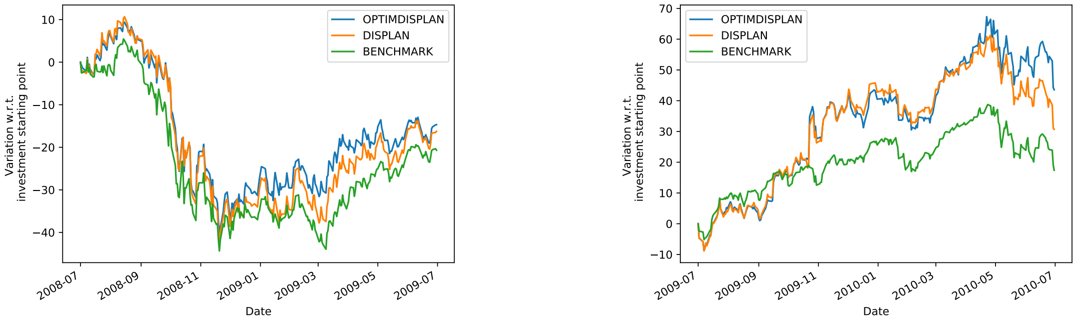
[5] Another critical fact is related to the data availability. Since these companies are well-known, it is possible to find their historical data by Yahoo! Finance



(a) Equity lines 2008 $\lambda = 0.5$

(b) Equity lines 2009 $\lambda = 0.5$

Figure 5.5: Average percentage variation w.r.t. 2008-2009. Comparison with DISPLAN and NASDAQ-100. $\lambda = 0.5$



(a) Equity lines 2008 $\lambda = 0.8$, no MA

(b) Equity lines 2009 $\lambda = 0.8$, no MA

Figure 5.6: Average percentage variation w.r.t. 2008-2009. Comparison with DISPLAN and NASDAQ-100. $\lambda = 0.8$ and no Moving Averages

- JP Morgan Chase & Co
- Morgan Stanley
- Amundi Pioneer

For these three firms, I have selected two portfolios *Large Cap Growth* and one *Large Blend* , respectivey:

- MSEGX-Morgan Stanley Inst Growth A
- OLGAX-JPMorgan Large Cap Growth A
- PIODX-Pioneer Fund Class A

Let us assume that OPTIMDISPLAN's portfolio is managed once a year according to the standard settings ($\lambda = 0.5$) and let us compare what would have been its performance during a five years time horizon.

The average percentage variation is represented in Figure 5.7. It would have yielded more

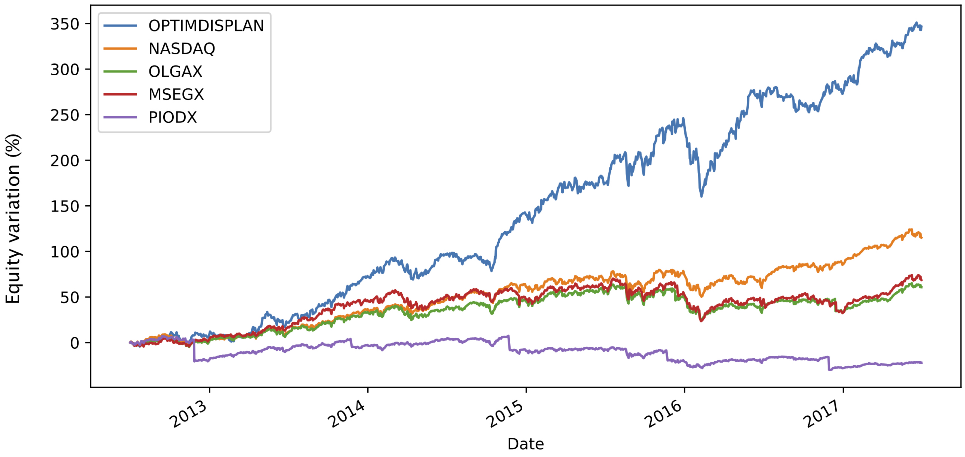


Figure 5.7: Average percentage variation 2012-2017. Comparison with Benchmark and Financial companies. $\lambda = 0.5$

than three times the original investment.

Concerning the chosen benchmarks, the volatility of our system is quite higher, but it is repaid by payoffs.

The reader could wonder which are the financial companies results with regards to OPTIMDISPLAN during an unfavourable market condition like the 2008 crisis, hence, in Figure 5.8, it is shown the results of a portfolio bought in 2008 and managed for two years according to the previous discussion 5.3.2.

It is not a secret that the performances are quite good respect to the benchmark and positive compared to other financial companies.

5.5 Validation of Statistical Significance of the performance improvements

One of the usual question in financial management is whether the investment returns are luckily episodes or they really involve skills and innovations. Academically, this problem matters and some discussed example are Heyman et al. [23] or Cuthbertson et al. [16] for the UK funds market.

It turns out necessary to better investigate the OPTIMDISPLAN performance with statistical tests and distribution analysis.

The first question regards the quantitative measures that are going to be used. On this purpose, although several possibilities have been introduced during this chapter, I

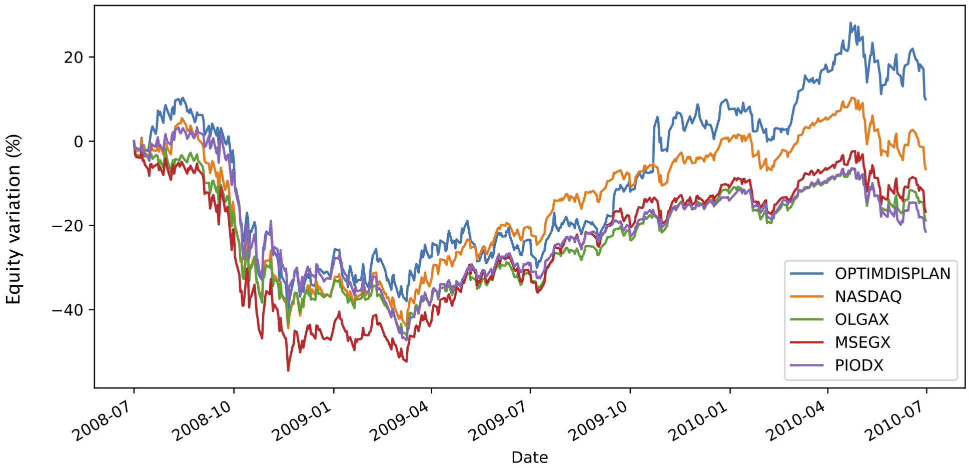


Figure 5.8: Average percentage variation 2008-2010. Comparison with Benchmark and Financial companies. $\lambda = 0.8$

opted for the yearly returns and the V@R$_{.95}$. These two choices depend firstly on the easy comprehensibility, but it is a matter of fact that Gaussian behaviours are well appreciated in statistical tests and yearly returns are the most suitable statistics to get close to a normal distribution.

Experiment design Concerning the experiment design, I decided to test OPTIMDISPLAN with its standard configuration varying the time horizon from 2008 to 2016. More specifically, the portfolios are built every 45 days, producing 63 portfolios that are validated in one year starting with an equally 45-days translated period.
It must be remarked that sometimes no portfolios are suggested, this is due to the lack of positive trend portfolios in a certain period or the w-support excessively high. Ultimately, 52 portfolios and relative time horizons are investigated.

Experiment results In Figure 5.10, on one side it is represented a triple histogram that delineates the not normalized empirical distributions of the yearly returns of DISPLAN, OPTIMDISPLAN and NASDAQ, on the other side a similar image regards the V@R$_{.95}$. Furthermore, the sample means are shown as vertical lines.
It is easily noticeable that the payoffs mean is higher than the benchmark both in OPTIMDISPLAN and DISPLAN, but OPTIMDISPLAN overperforms the pure data-mining approach in the yearly returns and also with regards to the Value at Risk, indeed this value is also lower on average.
Moving on statistical tests, a preliminary step concerns a goodness-of-fit Shapiro test for the returns distribution normality. The results are summarize in Table 5.4, thus, fixed a typical choice for the Type I error threshold $\alpha = 0.05$ we cannot reject the null hypothesis

that the data come from a Gaussian distribution, moreover a Q-Q plot is provided to better realize to what extent our data can be considered normally distributed in Figure 5.9.
Now I am going to compute the Kolmogorov-Smirnov two-sided test to assert a significant distance between the empirical returns distributions of the benchmark (NASDAQ) and our software. The result in Table 5.4 allows to reject the null hypothesis that the two distributions are the same, hence the OPTIMDISPLAN portfolios are distributed differently with regards to NASDAQ.
Lastly, focusing on distribution means, a t and a wilcoxon test results are provided in Table 5.4 and they both reject the null hypothesis too, thence the means differs significantly.
Summing up the previous pieces of evidence, what comes up is that returns are materially improved, furthermore the risk, compared to DISPLAN, has been reducted.

	Statistic	p_value
Shapiro for returns	0.964	0.120
Shapiro for NASDAQ	0.960	0.0795
K-S NASDAQ vs OPTIMDISPLAN	0.326	0.007
t-test NASDAQ vs OPTIMDISPLAN	2.291	0.023
Wilcoxon NASDAQ vs OPTIMDISPLAN	386.0	0.005

Table 5.4: Test results for the returns distribution

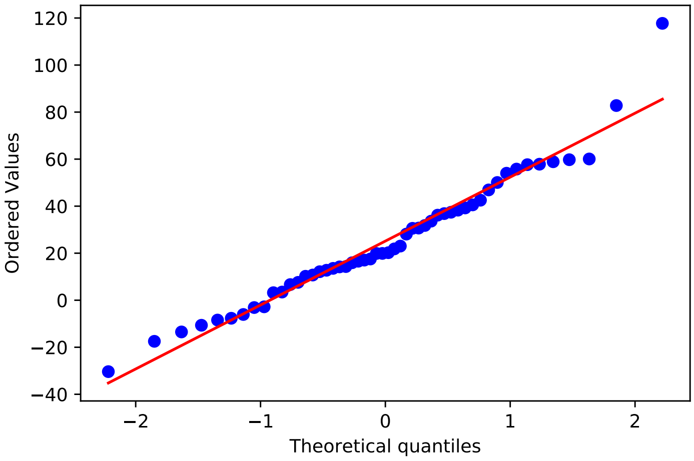
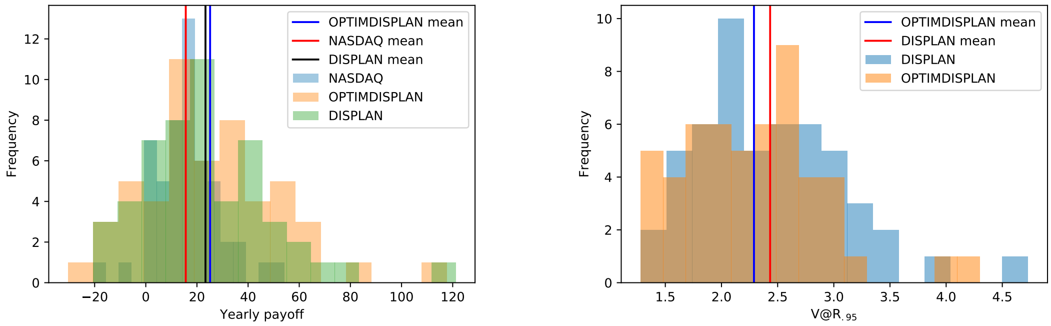


Figure 5.9: Q-Q normality plot of the yearly returns distribution of OPTIMDISPLAN

5.6 Parameter analysis

The examined financial tool deals with a substantial quantity of parameters. It is normal to wonder what is the effect of them, thus I am now going to analyze the alteration of some



(a) Yearly payoff

(b) V@R$_{.95}$

Figure 5.10: Yearly payoff and V@R$_{.95}$ statistical investigation

of the previously mentioned measures varying only one parameter, w.r.t. the standard conditions per experiment family.

5.6.1 Risk aversion effect

The first value that will be analyzed is risk aversion. More specifically, the aim is to investigate on 5 different possibilities:

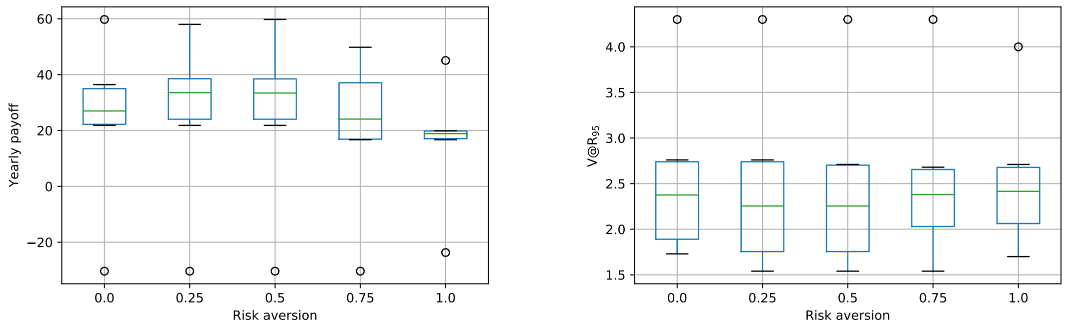
$$\lambda = [0{,}0.25{,}0.5{,}0.75{,}1]$$

uniformly selected with regards to the V@R risk measure. On this purpose, I will consider 6 different years representing both the payoff at the end of the validation period and the daily based Value at Risk.
In Figure 5.12, the results identify a decreasing of the risk and a payoff peak for 2008 (unfavourable market condition), whereas, when the trend is bullish(2013), the stock oscillation does not well reflect what was expected, but no losses are involved in this case.
Looking at Figure 5.11, the 2008 crisis is pointwise identified as extreme, whereas it is advisable a risk aversion value of 0.5 for higher payoffs overall.
From a general point of view, the payoffs go down when the risk aversion goes up, but the Value at Risk seems to be less effective when the market is bullish.

5.6.2 Moving average interval effect

Changing the subject to the moving averages, let us now study what could happen to vary the interval that this indicator uses for its computations.
As mentioned by Murphy [34], usually, to investigate trends, a value of 50 days is preferred.
What I will now test concerns in 4 different options (with $\lambda = 0.5$) that are well-known by technical analysts[6].

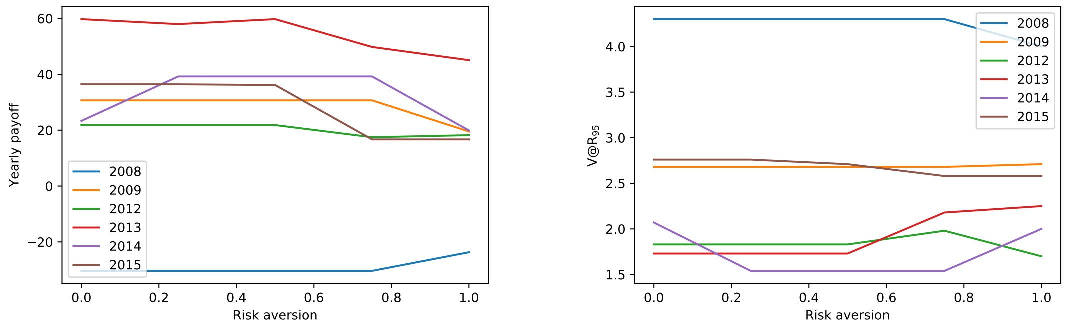
[6] Since OPTIMDISPLAN uses 6 months for training, the higher possible value is 125 (25 weeks)



(a) Yearly payoff

(b) V@R$_{.95}$

Figure 5.11: Yearly payoff and V@R$_{.95}$ Box-plots w.r.t. λ



(a) Yearly payoff

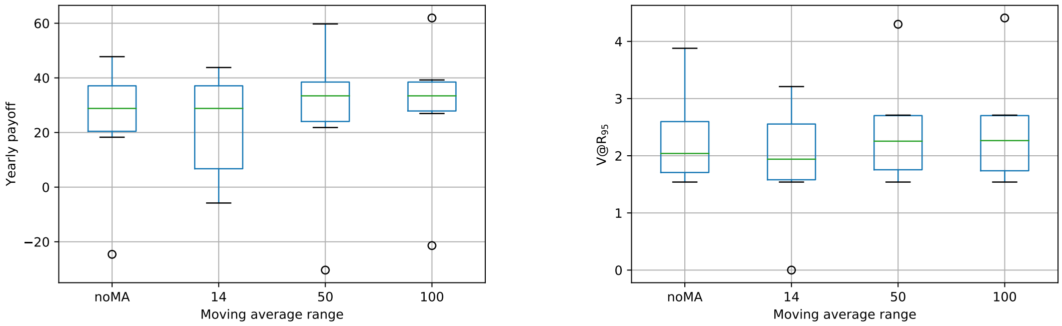
(b) V@R$_{.95}$

Figure 5.12: Yearly payoff and V@R$_{.95}$ w.r.t. λ

- no moving averages
- 14 days
- 50 days
- 100 days (20 weeks)

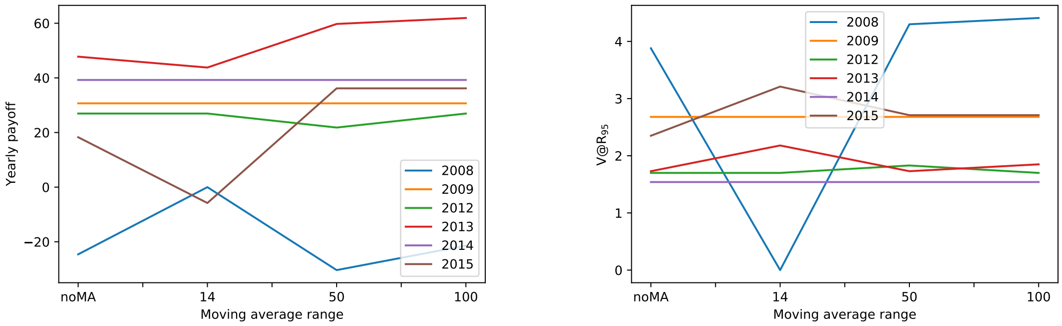
In Figure 5.14, the results are reported following the same logic of 5.6.1, but, since all the evaluated top portfolios may have no positive trend (e.g. 2008's financial crisis), when this eventuality happens, both the V@R$_{.95}$ and the payoffs will be indicated as 0 because it is supposed no investments.

What comes out from the visualization is that a mid-short range (14) could impact negatively in finding an investment when the market is unfavourable (2008-14SMA produces no possible portfolios), this is due to the negative general trend that makes difficult to have a positive indicator. Moreover, the overall behaviour when this range is selected seems to be a bit lower in payoffs. This result for payoffs is even clearer thanks to box-plots in Figure 5.13. A larger range (100) tends to give good payoffs



(a) Yearly payoff (b) V@R$_{.95}$

Figure 5.13: Yearly payoff and V@R$_{.95}$ Box-plots w.r.t. SMA



(a) Yearly payoff (b) V@R$_{.95}$

Figure 5.14: Yearly payoff and V@R$_{.95}$ w.r.t. SMA

Regarding the no moving averages option, it has good performances but the risk could increase drastically.
It must be remarked that technical indicator is considered to be an art, thus this should be considered as the most personalizable value, anyway, a mid-short bet may not be the best choice.

5.7 The optimization effect

Unlike the different parameters discussed so far, the optimization concept is intrinsically part of the tool. It is possible to argue on what is its role.

Taking as an example the standard conditions with $\lambda = 0.5$ applied to 2013, the Figure 5.15 shows how effectively all the proposed portfolios are higher than the benchmark, moreover, the tops overperform those who were at the bottom of the optimization classification.

Let us try to abstract statistically the optimization effect through some distributions and tests.

I will take into account all the years that have been analyzed so far (2008-2009, 2012-2017) and, more specifically, maintaining the standard conditions for the constraints, I am going to compute the empirical distributions of payoffs and V@R$_{.95}$ for the two top portfolios and the two bottom ones in 3 different values of lambda.

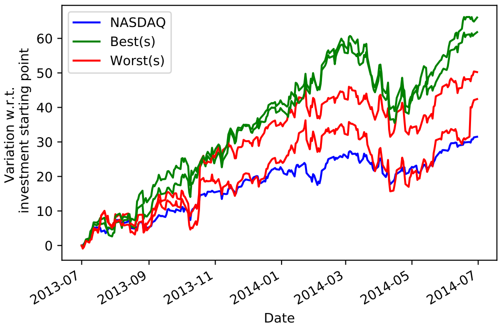


Figure 5.15: Average percentage variation 2013. Comparisons between Top(s) and Bottom(s)

$$\lambda = [0.25,0.5,0.75]$$

This latter choice has been made to avoid the extreme cases of the optimization where either no risk or no (partially) payoff is involved and also because the software is not so sensitive enough to vary portfolios for slight alterations of lambda.

Let us start from some data visualization. In Figure 5.16, the histograms of V@R and payoffs are plotted.

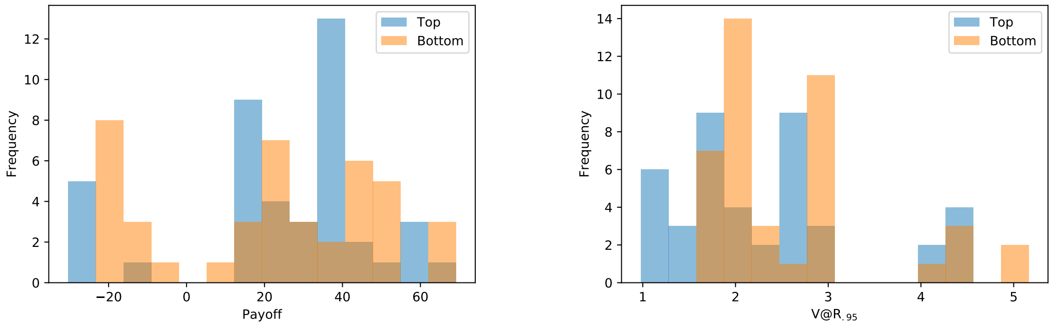
Concerning the V@R$_{.95}$ is quite evident that the optimization plays a role in minimizing it, but the aforesaid cannot be asserted for payoffs. This is neither so desired nor so unexpected.

The portfolios that are considered are filtered out by the w-support yet and they are defined 'worst' because their rank is far lower than the bests, but, for the sake of reliability, portfolios with single or very few stocks are filtered out[7].

Since a graphical point of view is not enough to prove mathematically what have been discussed above and optimization is an inseparable part of the tool, let us move on a Two-sample Kolmogorov–Smirnov test that allows testing if two underlying one-dimensional probability distributions differ.

As metioned in the previous tests, an usual choice for the Type I error threshold is $\alpha = 0.05$, hence I decided to operate with this value. The analytical results are tabulated in Table 5.5 and they authorize to reject the null hypothesis, that the V@R$_{.95}$ empirical distribution of the first and the second sample respectively are the same, at level $\alpha = 0.05$. Thus,

[7] I define 'worst' the last two portfolios across a sample that contains portfolios 1 to 100 according to OPTIMDISPLAN



(a) Yearly payoff histogram

(b) V@R$_{.95}$ histogram

Figure 5.16: Yearly payoff and V@R$_{.95}$ histograms. Comparisons between Top(s) and Bottom(s)

optimization plays a role in risk management. As awaited, I cannot say the same thing for

	Statistic	p_value
V@R$_{.95}$	0.31	0.03
Payoffs	0.24	0.18

Table 5.5: Yearly payoff and V@R$_{.95}$ distributional tests between Top(s) and Bottom(s)

payoffs, thus the null hypothesis is not rejectable at level α.

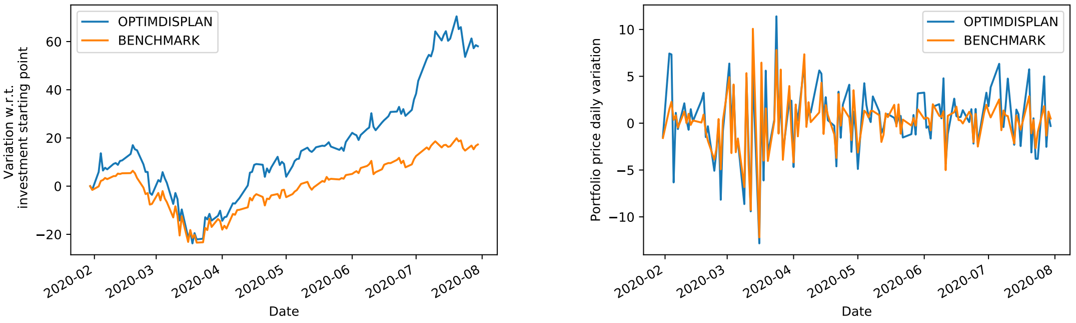
5.8 A recent case study: the COVID-19 pandemic

Meantime this thesis was being developed, the coronavirus pandemic, caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), exploded and it has had wide-ranging and severe impacts upon financial markets.
A natural investigation concern the OPTIMDISPLAN effect on a hypothetical portfolio selection before the outbreak.
On 30 January 2020, following the recommendations of the Emergency Committee, the WHO Director General declared that the outbreak constitutes a Public Health Emergency of International Concern (PHEIC) (World Health Organization [47]). It follows that I assumed to buy a portfolio precisely on that date, training the algorithm over the past six months and using the standard conditions of OPTIMIDISPLAN. Successively, because of the subsequent crisis, I also decided to test the algorithm tuning up the risk aversion as if the user was a priori risk-averse, even if I would rather concentrate on the standard condition due to the unpredictability of the outbreak.
Lastly, before the presentation of the results, it is necessary to specify that, the validation horizon will be shorter than the rest of the study because this chapter is being written on the 5-Aug-2020, moreover I sincerely hope that no other opportunity to test this kind of crisis will come.

5.8.1 The OPTIMDISPLAN financial recovery

Let us start in Figure 5.17 from the average percentage variation of the portfolio chosen by OPTIMDISPLAN against the NASDAQ benchmark used as stocks set and its respective volatility plot.

It is readily apparent that the financial downturn was not avoided, but usually, these bearish trends are heavily negative correlated across an indexes collection, however, it does not make matters worse.



(a) Average percentage variation.

(b) Volatility plot.

Figure 5.17: Average percentage variation and Volatility plot. Comparison with NASDAQ during COVID-19. $\lambda = 0.5$

Concerning the following uptrend, the behaviour is completely different an the losses are reset by the middle of April, whereas another month is needed for the benchmark.
Moving on some quantitative results, it is not possible to discuss the typical yearly based measures exploited in the rest of the tool's analysis, but, on top of the visual representation, it is useful to compare V@R$_{.95}$ and the to compute the Sharpe Ratio(5.4) with regard to the end of the validation period. These values are reported in Table 5.6.

	OPTIMDISPLAN	NASDAQ
Sharpe	0.98	0.45
V@R$_{.95}$	5.91	4.34

Table 5.6: Sharpe ratio and V@R$_{.95}$ OPTIMDISPLAN and NASDAQ during COVID-19

It is commonly said that an investment with a Sharpe ration near or over one it is considered a good one, hence the software selected portfolio may be seen as profitable even though the pandemic crisis. We cannot say that for the Benchmark.
Although the value at risk assumes a reasonable value, it is possible to test another possibility with a higher risk aversion on the tool that, looking at the experiments tried for the 2008 crisis, will be varied as $\lambda = 0.8$.

In Figure 5.18, it is shown that both the payoff and the resistance to the downturn seems to be lower than the standard condition settings, whereas the V@R$_{.95}$ is lower but faintly discernable than the previous one 5.6, thus it seems not so worthy to penalize the payoff

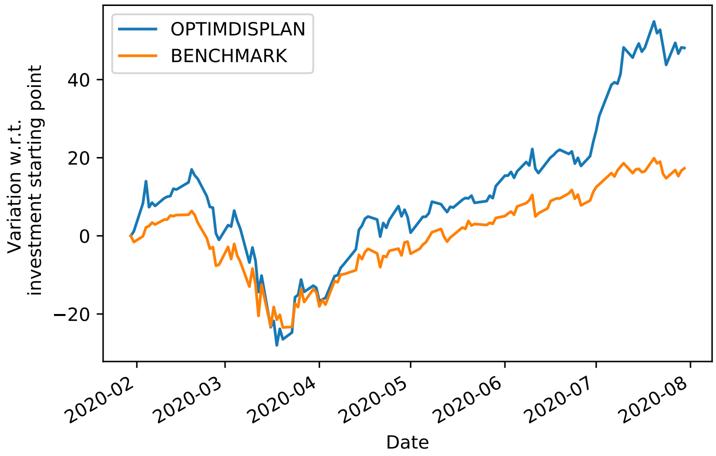


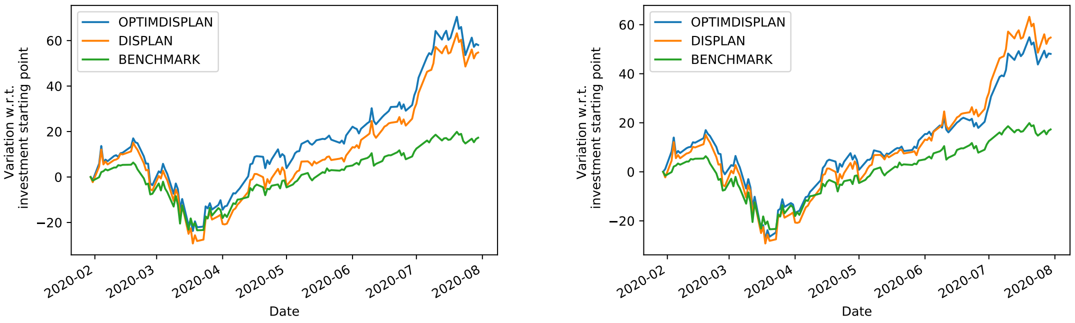
Figure 5.18: Average percentage variation. Comparison with NASDAQ during COVID-19. $\lambda = 0.8$

weight in a so correlated loss.

5.8.2 A pure data-mining selection

In most of the discussed section, a typically adopted benchmark have been DISPLAN. The purpose of this module is to test DISPLAN with the COVID-19 and to compare its performances with what has been exhibited so far.

From the Figure 5.19, it is plain that the downturn is even worse if only the data-mining algorithm is taken into account, moreover, about the comparison when lambda=0.5, the DISPLAN payoff never outperforms the optimization version. Also concerning the risk



(a) $\lambda = 0.5$

(b) $\lambda = 0.8$

Figure 5.19: Average percentage variation. Comparison with NASDAQ and DISPLAN during COVID-19. $\lambda = 0.5$ and $\lambda = 0.8$

point of view, the V@R is quite lower(Table 5.7).

	OPTIMDISPLAN	DISPLAN
V@R$_{.95}$ Daily	5.91	6.18
V@R$_{.95}$ Weekly	14.96	16.21

Table 5.7: V@R$_{.95}$ OPTIMDISPLAN and DISPLAN during COVID-19

Chapter 6

Conclusions

This thesis work focuses on studying and developing a decision support system for automated stock trading.
The proposed system, namely OPTIMDISPLAN, relies on a two-step buy-and-hold strategy, which comprises automatic portfolio generation, based on an itemset-based approach, and a portfolio selection strategy, based on an ad hoc optimization model.
Thanks to its customizable features, it is malleable and can reflect the experience of the user into its parameters.

The first step of this tool relied on a heuristic selection and suggestion of the portfolios. Making it part of a full decision support system, a real optimization based on the whole itemsets is developed considering both risk and payoff. Moreover, the optimal suggested solution examines several user preferences that reflect what the market and the investment perception is.
Although some combinations and internal values effects have been deeply discussed, a plethora of other possibilities are hidden into the software, but what seems to be clear is the profitability of the given result both in a higher payoff and in a lower risk.

Favourable and unfavourable market conditions have been widely investigated in terms of performances with a standard configuration of the DSS and objectively good investments are proposed. Nevertheless, also in a recent financial crisis like the COVID-19 pandemic, the software with its basic configuration gives a good yield.
Statistics in yearly returns and risk measures show that the model provides significant evidence of higher profits, furthermore, lots of other measures are produced by scripts and, even if they all have not been exploited into the analysis, these features can be useful to understand and to classify the suggested investments according to a professional user point of view.
The optimization seems to play a statistically significant role with regards to the risk. Whereas, regarding the yearly payoffs, the data-mining algorithm produces a high tier set of stocks yet that could be improved by the personalized tuning of risk aversion, technicals and fundamentals settings.

Future works Lots of further features could be implemented, but a trade-off between performances (in terms of possibilities) and comprehensibility should be taken into account. Some alternatives can deal with DISPLAN itself through the variation of the taxonomy or the complete substitution of the data-mining system that support OPTIMDISPLAN in favour of other unsupervised or supervised machine learning algorithms. This latter proposal could be especially considered when the software must deal with large stock collections, improving the data mining part both in speed and in necessary memory.
Some interesting stock collections could surely be S&P500 or even Russell2000, but tests with other markets that come from different continents such as South Asia (relatively new) or Europe (well-established market) could reveal good insights.
Lastly, but not leastly, the proposed system rely on an uniform investment strategy. Though this solution could be supported in case of extreme uncertainty or heavy risk aversion, a further enrichment of our tool should concern an optimized weight per stock.

Appendix A

Another Taxonomy: The cross-correlation

One of the central tools that OPTIMDISPLAN uses to mitigate the risk and achieve a profitable portfolio is diversification. As exposed previously 3.2, the diversification exploits a taxonomy that can be generalized according to different ways of clustering among firms. The default option is related to the GICS industries ([43]), but an innovative point of view has been recently developed by the Politecnico of Turin in the work of Fior et al. [20].

A.1 Price series cross-correlation analysis

The GICS diversification relies on the conjecture that same industrial sector usually provide correlation in the prices of the stocks belonging to it. This approach is quite conventional and could be biased, hence, even if some stocks belong to the same sector, it is typical to find variable and uncorrelated trends. On the other hand, stocks of different sectors may result to be fairly correlated with each other.

The studied cross-correlation analysis aims to identify groups of stocks showing similar temporal trends. This lead to a new taxonomy that will be used to diversify portfolios. Time series are cluster into a homogenous group based on their pairwise time-based similarity. The used cross-correlation is a statistical measure widely used for signal and image processing that will induce the relative distance to build clusters.

Neglecting the algorithm insights that are deeply inspected in Fior et al. [20], three different parameters drive the results:

- An initial number of clusters k
- A discretization threshold t
- A similarity threshold p

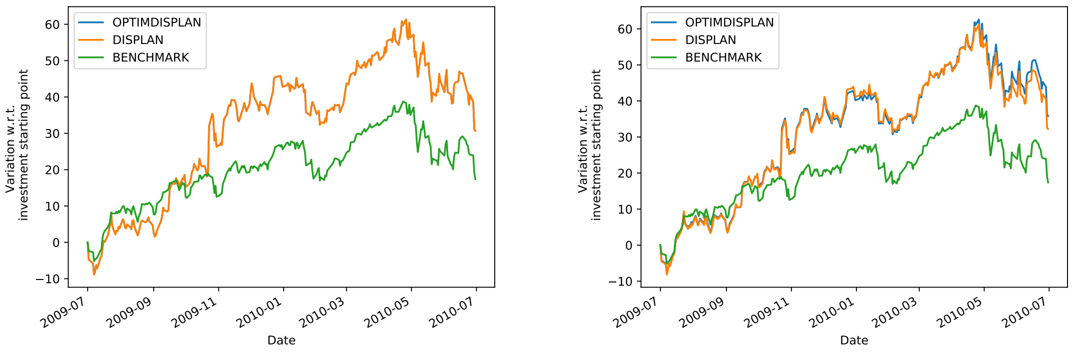
What I am going to choose are the suggested values, thus the number of initial clusters will be equal to the number of GICS sectors, whereas p and t will be 70% and 85% respectively.

A.2 Experimental results

The idea is to investigate the OPTIMDISPLAN behaviour with this new taxonomy in the market conditions discussed in Fior et al. [20], hence, similarly to Baralis et al. [8] unfavorable(2008-2009) and favorable(2013) condition are stressed, but also a mixed-trend in 2014-2016 is shown.
It must be remarked that the development of the clusters needs a training period that anticipates the OPTIMDISPLAN one, indeed it is meant that to train the decision support system, clusters must be known yet at its start. This necessity will shift the 2008 test onto 2009 due to a lack of data in our database, but the overall behaviour remains easily comparable with OPTIMDISPLAN results.

A.2.1 Unfavorable market conditions

Let us start from the 2008 crisis. The training period for the clustering algorithm will make OPTIMDISPLAN available in 2009. For the sake of simplicity, in Figure A.1 is reported the average daily variation w.r.t the starting point of the investment that has been previously presented during the performance analysis core. Alongside the standard set plot, it is shown the same validation equity line with the new taxonomy.

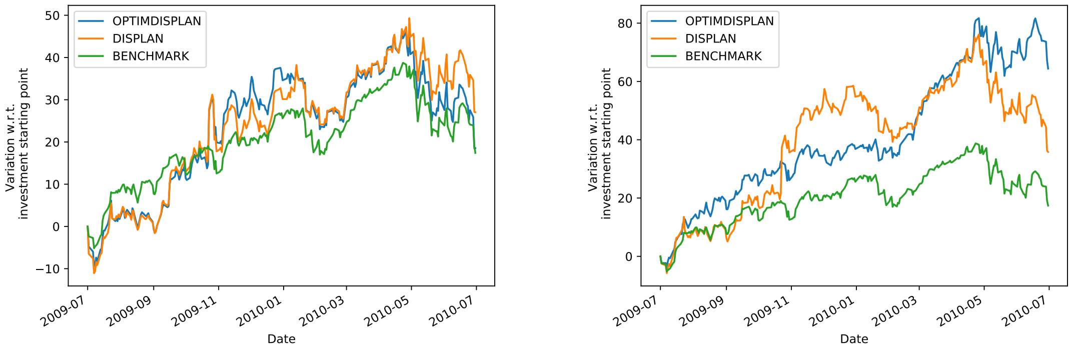


(a) Average percentage variations no Cross-Correlation

(b) Average percentage variations with Cross-Correlation

Figure A.1: Average percentage variation w.r.t. 2009 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.7.

Albeit the improvement it is not well-touchable, it is present and a strong result is achieved when the diversification constraint of OPTIMDISPLAN is raised. For example, in Figure A.2, it is reported the same comparison but with a diversification threshold of 0.9 instead of 0.7 and a more visible effect of the taxonomy appears.
It is remarkable that in all the presented experiments, the DISPLAN comparison deals with the complete version of the data-mining approach, hence the diversification method plays a role in this benchmark and the diversification threshold is tuned coherently. A quantitative resume is provided in Table A.1 and, besides the verification of higher performances with the



(a) Average percentage variations no Cross-Correlation

(b) Average percentage variations with Cross-Correlation

Figure A.2: Average percentage variation w.r.t. 2009 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.9.

innovative approach, it is perceptible an over-diversification effect when the new taxonomy is not exploited but the threshold is raised, indeed both payoffs and risk measures of OPTIMDISPLAN are deteriorated when div=0.9 and the diversification is sector based.

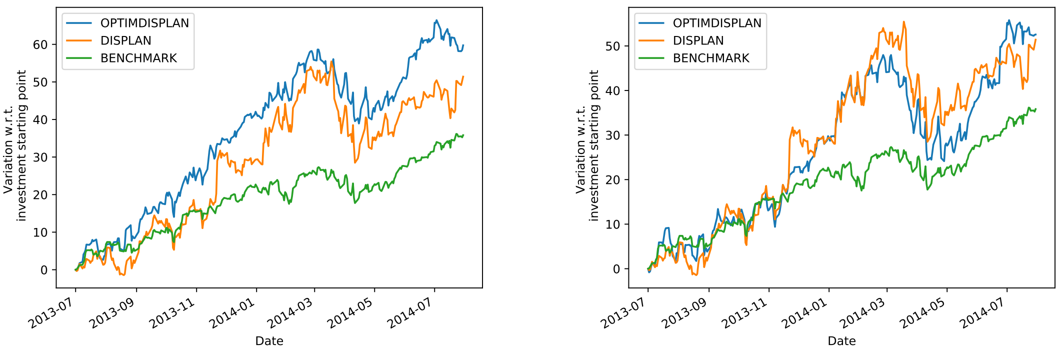
	OPTIMDISPLAN	DISPLAN	NASDAQ
Yearly payoff (div=0.7) New Tax	35.78	32.21	17.40
Yearly payoff (div=0.9) New Tax	64.39	35.80	17.40
V@R$_{.95}$ (div=0.7) New Tax	2.75	2.91	2.35
V@R$_{.95}$ (div=0.9) New Tax	2.52	2.76	2.35
Yearly payoff (div=0.7) Sectors	30.68	30.68	17.40
Yearly payoff (div=0.9) Sectors	18.53	27.04	17.40
V@R$_{.95}$ (div=0.7) Sectors	2.68	2.68	2.35
V@R$_{.95}$ (div=0.9) Sectors	2.85	2.67	2.35

Table A.1: Quantitative results w.r.t. cross-correlated taxonomy, year 2009

A.2.2 Favorable market conditions

Let us now discuss what happens if we test OPTIMDISPLAN in 2013 with a different taxonomy.
The standard configuration gives us an astonishing result that has been repeated in Figure A.3 for the sake of clarity and, even though these performances are not confirmed with regards to the new taxonomy, both the experiments well outperform the benchmark.

A fair question is to wonder if a higher diversification threshold produces a better result with the cross-correlated clusters and, similarly to the 2009 case, in Figure A.4 the respective equity lines are represented.



(a) Average percentage variations no Cross-Correlation

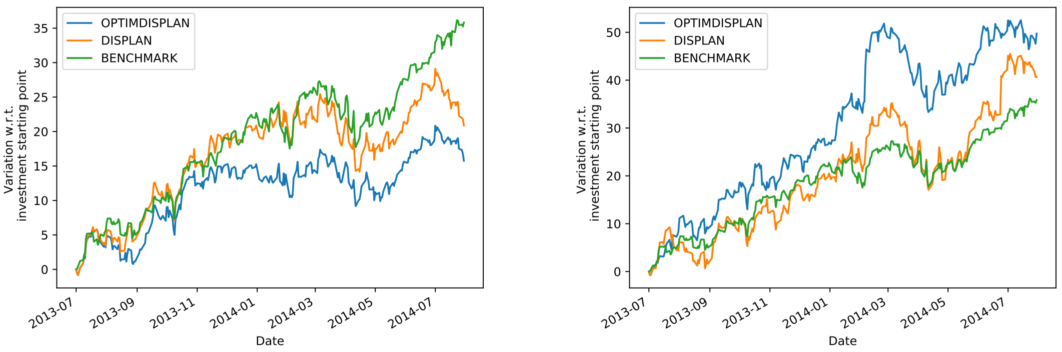
(b) Average percentage variations with Cross-Correlation

Figure A.3: Average percentage variation w.r.t. 2013 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.7.

What comes up from the visual performance is again a hint of over-diversification when the software is sectors based, whereas a better performance is clear with the new taxonomy. In Table A.2 some quantitative results allow to confirm the previous observations in terms of general behaviour, indeed the higher diversification case lose if we would consider only the yearly returns.
A naive justification of the worst behaviour concerning this new taxonomy could be related to the bullish market behaviour. It is not a secret that the sectors based version could heavier rely on a particular kind of stocks that well exploit the market trend, whereas a deeper diversification is useful to better manage risk when the market has downturns.
Lastly, could be useful to remark that the new taxonomy affects more the whole DSS than the pure data-mining algorithm, indeed its related portfolios always change, but the DISPLAN ones do not vary when the diversification threshold is 0.7.

	OPTIMDISPLAN	NASDAQ	DISPLAN
Yearly payoff (div=0.7) New Tax	52.56	35.82	51.40
Yearly payoff (div=0.9) New Tax	49.74	35.82	40.66
V@R$_{.95}$ (div=0.7) New Tax	2.31	1.28	2.59
V@R$_{.95}$ (div=0.9) New Tax	1.57	1.28	1.70
Yearly payoff (div=0.7) Sectors	59.75	35.82	51.40
Yearly payoff (div=0.9) Sectors	15.75	35.82	20.88
V@R$_{.95}$ (div=0.7) Sectors	1.73	1.28	2.59
V@R$_{.95}$ (div=0.9) Sectors	1.31	1.28	1.44

Table A.2: Quantitative results w.r.t. cross-correlated taxonomy, year 2013



(a) Average percentage variations no Cross-Correlation

(b) Average percentage variations with Cross-Correlation

Figure A.4: Average percentage variation w.r.t. 2013 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.9.

A.2.3 Mixed market conditions

The 2016-2017 period has not been discussed as deeply as 2008 or 2013, hence a preliminary analysis with regards to the standard conditions should be made.
In these experiments, as reported in Figure A.5, the standard configuration of OPTIMDISPLAN do not provide promising results, but some insights could well explain what phenomena are going on.

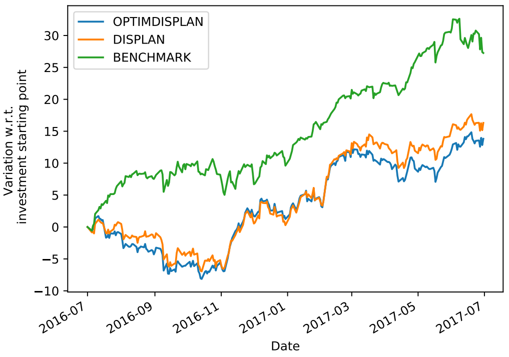
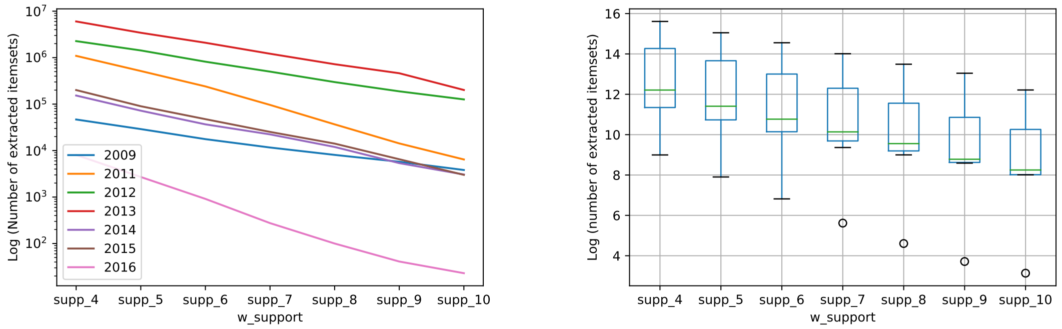


Figure A.5: Average percentage variation w.r.t. 2016 $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, w-support=8%, diversification=0.7.

Looking at Figure A.6, the number of extracted itemsets by the data-mining algorithm is extremely low if compared with other experiments with the same support, hence the suggested portfolios in the first stage of OPTIMDISPLAN do not show properly the market potential. When the user sets the DSS parameters he can easily notice that the overall performance are different from what the standard behaviour of the software is in terms of analyzed portfolios. Anyway, future works could involve automated management to provide a support cue for the user.
Focusing on the box-plot, concerning 2016, possible support could be 6%, indeed the number of extracted itemsets do not exceed the third quartile compared to the overall statistics, thus we will select this threshold to make some experiments.

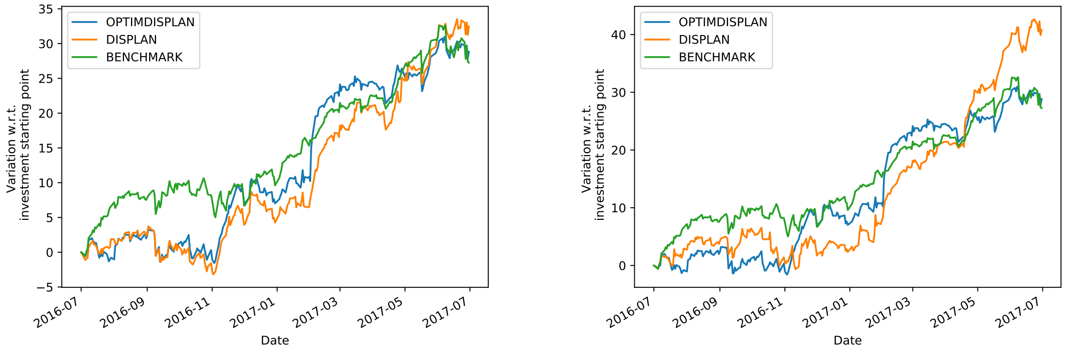


(a) Yearly lines

(b) Box-plot

Figure A.6: Logarithmic number of extracted itemstets by the first step of OPTIMDISPLAN w.r.t. the minimum w-support

Once defined the meta-parameters, the experiment design will follow the previously adopted setting, hence two different thresholds of diversification will be discussed with and without the innovative taxonomy, exploiting the standard conditions for OPTIMDISPLAN. In



(a) Average percentage variations no Cross-Correlation

(b) Average percentage variations with Cross-Correlation

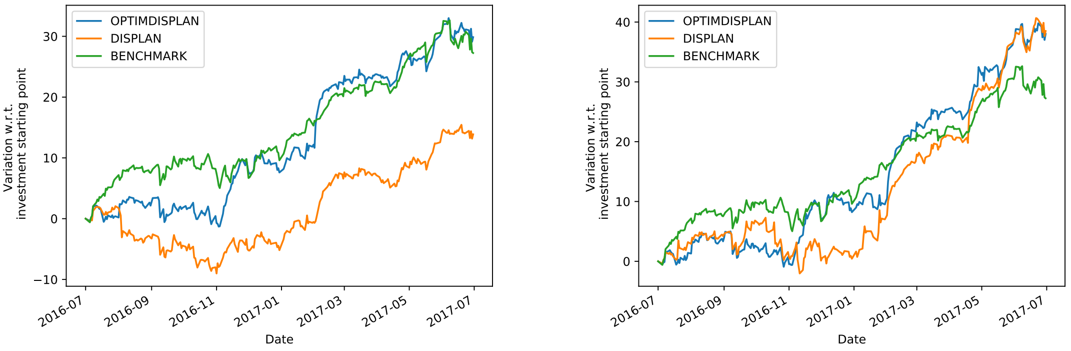
Figure A.7: Average percentage variation w.r.t. 2016 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.7.

Figure A.7 and Figure A.8, it is represented what are the visual differences and several comments should be made.

First of all, the pure data-mining approach strongly varies up and down with regards to the benchmark, whereas the DSS solutions are more stable between the experiments.

Regarding the performances, DISPLAN utterly suffers the over-diversification effect when the sectors taxonomy is adopted, but the whole DSS well manage these changes. When the cross-correlation approach is preferred, a higher diversification threshold continues to provide better results, however, a formal statistical investigation of this common event should be carried on.

In Table A.3, some insight related to the risk could be discovered, but the usual behaviour of OPTIMDISPLAN is shown.



(a) Average percentage variations no Cross-Correlation

(b) Average percentage variations with Cross-Correlation

Figure A.8: Average percentage variation w.r.t. 2016 (Cross-Correlated taxonomy) $\lambda = 0.5$. Comparison with DISPLAN and NASDAQ-100, diversification=0.9.

	OPTIMDISPLAN	NASDAQ	DISPLAN
Yearly payoff (div=0.7) New Tax	28.82	27.25	40.78
Yearly payoff (div=0.9) New Tax	38.01	27.25	38.50
$V@R_{.95}$ (div=0.7) New Tax	1.03	0.94	1.28
$V@R_{.95}$ (div=0.9) New Tax	1.23	0.94	1.24
Yearly payoff (div=0.7) Sectors	28.82	27.25	32.47
Yearly payoff (div=0.9) Sectors	29.86	27.25	13.87
$V@R_{.95}$ (div=0.7) Sectors	1.03	0.94	1.23
$V@R_{.95}$ (div=0.9) Sectors	1.04	0.94	1.04

Table A.3: Quantitative results w.r.t. cross-correlated taxonomy, year 2016

Appendix B

Selected portfolios

B.1 Selection pattern

All the portfolios that have been selected by OPTMIDSPLAN in this study, except those that concern the Validation of Statistical Significance of the performance improvements 5.5, will be summarized hereafter. The tuned parameters of the DSS are schematically divided into :

- The training period
- The risk aversion coefficient
- The moving average interval and type
- The fundamental indicator selection (or `ON` if all are supposed reliable)

For each combination, the portfolio is composed of the stock symbols. For the sake of readability, the whole company name is avoided, but if the reader would like to deeper investigate the firm names, it is possible to search for them into the NASDAQ official website `https://www.nasdaq.com/market-activity/stocks/screener`.

B.2 Portfolios

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2008-30/06/2008	0.5	SMA(50)	ON	HAS ILMN TTWO ROST CSX
1/1/2009-30/06/2009	0.5	SMA(50)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.5	SMA(50)	ON	JBHT ALXN ILMN AAPL MNST ULTA REGN

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2013-30/06/2013	0.5	SMA(50)	ON	AMAT CSX WDC HAS INCY NFLX MU
1/1/2014-30/06/2014	0.5	SMA(50)	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.5	SMA(50)	ON	ULTA HAS MNST INCY AMZN SWKS EA
1/1/2016-30/06/2016	0.5	SMA(50)	ON	FISV PCAR XEL HAS
1/1/2008-30/06/2008	0.25	SMA(50)	ON	HAS ILMN TTWO ROST CSX
1/1/2009-30/06/2009	0.25	SMA(50)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.25	SMA(50)	ON	JBHT ALXN ILMN AAPL MNST ULTA REGN
1/1/2013-30/06/2013	0.25	SMA(50)	ON	CSX WDC EA INCY HAS MU NFLX
1/1/2014-30/06/2014	0.25	SMA(50)	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.25	SMA(50)	ON	ULTA AMZN MNST SBUX INCY EA SWKS
1/1/2016-30/06/2016	0.25	SMA(50)	ON	FISV FAST JBHT HAS XEL
1/1/2008-30/06/2008	0.75	SMA(50)	ON	HAS ILMN TTWO ROST CSX

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2009-30/06/2009	0.75	SMA(50)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.75	SMA(50)	ON	CMCSA JBHT AAPL ALXN MNST ULTA REGN
1/1/2013-30/06/2013	0.75	SMA(50)	ON	ORLY BIIB INCY UAL HAS MU NFLX
1/1/2014-30/06/2014	0.75	SMA(50)	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.75	SMA(50)	ON	ULTA HAS MNST NFLX INCY EA SWKS
1/1/2016-30/06/2016	0.75	SMA(50)	ON	FISV PCAR XEL HAS
1/1/2008-30/06/2008	0	SMA(50)	ON	HAS ILMN TTWO ROST CSX
1/1/2009-30/06/2009	0	SMA(50)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0	SMA(50)	ON	JBHT ALXN ILMN AAPL MNST ULTA REGN
1/1/2013-30/06/2013	0	SMA(50)	ON	CSX WDC EA INCY HAS MU NFLX
1/1/2014-30/06/2014	0	SMA(50)	ON	NVDA ALXN EA IDXX WBA SWKS AAL
1/1/2015-30/06/2015	0	SMA(50)	ON	ULTA AMZN MNST SBUX INCY EA SWKS

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2016-30/06/2016	0	SMA(50)	ON	FISV FAST PCAR HAS XEL
1/1/2008-30/06/2008	1	SMA(50)	ON	DLTR TTWO ILMN HAS CSX
1/1/2009-30/06/2009	1	SMA(50)	ON	ATVI MYL AMZN ILMN WDC
1/1/2012-30/06/2012	1	SMA(50)	ON	CMCSA JBHT ROST CTXS AAPL ULTA REGN
1/1/2013-30/06/2013	1	SMA(50)	ON	ORLY BIIB INCY UAL HAS MU NFLX
1/1/2014-30/06/2014	1	SMA(50)	ON	MAR TTWO MXIM IDXX WBA SWKS AAL
1/1/2015-30/06/2015	1	SMA(50)	ON	ULTA HAS MNST NFLX INCY EA SWKS
1/1/2016-30/06/2016	1	SMA(50)	ON	FISV FAST XEL HAS
1/1/2008-30/06/2008	0.5	no MA	ON	XLNX TTWO DLTR HAS ILMN CSX
1/1/2009-30/06/2009	0.5	no MA	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.5	no MA	ON	JBHT INCY AAPL ILMN MNST ULTA REGN
1/1/2013-30/06/2013	0.5	no MA	ON	WBA WDC HAS VRTX UAL MU NFLX

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2014-30/06/2014	0.5	no MA	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.5	no MA	ON	ULTA HAS MNST INCY SBUX SWKS EA
1/1/2008-30/06/2008	0.5	SMA(14)	ON	-
1/1/2009-30/06/2009	0.5	SMA(14)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.5	SMA(14)	ON	JBHT INCY AAPL ILMN MNST ULTA REGN
1/1/2013-30/06/2013	0.5	SMA(14)	ON	CSX VRTX UAL HAS INCY MU NFLX
1/1/2014-30/06/2014	0.5	SMA(14)	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.5	SMA(14)	ON	BIIB SBUX INCY AMZN MNST EA SWKS
1/1/2008-30/06/2008	0.5	SMA(100)	ON	XLNX TTWO ROST HAS ILMN JBHT
1/1/2009-30/06/2009	0.5	SMA(100)	ON	MYL ILMN AMZN NFLX WDC
1/1/2012-30/06/2012	0.5	SMA(100)	ON	JBHT INCY AAPL ILMN MNST ULTA REGN
1/1/2013-30/06/2013	0.5	SMA(100)	ON	WBA WDC INCY UAL HAS MU NFLX

Training	λ-Risk Aversion	MA	Fundamentals	Portfolio
1/1/2014-30/06/2014	0.5	SMA(100)	ON	ORLY MXIM EA WBA IDXX SWKS AAL
1/1/2015-30/06/2015	0.5	SMA(100)	ON	ULTA HAS MNST INCY AMZN SWKS EA
1/1/2008-30/06/2008	0.8	SMA(50)	ON	DLTR TTWO ILMN HAS CSX
30/7/2019-29/01/2020	0.5	SMA(50)	ON	VRTX TSLA AMGN AAPL NTES
30/7/2019-29/01/2020	0.8	SMA(50)	ON	CHTR TSLA LRCX AMGN NTES

Table B.1: Experiments summary

Bibliography

[1] Jeffery S. Abarbanell and Brian J. Bushee. Abnormal returns to a fundamental analysis strategy. *The Accounting Review*, 73(1):19–45, January 1998.

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. *Mining Association Rules Between Sets of Items in Large Databases, SIGMOD Conference*, volume 22, pages 207–. 06 1993. doi: 10.1145/170036.170072.

[3] Renato Aguiar and Roberto Sales. Analysis of overreaction and underreaction in the american stock market using fuzzy clustering means algorithm. pages 376–380, 09 2010. ISBN 978-1-4244-6927-7. doi: 10.1109/ICIFE.2010.5609382.

[4] S. Al-augby, S. Majewski, K. Nermend, and A. Majewska. Proposed investment decision support system for stock exchange using text mining method. In *2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA)*, pages 1–6, 2016.

[5] K. A. Althelaya, E. M. El-Alfy, and S. Mohammed. Evaluation of bidirectional lstm for short-and long-term stock market prediction. In *2018 9th International Conference on Information and Communication Systems (ICICS)*, pages 151–156, April 2018.

[6] E.J. Anderson. *Business Risk Management: Models and Analysis*. Wiley, 2013. ISBN 9781118749364. URL https://books.google.it/books?id=wRa5AQAAQBAJ.

[7] Tanvir Ansari, Manoj Kumar, Anupam Shukla, Joydip Dhar, and Ritu Tiwari. Sequential combination of statistics, econometrics and adaptive neural-fuzzy interface for stock market prediction. *Expert Syst. Appl.*, 37(7):5116–5125, July 2010. ISSN 0957-4174. doi: 10.1016/j.eswa.2009.12.083. URL https://doi.org/10.1016/j.eswa.2009.12.083.

[8] Elena Baralis, Luca Cagliero, and Paolo Garza. Planning stock portfolios by means of weighted frequent itemsets. *Expert Systems with Applications*, 86:1 – 17, 2017. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2017.05.051. URL http://www.sciencedirect.com/science/article/pii/S0957417417303731.

[9] Messod D. Beneish. The detection of earnings manipulation. *Financial Analysts Journal*, 55(5):24–36, 1999. doi: 10.2469/faj.v55.n5.2296. URL https://doi.org/10.2469/faj.v55.n5.2296.

[10] Messod D. Beneish, Charles M. C. Lee, and Robin L. Tarpley. Contextual fundamental analysis through the prediction of extreme returns. *Review of Accounting Studies*, 6(2):

165–189, 2001. doi: 10.1023/A:1011654624255. URL https://doi.org/10.1023/A:1011654624255.

[11] Victor L. Bernard and Jacob K. Thomas. Post-earnings-announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, 27:1–36, 1989. ISSN 00218456, 1475679X. URL http://www.jstor.org/stable/2491062.

[12] O Bustos and A. Pomares-Quimbaya. Stock market movement forecast: A systematic review. *Expert Systems with Applications*, 156:113464, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.113464. URL http://www.sciencedirect.com/science/article/pii/S0957417420302888.

[13] Luca Cagliero and Paolo Garza. Infrequent weighted itemset mining using frequent pattern growth. *Knowledge and Data Engineering, IEEE Transactions on*, 26:903–915, 04 2014. doi: 10.1109/TKDE.2013.69.

[14] Gen-Huey Chen, Ming-Yang Kao, Yuh-Dauh Lyuu, and Hsing-Kuo Wong. Optimal buy-and-hold strategies for financial markets with bounded daily returns. *SIAM Journal on Computing*, 31(2):447–459, 2001. doi: 10.1137/S0097539799358847. URL https://doi.org/10.1137/S0097539799358847.

[15] Y. Chen, S. Mabu, K. Hirasawa, and J. Hu. Genetic network programming with sarsa learning and its application to creating stock trading rules. In *2007 IEEE Congress on Evolutionary Computation*, pages 220–227, 2007.

[16] Keith Cuthbertson, Dirk Nitzsche, and Niall O'Sullivan. Uk mutual fund performance: Skill or luck? *Journal of Empirical Finance*, 15(4):613 – 634, 2008. ISSN 0927-5398. doi: https://doi.org/10.1016/j.jempfin.2007.09.005. URL http://www.sciencedirect.com/science/article/pii/S092753980700103X.

[17] Harris Dellas and Martin Hess. Financial development and stock returns: A cross-country analysis. *Journal of International Money and Finance*, 24(6):891 – 912, 2005. ISSN 0261-5606. doi: https://doi.org/10.1016/j.jimonfin.2005.07.002. URL http://www.sciencedirect.com/science/article/pii/S0261560605000690.

[18] Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2325486.

[19] EUGENE F. FAMA and KENNETH R. FRENCH. Size and book-to-market factors in earnings and returns. *The Journal of Finance*, 50(1):131–155, 1995. doi: 10.1111/j.1540-6261.1995.tb05169.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1995.tb05169.x.

[20] Jacopo Fior, Luca Cagliero, and Paolo Garza. Price series cross-correlation analysis to enhance the diversification of itemset-based stock portfolios. In *Proceedings of the Sixth International Workshop on Data Science for Macro-Modeling*, DSMM '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380300. doi: 10.1145/3401832.3402680. URL https://doi.org/10.1145/3401832.3402680.

[21] Fu-lai Chung, Tak-chung Fu, R. Luk, and V. Ng. Evolutionary time series segmentation for stock data mining. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 83–90, 2002.

[22] Jarg Gottschlich and Oliver Hinz. A decision support system for stock investment recommendations using collective wisdom. *Decision Support Systems*, 59:52 – 62, 2014. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2013.10.005. URL http://www.sciencedirect.com/science/article/pii/S0167923613002522.

[23] Dries Heyman, Koen Inghelbrecht, and Stefaan Pauwels. Good luck, bad luck. can mutual funds really pick stocks? *SSRN Electronic Journal*, 01 2014. doi: 10.2139/ssrn.2395955.

[24] Po-Hsuan Hsu, Chung-Ming Kuan, et al. Re-examining the profitability of technical analysis with white's reality check. *Analysis of High-Frequency Financial Data and Market Microstructure*, 2004.

[25] R. A. Kamble. Short and long term stock trend prediction using decision tree. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1371–1375, June 2017. doi: 10.1109/ICCONS.2017.8250694.

[26] Mehmed Kantardzic, Pedram Sadeghian, and Chun Shen. The time diversification monitoring of a stock portfolio: An approach based on the fractal dimension. volume 1, pages 637–641, 01 2004. doi: 10.1145/967900.968034.

[27] V. Kedia, Z. Khalid, S. Goswami, N. Sharma, and K. Suryawanshi. Portfolio generation for indian stock markets using unsupervised machine learning. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–5, 2018.

[28] Monica Lam. Neural network techniques for financial performance prediction: integrating fundamental and technical analysis. *Decision Support Systems*, 37(4):567 – 581, 2004. ISSN 0167-9236. doi: https://doi.org/10.1016/S0167-9236(03)00088-5. URL http://www.sciencedirect.com/science/article/pii/S0167923603000885. Data mining for financial decision making.

[29] Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/2975974.

[30] H.M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. Monograph / Cowles Foundation for Research in Economics at Yale University. Wiley, 1991. ISBN 9781557861085. URL https://books.google.it/books?id=T2PHRWxp_RkC.

[31] Lukas Menkhoff. The use of technical analysis by fund managers: International evidence. *Journal of Banking & Finance*, 34:2573–2586, 11 2010. doi: 10.1016/j.jbankfin.2010.04.014.

[32] Danilo Alcantara Milhomem and Maria José Pereira Dantas. Analysis of new approaches used in portfolio optimization: a systematic literature review. *Production*, 30, 00 2020. ISSN 0103-6513. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-65132020000100404&nrm=iso.

[33] Partha S. Mohanram. Separating winners from losers among lowbook-to-market stocks using financial statement analysis. *Review of Accounting Studies*, 10(2): 133–170, 2005. doi: 10.1007/s11142-005-1526-4. URL https://doi.org/10.1007/s11142-005-1526-4.

[34] J.J. Murphy. *Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications*. New York Institute of Finance Series. New York Institute of Finance, 1999. ISBN 9780735200661. URL https://books.google.it/books?id=5zhXEqdr_IcC.

[35] Isaac kofi Nti, Adebayo Adekoya, and Benjamin Weyori. A systematic review of fundamental and technical analysis of stock market predictions. *Artificial Intelligence Review*, 08 2019. doi: 10.1007/s10462-019-09754-z.

[36] Felipe Dias Paiva, Rodrigo Tomás Nogueira Cardoso, Gustavo Peixoto Hanaoka, and Wendel Moreira Duarte. Decision-making for financial trading: A fusion approach of machine learning and portfolio selection. *Expert Systems with Applications*, 115:635 – 655, 2019. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.08.003. URL http://www.sciencedirect.com/science/article/pii/S0957417418305037.

[37] Joseph Piotroski. Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, 38, 01 2001. doi: 10.2307/2672906.

[38] Yue Qi, Bin Zhou, and Chao Wang. A computational analysis of the contradiction of mean-variance efficiency and diversification of portfolio selection and management. *Proceedings - 2009 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2009*, 2, 11 2009. doi: 10.1109/ICICISYS.2009.5358430.

[39] European securities and markets authority. *Annex to CESR'S technical advice on the level 2 measures related to the format and content of KID disclosures for UCITS: methodology for the calculation of the synthetic risk and reward indicator*, 2009.

[40] Ana Paula Serra. Country and industry factors in returns: evidence from emerging markets stocks. *Emerging Markets Review*, 1(2):127 – 151, 2000. ISSN 1566-0141. doi: https://doi.org/10.1016/S1566-0141(00)00007-8. URL http://www.sciencedirect.com/science/article/pii/S1566014100000078.

[41] William F. Sharpe. The sharpe ratio. *The Journal of Portfolio Management*, 21(1): 49–58, 1994. ISSN 0095-4918. doi: 10.3905/jpm.1994.409501. URL https://jpm.pm-research.com/content/21/1/49.

[42] Richard G. Sloan. Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review*, 71(3):289–315, 1996. ISSN 00014826. URL http://www.jstor.org/stable/248290.

[43] Standard & Poor's Financial Services LLC (S&P) and MSCI. *Global Industry Classification Standard*, 2018.

[44] Chandima D. Tilakaratne, Musa A. Mammadov, and Sidney A. Morris. Predicting trading signals of stock market indices using neural networks. In Wayne Wobcke and Mengjie Zhang, editors, *AI 2008: Advances in Artificial Intelligence*, pages 522–531, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[45] Stan Uryasev. Conditional value-at-risk: optimization algorithms and applications. volume 14, pages 49 – 57, 02 2000. ISBN 0-7803-6429-5. doi: 10.1109/CIFER.2000.844598.

[46] T. Williams and V. Turton. *Trading Economics: A Guide to Economic Statistics for Practitioners and Students*. The Wiley Finance Series. Wiley, 2014. ISBN 9781118766415. URL https://books.google.it/books?id=vYlPAwAAQBAJ.

[47] (WHO) World Health Organization. Covid-19 public health emergency of international concern (pheic) global research and innovation forum. 2020. URL https://www.who.int/docs/default-source/coronaviruse/global-research-and-innovation-forum-towards-a-research-roadmap.pdf?sfvrsn=a7fdb05b_1&download=true.

[48] Fengmei Yang, Zhiwen Chen, Jingjing Li, and Ling Tang. A novel hybrid stock selection method with stock prediction. *Applied Soft Computing*, 80:820 – 831, 2019. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2019.03.028. URL http://www.sciencedirect.com/science/article/pii/S1568494619301462.