UNIVERSITÀ DEGLI STUDI FIRENZE

Scuola di
Economia e Management
Corso di Laurea Magistrale in
Finance and Risk Management

Artificial Neural Networks and Genetic Algorithm for Value-at-Risk Model Combination

A Practical Application for Pension Funds' ETF Investments

Thesis Adivsor
Ilaria Colivicchi

Co-Advisor
Daniele Tantari

Candidate
Alessandro D'Amico

Academic Year 2018/2019

Table of Contents:

Abstract

In the framework of a growing importance for complementary pension plans, that could be able to integrate Social Security plans, Defined Benefit appears to be the most balanced solution both for members and providers. It becomes hence fundamental for Pension Fund managers to find alternative investments in order to guarantee acceptable returns, mitigating at the same time the risk. A recently developed solution is the use of Exchange Traded Funds and this is obtaining a pivotal role in building long term and balanced portfolios.

This work starts from these considerations (and from the consequent need for portfolio managers to monitor and reduce market risk) to present an alternative solution in Market Risk Management, that can avoid problems in model selection, based on the use of Artificial Neural Networks trained to combine two of the most common models of Value-at-Risk: the Historical Simulation and the GARCH model.

Due to the peculiar features of the loss function to be minimized, the Neural Network could not be trained through the standard method of Backpropagation of Error and Gradient Descent but employing a Genetic Algorithm based on reproducing the process of natural selection.

After having explained the theoretical basis of the study, we show how the Network trained to minimize a specific loss function is able to produce a third model that can pass the most common Backtesting criteria used to assess the reliability of VaR models.

We conclude underlining how this approach can be employed in other fields related to the Market Risk Management not only related to Pension Funds.

Introduction

In these years the entire pension system has been questioned, especially in the western developed economies where the ageing of society is forcing the legislators to redesign pension schemes to avoid the collapse of public finances. In this framework, the Defined Contribution scheme appears to be the most balanced in terms of risk-return. However, the issue of managing the market risk of the investment remains very important as well as for other financial institutions.

In recent years, a new financial instrument was developed on the idea of not trying to beat the market but to follow it: the Exchange Traded Fund. These ETFs had the main objective to track an index keeping low the costs. They hence provide remarkable results in the long run with very few commissions. An ideal combination for the typical member of a Pension Plan which is usually focused long term investments with low risks and fees.

ETFs were considered in this study for two main reasons: they are a core component for every Pension Funds' portfolio but they also represent already a well-diversified portfolio, easing the computation and the problem in building one from scratch.

Indeed, the core of this study is the presentation of an alternative approach to Market Risk Management: considering the Value-at-Risk (VaR) introduced by J.P. Morgan in 1996 as the main tool for this type of risk, we describe how the use of Artificial Neural Networks can combine two of the most important models into a third one presenting the positive features of the two inputs.

VaR is generally defined as the maximum possible loss for a portfolio within a certain confidence level and a time horizon and the candidate models of this study are the Historical Simulation and the GARCH model (assuming the normality of returns). The choice is affected by the fact that the first is the most common non-parametric approach being based on an empirical quantile while the second is a parametric model assuming normal returns: their combination is hence useful as we are employing non-overlapping informations. While HS adapts more slowly to structural breaks but it is quite

accurate in the long-run, GARCH is able to capture rapid changes in volatility being more precise in the short run.

Thus, this approach allows us to avoid all the problems related to model selection since each individual model shows advantages but also biases.

The theory of forecast combination was first developed by Bates and Granger[1] and it allows to absorb different VaR models' adaptability, reducing the forecast error uncertainty. This practice is accepted by practitioners since we are searching for the "best" model and not for the "correct" one.

Neural Networks are a branch of Machine Learning that is showing remarkable growth since abundant literature has shown how they can approximate a large class of functions being a generalization of non-linear regressions. For this reason, they are potentially suited for the problem of forecast combination since the solution is likely to be non-linear.

Since the VaR is a quantile, the common training procedure for the Neural Networks based on Gradient Descent would not be applicable as the loss function to be minimized is asymmetric and non-differentiable. Thus, we trained the ANN using a Genetic Algorithm whose process replicates the natural selection.

We are then comparing the ANN-VaR's performances with the ones of the HS and GARCH applying the common backtesting procedure based on Violation Ratio, Proportion Of Failures, Time Under First Failure, Christoffersen's Interval Forecast Test, Basel Traffic Lights and Mean 'Tick' Loss Function.

The study is hence organized as follows:

In Chapter 1 we are introducing aims and types of Pension Funds, considering also the problems related to their Risk Management.

In Chapter 2 we are going to describe ETFs starting from the issuance to their compositions and structures.

Chapter 3 will be dedicated to the description of VaR and the different types of models (each one presenting several pros and cons).

[1] J.M. Bates, and C.W.J. Granger (1969), The combination of forecasts. Operations Research Quarterly, 20, 451-468

In Chapter 4 we are presenting the topic of Artificial Neural Network analysing their applications, structures and training. This last point will also see the description of the Genetic Algorithm as an alternative to more common training procedures.

Finally, in Chapter 5 we are combining all the theoretical knowledge listed along the previous chapters to show an empirical application in which we first estimate two VaR models (HS and GARCH) and then we combine them through the Artificial Neural Network obtaining a third sounder model, comparing it to the inputs thanks the use of VaR Backtesting criteria.

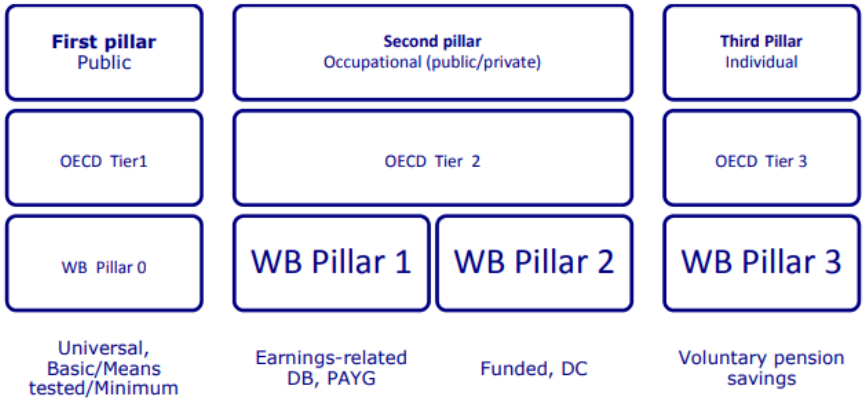
The entire process is repeated for 3 different ETFs.

Chapter 1

Pension Plans

Pension Plans are systemic plans through which individuals (referred as *members*) can accumulate resources during their working life and benefit from an income in the future retirement period. They are part of what is called *three-pillar pension system*, including the compulsory Social Security Plan (First), Group Pension Plans (Second), Individual Pension Plans (Third). The second and the third are the private pension solutions. While anyone can join an Individual Pension Plan, the Group Pension Plan covers subjects sharing common features such as same employer or same economic sector.

Figure 1.1: Different types of retirement income provisions



Sources: OECD (2005b, 2013), World Bank (1994)

These private solutions are extremely widespread in the Anglo-Saxon countries, for instance in 2008 according to the department of labour of the United States, there were 679.000 private retirement plans covering 117 million participants and managing 5 trillion dollars of assets.[2] Especially in European countries, after many reforms of the pension system to preserve the financial stability of public finances, now private solutions are developing to integrate the state pension. Indeed this one, being unfunded (also called *pay-as-you-go pension plan*) is based on the fact that benefits currently paid are

[2] D. B. Loeper (2008), The four pillar of retirement plans

not accumulated in a personal account but are used to pay the current benefits of retired members. In the past this solidarity combined with Defined Benefit plans showed to be unsustainable in the long run forcing legislators to reform pension systems.

The main distinction is between Defined Benefit (DB) and Defined Contribution (DC) pension plans.

1.1 Defined Benefit (DB)

In the former, a rule is set up for the benefit definition (usually a percentage of the member's salary during the working period) and the contributions are then computed in order to achieve the balance. The balance can be realized on an individual basis, producing an arrangement similar to the one of a life insurance contract with fixed benefits. The actuarial balance would be as follows:

$$Prem(0,r) = Ben(0,+\infty) \tag{1.1}$$

where $Prem(0,r)$ represents the expected present value at time 0 of the contributions in the interval up to retirement while $Ben(0,+\infty)$ the expected present value of the benefit that will be paid to the individual. In this framework, assumptions are required relatively to interest and mortality rates to discount future contributions and benefits. Several risks arise for the provider, particularly the investment and longevity risks. We note that (2.1) implies the accumulation of a fund used after retirement for paying out the benefits. Thus, at time *t=0,1, 2, ... r-1*, the following balance must be fulfilled:

$$Prem(t,r) + V\mathrm{t} = Ben(t,+\infty) \tag{1.2}$$

While if the balance is realized on a group basis, the condition to be satisfied becomes:

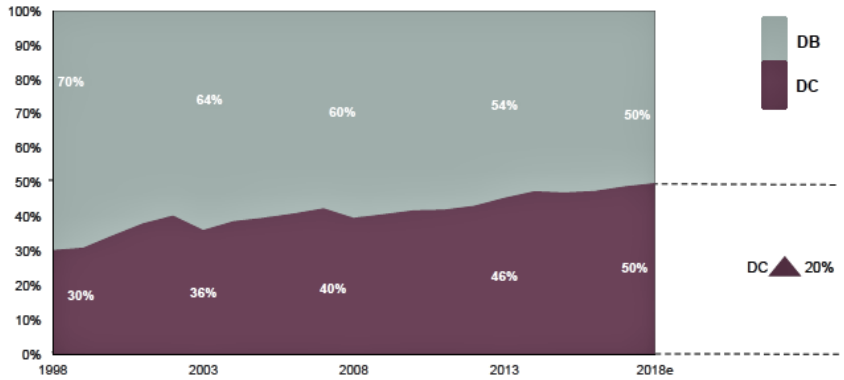
$$Prem^{[P]}(t,t+T) + V_t^{[P]} = Ben^{[P]}(t,t+T) \tag{1.3}$$

Where the values are expressed in terms of a portfolio of several positions held by the pension fund and not on an individual basis. It is important to note how usually (1.3) implies less guarantees than (1.1).

1.2 Defined Contribution (DC)

In the Defined Contribution (DC) case a rule is defined for the calculation of contributions (usually a portion of the member salary) which cumulate along the years in an individual account used to pay the pension income at retirement. In this type of arrangement no guarantee is implied unless the underwriting of ancillary benefits providing for instance capital protection or death benefits. In recent years DC pension plans have become more popular given their higher degree of flexibility and the lower risk for the provider
Nowadays, slightly more than 50% of retirement plans' assets in the seven main markets for asset allocation (Australia, Canada, Japan, Netherlands, Switzerland, USA and UK) belongs to Defined Contribution plans, showing an evidently increasing trend.[3]

Figure 1.2: DB vs DC Plans



Source: Thinking Ahead Institute

Defined Benefits presents an investment risk borne by the provider that, if for instance the asset take a hit on the stock market, has to make a shortfall

[3] R. Urwin, T. Hodgson, B. Collie, L. Yin, M. Hall (2019), Global Pension Assets Study. Thinking Ahead Institute - Willis Tower Watson

over the operating account. Defined Contribution plans transfer the risk on participant being more predictable for the employer that will just have to pay administrative costs and contributions to workers. Another important difference is that DB plans are not transferable (so if the employee leaves the company, his/her account cannot be transferred) unlike the DC.
In conclusion, we can summarize the different features of the two plans in the following table:

Figure 1.3: **Defined Benefits vs Defined Contributions differences**

DB plans	DC plans
Specifies the benefit to be received	Specifies the contribution to be made
Assets are usually pooled together	Assets are held in individual accounts
Investment risk borne by the employer	Investment risk borne by the participant
Unpredictable cost for the employer	Predictable cost for the employer
Costly to administer	Less costly to administer
Provision for past service	No provision for past service
Not portable between employers	Portable between employers

Source: J. Robbins (2014)

The provider of the service can be the employer itself (and in this case it would be forced underwrite a an insurance contract in order to hedge the risk) or, in the more common solution, it can a third institution specifically created to manage the large resources and called Pension Fund. It was estimated that the Pension Funds of the 22 main markets around the world managed assets for more than 40.173 billion USD in 2018.[4] It hence becomes of primary importance understanding the principles ruling their investments.

1.3 Pension Funds' Investment policy

Thanks to the flexibility allowed by the DC pension scheme, members can decide how to allocate the resources according to their preference and risk

[4] R. Urwin, T. Hodgson, B. Collie, L. Yin, M. Hall (2019), Global Pension Assets Study. Thinking Ahead Institute - Willis Tower Watson

appetite. In particular, it is known how younger subjects tend to concentrate on stocks in order to maximize the return, while members closer to retirement adopt more conservative approaches focusing on bonds (the so-called "lifestyle investment strategy"). Furthermore, each country presents its own historical features: for instance US and Australia have higher investments in equities than the rest of the market, whilst Japan, Netherlands and Switzerland have higher allocation to bonds. Following the classification designed by R.Ferri (2012), we can identify four life phases of investing in retirement:[5]

- **Early Savers**: subjects ate the beginning of their careers (between 20 and 39 years old) and often they do not have to take care of a family. Usually they open a saving plan with a few assets and the pension funds will propose them to be aggressive because the risk for them is overextending.
- **Mid-Life Accumulators**: types of investors that have a good career and a family to think about. With an age between 40 and 59, they usually own also cars, houses or have children. They tend to reduce their exposure to risk.
- **New Retirees**: they are people close to retirement (age 60-75). The distribution of wealth in the form of pension income is near to begin so their risk tends to be very low.
- **Mature Retirees**: they are the fully retired investors or people not as active as they used to be on their job. Having other needs with respect to the common investors, they range from long-term cares to estate planning issue. Sometimes this stage decisions are taken also with their children or their family members. Money for heirs or charities are common at this stage.

The ultimate objective of a pension fund is to obtain a competitive rate of return on portfolio assets balanced with prudent investment rules. The responsibility is to provide retirement benefits for members, retirees and beneficiaries but without the complete assurance of incremental returns because of the performances of assets classes, asset allocation and the market.

[5] R. Ferri (2012), All you need to know about ETFs. Don Phillips

There are some general policies that the institutions which manage retirement plans have to follow and the most important are:

- **Investment Diversification**: it is required in order to minimize the risk of large losses. It must be prudent and in accordance with the objectives of asset allocation.
- **Safety of principal and risk limitation**: this policy regards in particular equities and fixed income investment. Financial risk must be managed because the default (for fixed income) and price fluctuation (for equity) can compromise the financial integrity of the client's portfolio. This can be done paying attention to diversification, issuers and maturities.
- **Maintain portfolio in highly marketable assets**: it can allow the fund to have unforeseen cash requirements and basic restructuring of the portfolio if the institution needs to change it overtime.
- **Asset allocation:** the fund or the financial institution must determine its portfolio as a composition of capital market theory, financial and fiduciary requirements and liquidity needs, considering also the type of liabilities. Each class of assets has a set of risk and returns correlated with their characteristics. Changes in asset classes should happen not very frequently as short-term market fluctuations may erode the asset mix even if some rebalancing might be necessary during the fund life cycle.

Retirement funds must appropriately follow all the investment policy and guidelines as well as make a final decision pertaining the investment of the assets including implementation and compliance. Funds must maintain high quality investments particularly on quality, freely tradable and liquid investment. Performances are generally reviewed quarterly at the fund level to determine the practicability of the investment targets.

Given all these characteristics it is evident the difficulty of the portfolio selection in order to guarantee acceptable returns but diversifying the investment to maintain low levels of risk. In recent years, one of the main solutions appreciated by Pension Fund managers also for its cheapness has been the inclusion of Exchange Traded Funds in their portfolios. Nowadays the average pension fund holds 32% of its investments in passive products, in

the form of index funds and ETFs while 66% of pension fund managers regard passive investments as an established, mature part of their portfolio.[6]

1.4 Pension Funds' Risk Management

Pension supervisory authorities have been following worldwide other financial sectors in moving towards a risk-based approach to pension supervision. This can be identified as an organized process aiming to recognize the most critical risks faced by each Pension Fund and, through a focused review by the supervisor, assessing the management of those risks considering also the vulnerability to potential adverse events.[7] One of the key objectives of risk based supervision is to ensure sound risk management at the institutional level taking into account both the accuracy of the risk assessment and the quality of the handled risk. A risk-based regulation often allows Pension Funds a freer range of investments than a strict rule-based approach (even though some quantitative limits may be applied). Risk Management frameworks may be defined as the process designed to provide a decisive assurance regarding the achievement of objectives in terms of effectiveness and efficiency of operations; reliability of financial reporting; and compliance with laws and regulations.

The OECD Core Principles of Occupational Pension Regulation (OECD 2004) state that: *"Pension entities should have adequate risk control mechanisms in place to address investment, operational and governance risks, as well as internal reporting and auditing mechanism."*

In general, the broad risk management requirements among the different financial sectors are quite similar. It is hence unsurprising that the detailed guidance provided by several international authorities are comparable and fundamentally based on four categories (with guidance for how to implement each aspect):

[6] DWS (2018), Passive Investing: reshaping the global investment landscape

[7] F. Stewart (2010), Pension Funds' Risk Management Framework: Regulation and Supervisory Oversight. OECD Working Paper No. 40

- Management Oversight and Culture
- Strategy and Risk Assessment
- Control Systems
- Information, Reporting and Communication

Focusing for the purpose of this study on the risk related to the investment strategy, being it the major challenge for any fund, we can see how the OECD Guidelines on Pension Fund Asset Management (OECD 2006) provide details on Investment objectives, Asset allocation, diversification, use and monitoring of derivatives, Asset Liability Matching targets, performance measurement and risk monitoring procedures.

Especially after the financial crisis the sudden turnaround, for the Anglo-American Pension Funds in particular, from surplus to deficit served as catalyst for calls for "better risk management" of Pension Funds introducing analogous tools to those applied in other sector of financial industry such as securities firms and banks.[8] Nowadays, Pension Funds estimate Value-at-Risk (VaR), apply risk budgeting concepts and analyse fat-tails. Also Asset-Liability-Management (ALM) is routinely applied as a tool for strategic risk management.

In the following chapters, we are going to focus on Value-at-Risk as the main measure indicating the riskiness of Pension Funds' investments.

[8] D. Franzen (2010), Managing Investment Risk in Defined Benefit Pension Plan. OECD Working Paper No. 38

Chapter 2

Exchange Traded Funds

Exchange Traded Fund (or ETF) is a quite new instrument, introduced in Canada in the early 1980s and in the US in the 1990s, that is showing remarkable growth rate in the last decade. ETF can be considered as a sort of Index Fund since it presents the same objective: it tries to provide the investors with a benchmark return having minimal cost since while indexes are very expensive, ETFs are often commission-free. When investors buy a share of an ETF, they buy a share of a portfolio tracking returns and yields of the underlying index. In fact, usually ETF's aim is to replicate the performances of the index rather than outperforming it.

As Seddik (2006) affirmed '*They do not try to beat the market, they try to be the market'*.[9]

However, exactly like in conventional index investments, they allow investors to decide to be active or passive. They can hence decide the composition of their portfolios using plain-vanilla ETFs offering different exposures to Bonds and Stocks or they can choose to combine more sectors of ETF. They also have more advantages than disadvantages over traditional mutual funds. The advantages are in the low fees of the investment and the greater flexibility. Disadvantages are more related to the pricing since the intraday pricing might be overkill because of the large bid ask spread and much higher cost with respect to a standard stock. In general, the practice is to compare ETF to mutual funds so that the cost appears to be low, while actually it is not. This is particularly due to the fact that, even if they seem to be very simple instruments, they present a more complex operating structure that requires deep investment analysis.

[9] A. Seddik (2006), Exchange Traded Fund as an Investment Option. PALGRAVE MACMILLAN.

2.1 Development of an ETF

ETFs have a peculiar creation process that makes them very different from traditional mutual funds. Developing an ETF involves decisions for the company that wants to manage it. Usually those who want to implement this process are called Authorized Participants (AP) and they can be institutional investors, specialists or market makers. To obtain the authorization the must fill and agreement specifying what kind of ETF sponsor or distributor they want to be.

An institution can be interested in issuing ETFs for several reasons. For instance, the APs benefit from arbitrage opportunities resulting from tracking errors or differences between the underlying prices of the securities making up the ETF and the ETF's share itself. This is possible because the participants can sell the created ETF on a retail basis in smaller increments. They might also be interested in creating ETFs to introduce liquidity on the market.[10]

In the product development, the first thing is what kind of market exposure will be offered by the ETF and then, the method of exposure for clients. Probably the most important step of the process is the definition of the basket of shares included in the fund's underlying. This basket should be transparent, liquid and easy to trade. The ETF has to trade close to its underlying Net Asset Value (NAV).

The ETF's price point is important for product positioning and it starts determining the amount. Then the efficiency of the basket's constituents and average trading volumes must be evaluated. When the required shares match the Creation Unit, they are delivered to the issuer with the cash component and the issuer delivers them back right after.

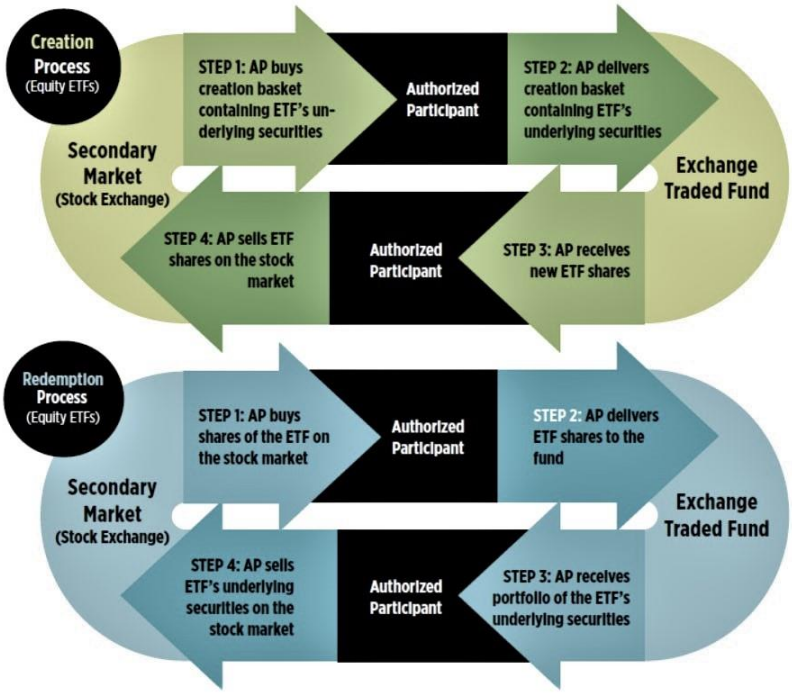
In practice, the issuer does not keep the shares but it has to deliver them back to the AP as part of the creation process and in this way, it creates new ETF shares.

[10] D. J. Abner (2010), The ETF. John Wiley & Sons, Inc.

The redemption process works in the opposite way: the AP receives a basket of stocks from the issuer and the issuer receives the shares of the ETF from the AP.

The creation process is entirely possible because the shares are fungible vehicles, interchangeable for each other. Furthermore, it removes the trading expenses from ETFs when there is growth or decline in the asset and it arises for its tax efficiency because investors are able to divest the portfolio without trading in the market (in this way generating a taxable event).

Figure 2.1: The Creation/Redemption Mechanism



Source: ETF.com

The main part of positive characteristics of ETFs derive from their creation process. Even though tax efficiency and trading expenses have been already mentioned, it is relevant to stress that the pricing and development is different from the main competitors such as Closed-End funds and Mutual funds.

The formers issue shares through an Initial Public Offering (IPO) listing on the stock market and if an event will cause the liquidation of the fund, the shares will remain listed on the exchange.

Instead, Mutual funds never list their shares on the exchange. They take cash directly from clients issuing the shares directly to them. In case of redemption,

the Mutual fund can return cash to costumers deleting the shares. Thus, summarizing the differences between ETFs, Closed-End Funds and Mutual Funds we have:

- ETFs do not trade into the basket of securities since they deliver them via AP
- Closed-End Funds use IPO to collect the resources used to trade in the market
- Mutual Funds issue every time new shares creating a taxable event

2.2 Pricing and Valuation of ETFs

Exchange Traded Funds, unlike the Mutual Funds, are traded through the day like a common stock hence computing their performance may be more complex. Furthermore, it is important to notice how an arbitrage relationship is made up by the creation and redemption mechanisms and it is often used to make some strategies. [11]

Below we will describe the valuation of an ETF through Intraday indicative value of the fund and the Net Asset Value (NAV) and the different kinds of structure of the product.

2.2.1 Domestic Constituent

ETF calculates the NAV every day based on the closing price of all the underlying assets and an actual accounting of cash inside the fund at the time of computation. It is possible also to compare the result to other funds in order to obtain the performances. The NAV will be:

$$NAV = \frac{A_j - L_j}{S_j^0} \qquad (2.1)$$

where:

A_j is the value of j assets for j = 1, ..., N

L_j is the value of j liabilities for j = 1, ..., N

[11] E. Hehn (2005), Exchange Traded Fund, Springer

S_j^0 is the value of j share outstanding for j = [1, ..., N]

The simplest way to calculate the NAV of an ETF is to take as a unit the Creation Unit (hereafter, CU) and the total cash published daily:

$$NAV = \frac{\sum_{i=1}^{N}(S_i\, w_i p_{ti})}{\frac{C_u S_i + C_t}{C_u S_i}} \qquad (2.2)$$

where:

S_i is the value of i different kind of stocks where i = 1, ..., N

w_i is the quantity of share per each component of stock i

p_{ti} is the most recent price at time t for stock i with t = 1, ..., N

C_u is the Creation Unit

C_t is the amount of total cash at time t

Thus, the NAV will be represented in the share price. The CU is a set of shares composed by a unit of the fund held by the trust underlying the ETF. The CU shows the basket to be delivered to the issuer in order to receive ETF shares during the creation and it also shows the basket of assets that would be received by an AP whenever it decides to make a redemption.

Many data providers such as Bloomberg Professional © can show the constituents of all ETFs available in the market, with a section in which we can see the number of shares per CU and their market value at the closing price of the stocks. In order to compute the ETF's value, we can just use the CU and then create also all the model around the price of the ETF. Some misconceptions can arise from the NAV concerning the discount and the premium calculation for the fund. For instance, in a Closed-end Fund and in ETFs the premium and discount factors have completely different meanings. There is a structural inconsistency in the discount and premium pattern which is very short-lived and a long-term pattern in Closed-end Funds. When we deal with domestic constituent the situation is considered in normal circumstances because premium and discount arising between NAV and trading price is the result of late market activity at the end of the trading day,

so it will narrow o the next market opening. The relevant fact that must be stressed is that with international stocks and constituents the baskets are essentially traded on expected valuation, they hence usually trade away from the NAV. There are many other circumstances that can push the ETF value away from its NAV determining either a premium or a discount if compared to the underlying basket of traded assets.

2.2.2 International Constituent

Many ETFs presents some international underlying stocks. The main difference between international constituents and domestic constituents is that the second is also a function of timing and currencies. If the underlying trade in a time zone different from the one of the ETF, the intraday value will remain fixed for the equity portion but will change relatively to a spot foreign exchange rate.

This type of funds is very important for investors aiming to diversify their portfolios. The computation of the NAV is very similar to the domestic constituent case and the CU as well as the total cash and number of shares are used:

$$NAV = \frac{\sum_{i=1}^{N}(S_i\, w_i p_{ti})}{\frac{R_f C_u S_i + C_t}{C_u S_i}} \qquad (2.3)$$

where:

R_f is the currency rate of the country f

and the other components are described in equation 2.2

We can notice from the formula that also the currency conversion has been taken into account.

The International ETF execution can be described through an example:

A US customer wants to receive a NAV execution in Dollars of an ETF with Japanese constituents.

- Customer gives to AP an order to buy 1 Million dollars of an ETF with Japanese constituent
- The AP buys the basket of Japanese constituents at the close of trading the following day in Japan.

- The AP has to borrow the Yen in order to create the basket and buy the Japanese basket
- At this point the AP can delivers the ETF to the agent and the payment of it will be in US dollars
- Now the AP have an up-and-down position in currencies because it results to be long in US dollar and short in Japanese Jen. The AP will buy Yen with the Dollars it receives from the client and will repay the loan given by the bank

This currency transaction is extremely important since in this way the price of the ETF is determined. On the basis the decisions taken by the AP considering all the variables as showed in the previous example, the price of the ETF will be done. This feature is important because it can show how the domestic constituent can have different prices with respect to the international constituent.

2.2.3 Fixed Income Constituent

The Fixed-Income field has been one of the most interesting in which the ETF makers wanted to work in the past but, at the same time, the hardest to get into. The first Fixed Income ETF has been developed by a small start-up called ETF Advisors. Nowadays, it is used by investors but still not so much consolidated as there are many problems arising from the structure of Fixed Income:

- **Price:** the pricing for Fixed-Income market is very different from normal stocks as they are usually traded on Over The Counter (OTC) and hence there is no official open or close. Furthermore, it must be considered the variability of market prices and the bid-ask spread that creates huge problems since it is inversely related to the bond's liquidity. Since ETFs require the closing price to estimate the NAV, ETF makers are forced to adapt in a context without exchange pricing.
- **Bonds vs Stocks:** while stock prices fluctuate based on supply and demand for the company shares, bonds are debt instrument with a stated term structure and maturity. Bonds prices do not fluctuate as stocks hence investors do not have the same benefit that equity market provides as there

is not the same accurate valuation for Treasuries, MBSs, Corporate Bonds and other Fixed Income assets facing the same risk.

- **Market:** another feature that must be taken into account is that the Fixed Income market is dominated by the institutional community and it is usually very large. This creates new problems regarding the bid and ask spread for ETFs on Fixed Income securities.

It is hence possible to conclude that this type of instrument is more complex if compared to other types of ETFs and this is why it is still developing and trying to overcome several problems, especially in pricing.

2.2.4 Commodity, Inverse, Future ETFs

Commodity and Inverse ETFs are used, in general, for statistical arbitrages in the investing community. They are commonly more short-term oriented products needing tactical strategies. They can also be used to hedge avoiding the use of futures roll. Commodity ETFs are an alternative way to access the market of commodities and they are called Exchange Traded Commodities (ETC). Their constituents can be a category of single commodities or baskets of commodities created according to several strategies and trading models. The main categories are usually:

- Agriculture
- Energy
- Physical commodity
- Metal
- Futures tracking single commodities
- Futures tracking baskets of commodities
- Equity with various forms of exposure to commodities

They can present in the basket holdings either domestic or international, with various weighting schemes providing exposure to companies with commodity-related activities.

Inverse ETFs usually hold swaps and futures to achieve their exposure. They are designed to profit from the decline of an underlying benchmark. Investing in this kind of product is similar to holding various short positions, which involves borrowing securities and selling them with the aim of repurchasing

them at a lower price. For this reason, they are also known as "Short ETF" or "Bear ETF."

Futures ETF are used as a preferred alternative to Futures on ETFs for many reasons. First of all, because they are a more flexible option being a non-derivative instrument. Secondly, Futures ETFs are bought by funds because it is not convenient to buy every single component of the index. Thirdly, ETFs enable the investors to take long and short position at the same time with less constraints than futures (also, being derivative instruments not all the investors are allowed to deal with them). Another main difference to stress is that while the use of Futures needs to have detailed legal documentation and a margin account as collateral, all these complexities are not necessary for ETFs. The ETFs used to replicate the Futures following the index has to roll continuously because otherwise the position would end at the expiration date. It hence replicates an index which has a future inside with short term expiration. Once the contract gets to maturity the ETF sells and buys in order to maintain the position. This mechanism has some complications due to the presence of Backwardation and Contango in Futures that may sharpen the tracking error on prices.

2.3 Advantages and Disadvantages

If compared to traditional Mutual Funds, ETFs provide new advantages for investors including, flexibility, transparency, lower operating costs and tax efficiency. At the same time, they present some drawbacks such as trading costs, complexity and tracking error. We are going to analyse in detail these characteristics in order to understand the importance of this new products.

2.3.1 Advantages

- **Flexibility:**

With Mutual Funds and Open-End Funds the investors wait the end of the day to know what price they paid for the new shares they sold. Even though this characteristic may appear irrelevant in the long term, for a short-term investor it is very important. In fact, ETFs are bought and sold during market day with

intraday variations so investors can be aware of how much they are paying or receiving in the same moment they trade in the market.

- **Risk Management:**

Since ETFs cover a variety of sectors and countries, they make easier to invest in many market segments. Buying an ETF make the portfolio already diversified with a positive effect on fees and risk management.

- **Cost:**

The requested costs are usually management fees, custody costs, administrative costs, marketing expenses, distribution expenses. Furthermore, Mutual Funds and Open-End Funds have to make monthly statements which are costly and charged to clients. An ETF overcome this problem since investors who decide to put money on it are buying a fund with all features included in a stock. The body responsible for these costs is the AP rather the ETF companies, making them cheaper than normal funds.

- **Tax Advantages:**

ETFs are tax efficient thanks to their low level of turnover, as they only make buy and sell orders to adjust the underlying benchmark. Both Mutual and Closed-End Funds try to increase their performances and reducing at the same time risk. This characteristic made them more taxable because of the higher capital gains distributed to investors. Another important feature that makes ETFs very attractive in terms of taxation is the fact that they are no-taxing entity. In US, for instance, there are two kinds of taxation: corporate income and individual income (on distributed dividends). Usually investment companies may not be subjected to the same taxation of corporates since they are not considered regulated investment companies if they distribute enough income to shareholders. The limit is 98 % of required distribution and for the ETFs is quite easy to realize. Almost all the available ETFs are not subjected to federal income so if the AP is able to manage them well, they can be considered tax-free entity. Finally, it must be considered that through the process of creation and redemption the unrealized capital gains can be erased.

2.3.2 Disadvantages

- **Brokerage Fees:**

Since ETFs are traded like stocks, investing on them will result in higher brokerage fees and commissions. Those who invest in ETFs through a brokerage firm will have higher trading costs while for those who have no-load funds directly will reduce them. Investors with a fund company which manages the money cannot buy ETFs on their own. They can only open an account and pay fees every time the broker decides to move the money. Brokers uses indexes as instrument designed for long-term investment securities and this implies frequent trading producing large commissions to be paid by the customer. While there is a prohibition to most of Mutual Funds of buying and selling in a relative short period, this is not applied to ETFs. Staying with a no-load open-end fund is better under this scenario.

- **Tracking Errors:**

The problem arises as the AP is the entity that must guarantee that the fund investment performances are in line with the tracked indexes but there are many cases in which this does not happen and it may be extremely costly for investors. When the price deviates from the NAV there can be arbitrage opportunities for traders. Since indexes do not hold cash while the ETFs do, the tracking error is expected for sure in an ETF. Furthermore, dividends are treated in a different way in the indexes and in the ETFs. Indexes attempt to reinvest the dividends the same day the company releases them, ETFs cannot reinvest the dividends in other securities, so ETFs hold the cash until dividends are effectively paid. Because of this problem the ETF would never be able to perfectly track the index. The same happens for the ETFs with currency and with ETFs on futures. Almost all the type of ETF has inside problem related to tracking errors. They may create higher portfolio turnover and increasing of costs.

2.4 Regulation in Europe

ETFs development in Europe have a very clear regulatory policy if compared to the one in the United States. Companies interested in introducing the product in the market have to design it with a less restrictive framework than the American Company Act 1940. European aim is in this case to simplify and harmonize the investment company the investment company regulation before ETFs accumulate too many assets in US rather than Europe.
In order to introduce an ETF in the European Union the AP needs a *passport* allowing the AP itself to meet the requirements of the home country regulator. In fact, in Italy the development and control of ETFs is made by Borsa Italiana. The European regulation acts by the Undertaking for Collective Investment in Transferable Securities (UCITS) which provide fund diversification requirements.[12] ETFs must comply with policy of every member state in which are sold. Since this can be a huge disadvantage for the spread of this product in Europe, Barclays asked the European Commission to create a unique European regime. European laws currently do not allow the fund to create their own ETF and give to them some limits on the possible investment amount. In Europe, there are several exchanges that compete for ETF activities. Despite the existence of UCITS which is working to facilitate the marketing and distribution of ETF in the country, ETFs are still primarily developed in the Deutsche Börse.

2.5 Selected ETFs

For the purpose of this study I selected three ETFs considering as key criteria the Asset Under Management (AUM) and the variety of assets held in order to reproduce as much as possible the portfolio diversification strategy implemented by Pension Funds. Thus, the choice fell on iShares Core MSCI

[12] ESMA (2014), ESMA/2014/937EN, Guidelines for competent authorities and UCITS management companies

World UCITS ETF, iShares Core MSCI Europe UCITS ETF and SPDR S&P 500 ETF.

- **iShares Core MSCI World UCITS ETF (SWDA)**

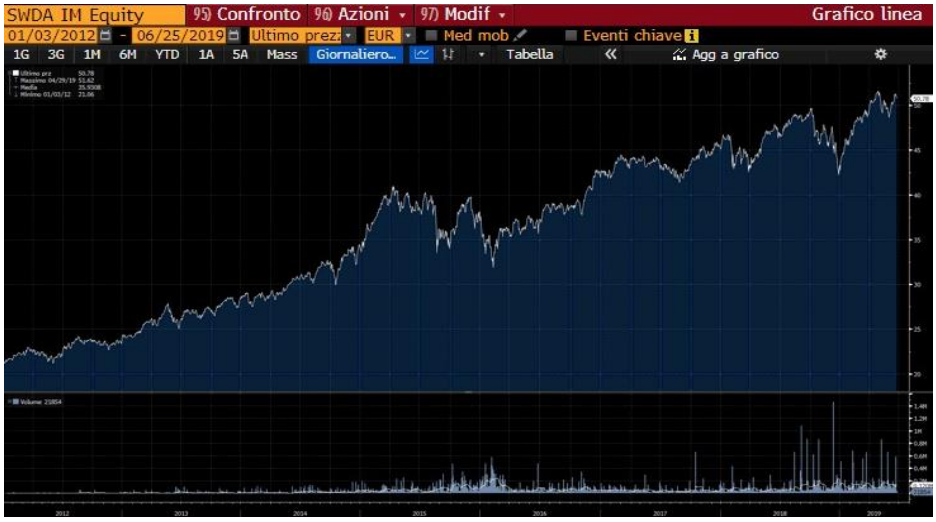


Figure 3.2 Source: Bloomberg

With an amount of Net Asset Fund equal to USD 19.578.284.637 this fund issued by iShares (managed by BlackRock, Inc.) aims to replicate the performances the composite index of the companies located in the developed countries. The exposure is hence towards an international and developed market with remarkable levels of stock liquidity and safety of the investment. Indeed, the benchmark index is MSCI World Index. The sectors presenting the main levels of exposure are: IT (15,97%), Financials (15,76%), Health Care (12,53), Industrials (11,10%). This product is listed on Borsa Italiana, London Stock Exchange, Euronext Amsterdam, Bolsa Mexicana De Valores, SIX Swiss Exchange, Deutsche Boerse Xetra.

- **iShares Core MSCI Europe UCITS ETF (IMEU)**

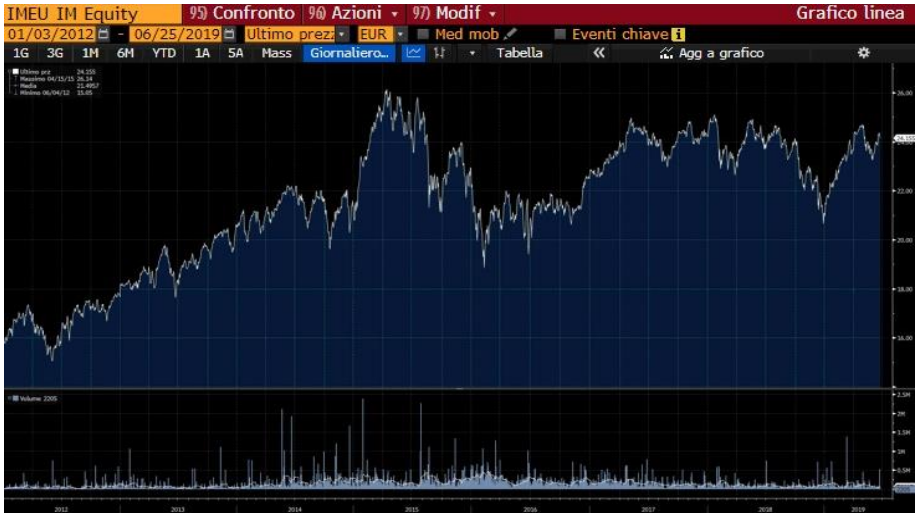


Figure 3.3 Source: Bloomberg

With an amount of Net Asset Fund equal to EUR 4.971.984.906, also this fund is issued by iShares providing a diversified exposure towards the main European corporations. The sectors presenting the main levels of exposure are: Financials (18,04%), Consumer Staples (14,82), Industrials (13,13%), Health Care (12,72%). This product is listed on Borsa Italiana, London Stock Exchange, Euronext Amsterdam, Bolsa Mexicana De Valores, SIX Swiss Exchange, Deutsche Boerse Xetra.

- **SPDR S&P 500 ETF (SPY)**

Figure 3.4 Source: Bloomberg

SPY is the best-recognized and oldest ETF and typically tops rankings for largest AUM and greatest trading volume. The fund tracks the massively

popular US index, the S&P 500. It is important to stress that S&P's index committee chooses 500 securities to represent the US large-cap space—not necessarily the 500 largest by market cap, which can lead to some omissions of single names. The ETF is issued by State Street Global Advisor.

Chapter 3

Value-at-Risk

In financial markets three main risks are present: Credit Risk, Liquidity Risk and Market Risk. The latter, for trading purpose and in particular considering the ETF-investing Pension Fund of this study, is the most important to be considered. Market Risk can be defined as the possibility of an investor to experience losses due to factors affecting the overall performance of the financial markets in which he or she is involved. Market Risk is also known as "systematic risk" and cannot be eliminated through diversification but hedged in other ways. This is in contrast with the so-called "unsystematic risk" which is unique to specific companies and can be reduced through diversification.[13] The main types of Market Risk are Interest Rate, Commodity and Currency Risk. However, the most important problem is to define the concept of risk in a quantitative measure. Within financial institutions the concept of Value-at-Risk (VaR) started to spread and now, despite its well documented flaws that will be described later in this chapter, it is probably the most common measure of risk available in the financial industry.

Introduced for the first time by J.P. Morgan in 1996, the VaR is defined, given a portfolio, time horizon and probability p as the maximum possible loss during that time after excluding all worse outcomes whose combined probability is at most p.[14] We hence have a probability of $(1-p)$ that the loss on a given trading day will be lower than the VaR.

The most common probability levels are 1% and 5% but can vary in practice. VaR is the quantile on the distribution of Profit & Loss (P/L) so letting y be the daily log-return for a given asset held in a long position, we have that:

[13] James Chen (2019), Market Risk. Investopedia

[14] Jorion, Philippe (2006). Value at Risk: The New Benchmark for Managing Financial Risk (3rd ed.). McGraw-Hill

$$\Pr[y \leq -VaR(p)] = p \qquad (3.1)$$

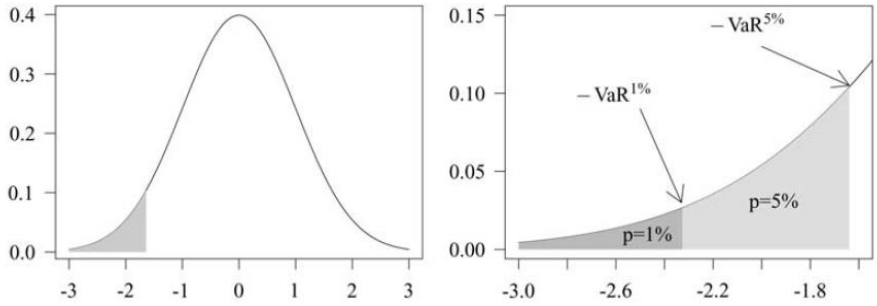
or

$$p = \int_{-\infty}^{-VaR(p)} f_q(x)dx \qquad (3.2)$$

For the holder of a short position, the loss will be in case $y \geq 0$ so the VAR is defined as:

$$\Pr[y \geq -VaR(p)] = p \qquad (3.3)$$

Figure 3.1: P/L density and VaR - Left Tail and VaR



Source: J. Danielsson (2011)

If the CDF is known, the VaR is simply its *p*th quantile. Since the Cumulative Density Function is unknown in practice, studies of VaR are basically focused on the estimation of CDF, especially its tail behaviour. In general, the computation of VaR involves the following elements:

1. probability *p*;
2. time horizon *t*;
3. CDF denoted as F*(x)* or its quantile.

It is easy to recognize the importance of F*(x)* in the VaR modeling.

This instrument has achieved a great popularity is essentially because of its conceptual simplicity: VaR can basically reduce the (market) risk associated with any portfolio to just one number representing the loss associated to a given probability. Value-at-Risk can have applications both in the risk management and for regulatory purposes, in particular the Basel Committee on Banking Supervision (1996) imposes to financial institutions to meet capital requirements using VaR. Providing accurate estimates is of

fundamental importance since, if the risk is not adequately assessed (both overestimated or underestimated) this may lead to a sub optimal capital allocation damaging financial institutions' profitability and stability.

While VaR is a very easy and intuitive concept, its estimation is an extremely challenging statistical problem following a common general structure that can be summarized in three points:

1. Mark-to-market the portfolio
2. Estimate the distribution of portfolio returns
3. Compute the VaR of the portfolio

The main issue distinguishing the different methods is related to point 2. Regarding this point, we can classify the existing models into two main categories (even though the number and types of approaches to VaR estimation is growing exponentially involving in particular Monte Carlo Simulation, Stress Testing and Extreme Value Theory):

- Nonparametric (Historical Simulation and Monte Carlo Simulation)
- Parametric (Risk Metrics and GARCH)

These methods may lead to very different results. T.S. Beder[15] applied eight common VaR methodologies to three portfolios and the results showed that the differences might be very large, with estimates varying by more than 14 times for the same portfolio. Thus, in order to choose the right methodology to apply, it is necessary to understand the underlying assumptions together with the quantitative techniques and the mathematical models used.

The fundamental inspiration of VaR methodologies usually comes from the particular features of financial data. These empirical facts about financial markets are very well known thanks to the works of Mandelbrot[16] and Fama [17] and they can be summarised as follows:

1. Financial return distributions are leptokurtotic, meaning that they have heavier tails and a higher peak than a normal distribution.

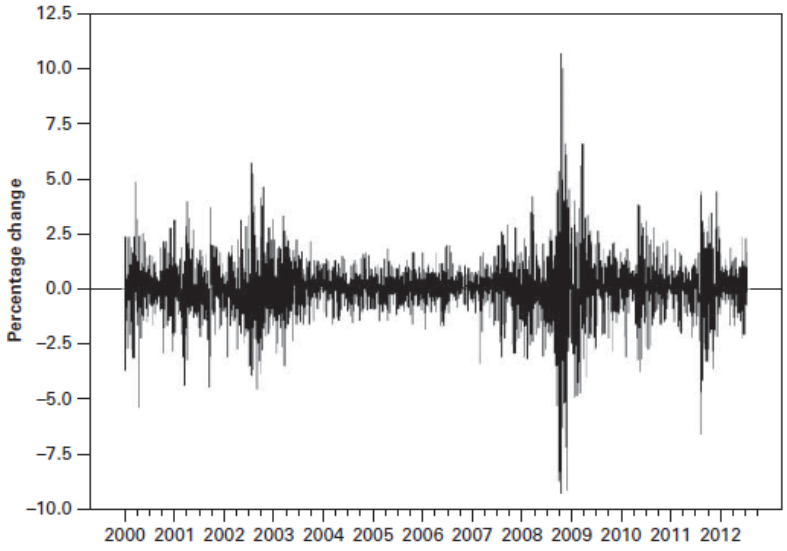
[15] T.S. Beder (1995), VaR: Seductive but Dangerous, Financial Analyst Journal, Sep-Oct, 12-24

[16] B. Mandelbrot (1963), The Variation of Certain Speculative Prices. The Journal of Business, Vol. 36, No. 4, pp. 394-419

[17] E. F. Fama (1965), The Behaviour of Stock-Market Prices. The Journal of Business, Vol. 38, No. 1. pp. 34-105.

2. Equity returns are typically negatively skewed.
3. Squared returns show significant autocorrelation, i.e. volatilities of market factors tend to cluster (as we can see from the below time series). This one is an important feature of financial returns, allowing researchers to consider market volatility as quasi-stable, changing in the long run, but stable in the short period. Most of the VaR models take into account this quasi-stability to evaluate market risk and capturing some of these empirical regularities.

Figure 3.2: Volatility Clustering in the NYSE U.S. 100



Source: W. Enders (2014)

3.1 Nonparametric Model

Very common methods for VaR computation are the Historical Simulation (HS) and the Monte Carlo Simulation. These nonparametric approaches remarkably simplify the computational procedure as they do not make any distributional assumption about asset returns.

We can go now more deeply in detail for the two methods analysing also pros and cons.

3.1.1 Historical Simulation

HS is based on the principle of rolling windows. First, it is necessary to choose an observation window (generally ranging from 6 months to two years), then asset returns within this window are sorted in ascending order and the *p*-quantile is given by the return leaving *p*% of observation on its left side and *(1-p)*% on the right side. Some interpolation rule might be required if this number falls between two consecutive returns. To compute the VaR the following day, the whole window is moved forward by one observation repeating the entire procedure.

Even though this approach does not require any specific assumptions on the distribution of returns, behind this procedure we can notice how an implicit assumption is implied: the distribution of returns does not change within the window. From this assumption several problems derive. First of all, this method presents a logical inconsistency since if all returns within the window are assumed to have the same distribution, then the logical consequence must be that all returns of the time series must have the same distribution: if $y_{t-window}, \dots, y_t$ and $y_{t+1-window}, \dots, y_{t+1}$ are i.i.d., then also y_{t+1} and $y_{t-window}$ must be i.i.d. because of the transitive property. Secondly, the empirical quantile estimator is consistent only if k, the window size, goes to infinity. The most important issue concerns the length of the window since forecasts of VaR are meaningful only if the distribution of the historical data used in the computation is roughly the same. In practice the volatility clustering period is difficult to identify. The length of the window must satisfy two opposite properties: it must be sufficiently large to make a significant statistical inference but at the same time it must be not too large in order to avoid the risk of taking observations outside the current volatility cluster. Clearly, this problem does not have an easy solution. Furthermore, let us consider a case in which the market is moving from a period of relatively low volatility to a period of relatively high volatility (or vice versa). In this case, VaR estimates based on Historical Simulation will be biased downwards (correspondingly upwards), since it will take more time before the observations of the low volatility period leave the window. This is the reason

for the characteristic box-shape behaviour typical of the HS method, presenting many jumps due to the discreteness of extreme returns. We can hence summarize the features of the Historical Simulation method as:

- **Pros:** It is straightforward in its implementation and does not require any parametric assumption. It is also very general and can be applied in many situations. Finally, there are no time-consuming parameter estimations.
- **Cons:** As we deeply investigate into the left tail, we always need a certain amount of data in order to get reliable estimates. As the time series of returns is not independent, their unconditional and conditional distributions do not coincide. This implies that, when the window size is large, the historical estimator captures the unconditional quantile, while it is practically more relevant to know the conditional one. Finally, all past returns have the same weight. If an outlier falls into the window, the subsequent estimations will include it, until it will exit the window resulting in the weird box-shaped behaviour.

A possible remedy to avoid the difficult choice of the window length and reduce the box-shaped behaviour is the so called Weighted Historical Method in which the idea is to attach higher weights to the most recent observations. Thus, fixing η for example at 0.95 we obtain the probability of each observation as:

$$p_m = \frac{1-\eta}{\eta(1-\eta^M)}\eta^m \qquad (3.4)$$

With $m = 1, \ldots, M$ as observation from the window of length M

It is important to note that p_m is a system of decreasing percentage weights. In this way the choice of the window length M becomes far less crucial with respect to the "classical" historical method, as the weights rapidly decrease. We get closer to a conditional estimate (the recent past counts more) and we get rid of box-shaped behaviours. The disadvantages of this method is that it plainly depends on η and choosing an η too low, favours recent data too much. In practice, values within the range $[0.95, 0.99]$ are most often used. At the

same time, it does not solve one of the main drawbacks of the historical method: it is insensitive to large positive returns.

3.1.2 Monte Carlo Simulation

The Monte Carlo Simulation is another member of the group of nonparametric models. It is probably the most popular approach when a powerful VaR system is needed but, at the same time, it is by far the most challenging to implement.[18]

This approach can be summarized in two steps: first, basing on historical data, it is necessary to estimate stochastic processes for financial variables considering also volatilities and correlations. Second, using these inputs it is possible to simulate thousands of times the price paths from which derive the returns and then the VaR estimates. The main strength of this simulation, like the Historical Simulation, is that no assumption about returns' distribution is required. Although the parameters are estimated from historical data, it is easily possible to add subjective judgements or other informations to improve the forecasting. The method is also capable of covering nonlinear instruments, such as options.[19] Also, it is important to underline that the Monte Carlo Simulation generates the entire distribution allowing in this way to compute also losses exceeding the VaR.

On the other side, the most significant problem deriving from this approach is the computational time as it requires a lot of resources. Indeed, the simulation converges to the true value of VaR as $\frac{1}{\sqrt{N}}$ where N is the number of simulations hence to increase the accuracy of the model by 10 times, one must run 100 times more simulations.[20] Nevertheless. Monte Carlo is continuously increasing in popularity given the growing computation power of computer programmes. A potential weakness is the *model risk* arising from

[18] K. Dowd, (1998) Beyond Value at Risk: The New Science of Risk Management. Wiley, New York.

[19] A. Damodaran (2007), Strategic Risk Taking: A Framework for Risk Management. Pearson Education, New Jersey.

[20] P. Jorion (2006), Value at Risk: The New Benchmark for Managing Financial Risk (3rd ed.). McGraw-Hill

wrong assumptions about the pricing models and underlying stochastic processes.

3.2 Parametric Models

A common, even though rarely realistic, assumption is that returns are normal. This can ease the computation since to estimate the Value-at-Risk the value of σ (volatility of returns) and μ (mean return, even though it is usually negligible for daily estimates) as the quantile of the normal distribution is easy to obtain once fixed *p*. At this point, it becomes of paramount importance the estimation and forecast of volatility. We are now going to focus on the so-called Risk Metrics (or EWMA) and on the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model to forecast future daily volatility.

3.2.1 Exponentially Weighted Moving Average (EWMA)

The first approach was introduced by J.P. Morgan's Risk Management unit Risk Metrics™ [21]which was subsequently spun off and it is now known as the Exponentially Weighted Moving Average (EWMA): similarly to the weighted historical method, we can estimate the volatility giving more weight to the more recent observations:

$$p_m = \frac{1-\lambda}{\lambda(1-\lambda^M)}\lambda^m \qquad (3.5)$$

By weighting the observations with p_m we obtain the following volatility estimate:

$$\sigma = \sqrt{\sum_{m=1}^{M} p_m y_{-m+1}^2} \qquad (3.6)$$

[21] Riskmetrics (1996). Technical Document. Technical report. J.P. Morgan

After a bit of algebra, we can obtain that:

$$\sigma_t^2 = \lambda\sigma_{t-1}^2 + (1-\lambda)\, y_{t-1}^2 \qquad (3.7)$$

Where σ_t^2 is the conditional volatility forecast on day t and λ is the decay factor. When the model was first proposed, it was suggested that λ be set at 0.94 for daily returns and this is the most common assumption. Although the methodology is quite simple, it represents the first attempt to implement VaR estimation in a consistent manner and it is still widespread in financial industry, in particular among small banks. By making distributional assumptions about residuals and setting the decay factor to 0.94 it is straightforward to estimate the volatility through this model. The main disadvantage is represented by the fact that λ is constant and identical for all assets and this is rather unrealistic. As a result, the EWMA model by definition gives inferior forecasts with respect to GARCH models, even though the difference may be rather small in many cases. Considering (3.7) it is easy to notice how the EWMA is a special case of GARCH model with the sum of parameters equal to 1.

3.2.2 GARCH model

The family of Autoregressive Conditional Heteroskedasticity (ARCH) models was introduced by Engle[22] but then generalized into the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model by Bollerslev in 1986 [23] and has been successfully applied to financial data. Given the particular persistency of volatility, the GARCH model overcome many problems of the ARCH introducing a second parameter β for the memory of the process (while α stands for the impact of the news). The benchmark GARCH (1,1) can be described as follows:

$$y_t = \sigma_t \varepsilon_t \qquad \varepsilon_t \sim i.i.d.(0,1) \qquad (3.8)$$
$$\sigma_t^2 = \omega + \alpha y_{t-1}^2 + \beta\sigma_{t-1}^2$$

[22] R.F Engle (1982), Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica 50: 987-1007

[23] T. Bollerslev (1986), ‘‘Generalized Autoregressive Conditional Heteroskedasticity,’’ Journal of Econometrics,31: 307–27.

Where the size of $(\alpha + \beta)$ determines how quickly the predictability of the process dies out: if the sum is close to zero the predictability will die out very quickly, otherwise slowly.

This model has two crucial elements: the particular specification of the variance equation and the assumption that the standardised residuals are i.i.d. The first element was inspired by the characteristics of financial data discussed above. The assumption of i.i.d. standardised residuals, instead, is just a necessary device to estimate the unknown parameters.

A further necessary step to implement any GARCH algorithm is the specification of the distribution of the ε_t . The most applied is in general the standard normal. In general, any GARCH (p,q) process can be described as:

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha y_{t-i}^2 + \sum_{j=1}^{q} \beta \sigma_{t-i}^2 \qquad (3.9)$$

The parameters of the model are typically estimated by Maximum Likelihood but since in this case the first order equation is not linear, the solution requires some sort of search algorithm exploiting various "hill-climbing" methods in order to find the parameters that maximize the Log-Likelihood function . GARCH models are more suitable in capturing volatility clustering and persistency. Furthermore, an important characteristic of both ARCH and GARCH is that, even if the standard normality of residuals is assumed, the unconditional distribution of the returns will present excess kurtosis being consistent with the "fat tails" of returns previously described. Once the time series of estimated variance is computed, the normal quantile is easily derived and we have the VaR:

$$VaR_{t(p)} = -\sigma_t \Phi^{-1}(p) \qquad (3.10)$$

General findings show how these approaches (both normal GARCH and Risk Metrics) tend to underestimate the Value at Risk, because the normality assumption of the standardised residuals does not seem consistent with the behaviour of financial returns. The main advantage of these methods is that they allow a complete characterisation of the distribution of returns and there

may be space for improving their performance by avoiding the normality assumption. Indeed, many other models based on GARCH have been developed such as Student-T GARCH, EGARCH, APARCH and so on even though the benchmark still remains the GARCH (1,1). On the other hand, both GARCH and Risk Metrics are subject to three different sources of misspecification: the specification of the variance equation and the distribution chosen to build the log-likelihood may be wrong, and the standardised residuals may not be i.i.d. Whether or not these misspecification issues are relevant for VaR estimation purposes is mainly an empirical issue.

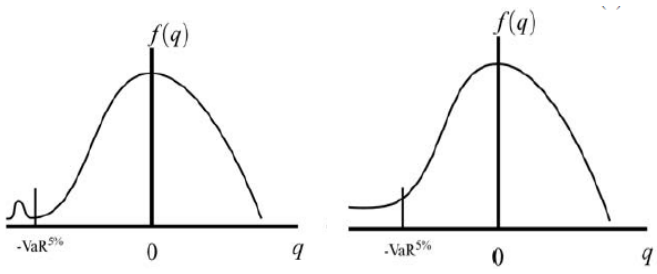
3.3 VaR Biases

Many authors strongly criticised the VaR adequacy as a measure of risk stressing in particular three main issues:

1. VaR is only a quantile on the P/L distribution:

VaR is estimated as the minimum potential loss that a portfolio can suffer in an adverse scenario but the problem is that this measure just provides the "best of worst-case scenario" and inevitably underestimates potential losses. If, for instance, we have a daily VaR at 95% confidence level this measure is incapable to capture what is beyond the 95%: extreme movements with lower probability.

Figure 4.3: VaR in unusual cases



Source: J. Danielsson (2011)

2. VaR is not a coherent risk measure:

Artzner et al. studied the properties that a risk measure should have in order to be considered a sensible and useful risk measure identifying four axioms

that a risk measure should adhere to: Monotonicity, Subadditivity, Positive Homogeneity and Translation Invariance. While the Positive Homogeneity is sometimes violated in practice, the most relevant issue is related to Subadditivity which is not satisfied unless in the case of normality where VaR is proportional to volatility, which is subadditive.[24] In practice, most assets do not have tails so fat that subadditivity is violated (e.g. equities, exchange rates and commodities) but there are some assets exposed to very large but infrequent negative returns for instance countries pegging their currency but subject to occasional devaluations, electricity prices and defaultable bonds.

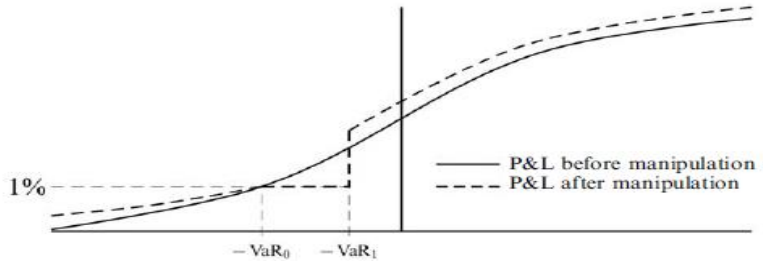
3. VaR is easy to manipulate

Another important weakness of VaR is how easily it can be manipulated. Since it is only a quantile on the distribution of profit and loss, a financial institution will often find it easy to move the quantile around manipulating the VaR. An easy way to lower the VaR is to reduce holdings of risky assets, but it can equally well be lowered by using simple trading strategies involving options. In this case, VaR could be lowered at the expense of overall profitability and even by increasing downside risk. Thus, the risk reduction implied by lower VaR is only an illusion: reported risk is reduced, but actual risk increases and profits decrease. One example of how this could be done is provided by Danielsson[25] who demonstrated how the use of put options can deliver any VaR desired. Suppose the VaR before any manipulation is VaR_0 and that a bank would really like the VaR to be VaR_1 where $0 > VaR_1 > VaR_0$ (see figure 3.4). One way to achieve this is to write a put option with a strike price below VaR_0 and buy one with a strike above VaR_1. The effect of this will be to lower expected profit and increase downside risk.

[24] Artzner P., Delbaen F., Eber J. and Heath D. (1999), Coherent measures of risk. Mathematical Finance, Vol. 9, pp. 203-228

[25] J. Danielsson (2011) Financial Risk Forecasting. Wiley Finance

Figure 3.4: Manipulation of P/L distribution



Source: J. Danielsson (2011)

VaR has also been criticized for its narrow focus. In its conventional form it is unable to account for any other risks than market risk and since it has also problems in estimating risk figures accurately for longer time horizons, as the results quickly deteriorate when moving for instance from monthly to annual measures.[26] Due to these factors, VaR is not a fool proof method. Tsai emphasizes that VaR estimates should therefore always be accompanied by other risk management techniques, such as stress testing, sensitivity analysis and scenario analysis in order to obtain a wider view of surrounding risks.[27]

3.4 Expected Shortfall

In order to overcome the problem of lack of subadditivity and provide more information regarding the tail shape Artzner et al.[28] propose, as an alternative measure of risk, the Expected Shortfall which represents the expected value of asset returns given that the threshold of VaR has been exceeded:

$$ES = -\mathrm{E}[y|y \leq -VaR(p)] \qquad (3.11)$$

Assuming that the distribution function of asset returns is continuous, the ES can distinguish between the level of riskiness in manipulated and non-manipulated assets as seen in figure 4.4. Since we are estimating the

[26] A. Damodaran (2007), Strategic Risk Taking: A Framework for Risk Management. Pearson Education, New Jersey.
[27] K.T. Tsai (2004), Risk Management Via Value at Risk, ICSA Bulletin, January 2004.
[28] Artzner P., Delbaen F., Eber J. and Heath D. (1999), Coherent measures of risk. Mathematical Finance, Vol. 9, pp. 203-228

expectation the Expected Shortfall, unlike the VaR, is aware of the tail distribution. Mathematically we have:

$$ES = -\int_{-\infty}^{-VaR(p)} x f_{VaR}(x)dx \qquad (3.12)$$

Despite the ES's higher degree of soundness both theoretically and intuitively, in practice the majority of financial institutions still employs VaR instead of ES. Essentially for two reasons:

- ES is estimated with a higher degree of uncertainty than VaR as the first step is ascertaining the VaR and then obtaining the expectation of tail observations. In this way, there are at least two sources of error in computing the ES.
- More importantly, ES is much harder to backtest than VaR since the former requires the estimate of tail expectation to compare with ES forecast. Thus, in backtesting, ES can only be compared with the output coming from a model while VaR can be compared with actual observations.

In conclusion, it is possible to state that although the evident deficiencies, the VaR is still considered the benchmark risk measure in financial industry.

3.5 Backtesting the Value-at-Risk

As showed in the previous paragraphs, many methods for the VaR estimation can be applied but the effectiveness in forecasting future risks must be evaluated. In order to assess their quality, the models should always be backtested with appropriate methods.

Backtesting is a statistical procedure where actual profits and losses are compared to corresponding VaR estimates. For instance, if the confidence level used for computing daily VaR is 95%, we expect an exception to occur five times every 100 days on average. In the backtesting procedure we can statistically examine whether the frequency of exceptions over the time interval is in line with the selected confidence level. These types of tests are

classified as *unconditional coverage* tests. However, in theory, a good VaR model does not only produce the 'right' amount of exceptions but also exceptions that are evenly spread over time i.e. they are independent of each other. When clustering of violations is present this indicates that the model does not accurately capture the changes in market volatility and correlations. For this reason, also tests of *conditional coverage* are required to examine conditioning in the data. We are now going to analyse different methods for backtesting.

3.5.1 Unconditional Coverage

A common and intuitive test for a VaR model is to count the number of exceptions (i.e. the number of days in which the portfolio losses exceed VaR estimates). If the number of exceptions is lower than the selected confidence level would indicate, the system overestimates risk. On the contrary, too many exceptions may denote underestimation of risk. In practice, it is rare to observe the exact amount of violations suggested by the confidence level. Thus, it is necessary to analyse statistically that amount is reasonable or not, determining the acceptance or rejection of the model.

Denoting the violation as η_t such that:

$$\eta_t = \begin{cases} 1 & if\ y_t \leq -VaR_t \\ 0 & if\ y_t > -VaR_t \end{cases} \qquad (3.13)$$

$$\upsilon_1 = \sum \eta_t \text{ and } \upsilon_0 = W_T - \upsilon_1$$

Where W_T is the testing window to be distinguished from W_E as the estimation window in which the model is set up.

We may define the Violation Ratio as the ratio between the observed number of violations and the expected number of violations:

$$VR = \frac{\upsilon_1}{p\ x\ W_T} \qquad (3.14)$$

Intuitively, if the Violation Ratio is greater than one the VaR model *underforecasts* the risk while if it is lower than one it *overforecasts.* Even though a useful rule of thumb is that the VR should stay in range between 0.8 and 1.2, we must assess the statistical significance of the estimates. Violations over time are a sequence of ones and zeros, often called *hit sequence* denoted

by $\{\eta_t\}$ is Bernoulli distributed. Thus, the null hypothesis for VaR violations is:

$$H_0 : \eta \sim B(p)$$

Where B is the Bernoulli distribution which density is given by:

$$(1-p)^{1-\eta_t} p^{\eta_t}, \qquad \eta_t = 0,1$$

Based on this principle is the Kupiec test[29], the most widely known test on the Violation Ratio (or failure rate), also known as the **POF-test** (Proportion Of Failures), it measures whether the number of violations is consistent with the confidence level under the null previously discussed of the model being 'correct' if the violations follow a binomial distribution. Thus, the informations required by the POF-test are the number of observations (W_T), the number of violations (υ_1), and the confidence level (p).

The null hypothesis for the POF-test is:

$$H_0 : p = \hat{p} = \frac{\upsilon_1}{W_T} \qquad (3.15)$$

The test aims to find out whether the observed violation rate $\hat{p}$ is significantly different from the rate p suggested by the confidence level.

According to Kupiec (1995), the POF-test is best conducted as a likelihood-ratio (LR) test. The test statistic takes the form:

$$LR_{POF} = -2ln\left(\frac{(1-p)^{\upsilon_0} p^{\upsilon_1}}{(1-\hat{p})^{\upsilon_0} \hat{p}^{\upsilon_1}}\right) \sim \chi^2(1) \qquad (3.16)$$

Under the null of a correct model, the LR_{POF} is asymptotically χ^2(chi-square) distributed with one degree of freedom. It is common to choose as confidence level 95% in which the null is rejected for $LR_{POF} > 3.84$. The Kupiec's POF-test presents two main issues: first, it is statistically sound only with large sample size (hence for instance it is weaker within the regulatory framework of one year specified in Basel Accords). Also, POF-test only considers the frequency of losses and not the time when they occur, failing to reject a model

[29] P. H. Kupiec (1995), Techniques for Verifying the Accuracy of Risk Measurement Models. The Journal of Derivatives Winter, 3 (2) 73-84;

that produces clustered violations. For these reasons, model backtesting should not rely exclusively on unconditional coverage tests.

Kupiec (1995) suggested also another type of backtest, called the **TUFF-test** (Time Until First Failure). This test measures the time (τ) it takes for the first violation to occur and it is based on similar assumptions as the POF-test. The test statistic is:

$$LR_{TUFF} = -2ln\left(\frac{p(1-p)^{\tau-1}}{\frac{1}{\tau}\left(1-\frac{1}{\tau}\right)^{\tau-1}}\right) \sim \chi^2(1) \qquad (3.17)$$

Again, the LR_{TUFF} is distributed as χ^2 with one degree of freedom. If the test statistic falls below the critical value the model is accepted. The main problem in this kind of test is its low power in identifying bad VaR models. For instance, if we compute daily VaR estimates at 99% confidence level and observe an exception already on day 7, the model is still not rejected (Dowd, 1998). Due to this evident lack of power, it is not advisable to use TUFF-test in backtesting when there are more powerful methods available. As Dowd [30] underlines, the TUFF-test is best used only as a preliminary to the POF-test especially when there is no larger set of data available.

3.5.2 Basel Traffic Light

Financial Institutions under Basel Accords are required to set aside a certain amount of capital to cover potential losses in their trading portfolios. The size of this capital related to market risk is defined by banks' VaR estimates. The regulatory framework requires banks to compute VaR for a 10-day horizon using a confidence level of 99%[31]. Under this framework, a strict backtesting mechanism is required to prevent banks from underestimating their risk exposure. The backtesting process is implemented by comparing the last 250 daily 99% VaR estimates with corresponding daily trading outcomes. The accuracy of the model is hence assessed by counting the number of violations

[30] K. Dowd (1998), Beyond Value at Risk: The New Science of Risk Management. Wiley, New York.
[31] Basel Committee of Banking Supervision (1996), Supervisory Framework for The Use of "Backtesting" in Conjunction with The Internal Models Approach to Market Risk Capital Requirements

during the period. [32] The size of the risk capital requirement rises as portfolio risk increases. Furthermore, the risk capital requirement depends on the outcome of the backtesting procedure.

$$S_t = \begin{cases} 3 & if\ x \leq 4 \\ 3 + 0.2(x - 4) & if\ 5 \leq x \leq 9 \\ 4 & if\ 10 \leq x \end{cases} \quad (3.18)$$

Here S_t is the scaling factor of market risk capital requirement and x the number of exceptions over 250 trading days. Basel Committee (1996) classifies backtesting outcomes into three categories: *green*, *yellow* and *red* zones. The categories are chosen in order to balance between errors of type 1 and 2. In the table below the cumulative probability is shown:

Figure 3.5: Traffic light approach

Zone	Number of exceptions	Increase in scaling factor	Cumulative probability
Green Zone	0	0.00	8.11 %
	1	0.00	28.58 %
	2	0.00	54.32 %
	3	0.00	75.81 %
	4	0.00	89.22 %
Yellow Zone	5	0.40	95.88 %
	6	0.50	98.63 %
	7	0.65	99.60 %
	8	0.75	99.89 %
	9	0.85	99.97 %
Red Zone	10 or more	1.00	99.99 %

Source: Basel Committee (1996)

Assuming the model is correctly specified, the expected number of violations is 2.5. With zero to four exceptions observed, it falls into the *green zone* where the probability of accepting an incorrect model is quite low.

The yellow zone consists of violations from five to nine. These results might be produced both by accurate and inaccurate models with relatively high probability, even if they are more likely to be inaccurate models. The backtests resulting in this zone generally cause an increase in the multiplication factor, according to the number of exceptions. However, the increase is not automatic since if the bank is able to demonstrate that the VaR

[32] Basel Committee of Banking Supervision (1996), Supervisory Framework for The Use of "Backtesting" in Conjunction with The Internal Models Approach to Market Risk Capital Requirements

model is fundamentally sound suffering for instance from bad luck (as in theory the yellow zone does not imply an inaccurate model) the supervisors can consider revisiting the requirements. Indeed, Basel Committee classifies the reasons for a backtesting failure into the following categories:

- *Basic integrity of the model:* the entire system is not able to capture the risk of the positions or there is a problem in estimating volatilities and correlations.
- *Model's accuracy could be improved:* the risk of some instruments is not measured with enough precision.
- *Bad luck or markets moved in fashion unanticipated by the model:* for instance, volatilities and correlations turned out to be significantly different than predicted.
- *Intra-day trading:* there is a change in position after the VaR estimates were computed.

The *red zone* generally indicates a clear problem with the VaR model. As can be noted from Figure 3.5, there is only a very small probability that an accurate model would results in 10 or more violations in a sample of 250 observations hence, this zone usually leads to an automatic rejection of the estimated VaR model. (Basel Committee, 1996). Haas (2001) reminds that the Basel traffic lights cannot effectively be used to assess the goodness of VaR model because it does not take into account important features such as the independence of violations.[33] Also, this framework has problems in distinguishing good models from bad ones. All these issues were recognized by the Committee itself.

Because of these evident drawbacks, the Basel framework is mostly used as a preliminary test for the accuracy of the model since in any type of credible model validation the traffic lights are simply inadequate and more advanced tests should be applied.

[33] M. Haas (2001), New Methods in Backtesting, Financial Engineering, Research Center Caesar, Bonn.

3.5.3 Conditional Coverage

Both Basel framework and Unconditional Coverage tests focus only on the number of violations but, in theory, we would expect these violations to be independently spread over the time horizon. Efficient VaR models can react to changing volatility and correlations in such a way that violations occur independently of each other while worse models tend to result in sequences of consecutive violations.

VaR users are extremely concerned about detecting clustering of violations since large losses occurring in rapid succession are more likely to lead to catastrophic financial events than individual exceptions taking place more rarely. [34] For this reason, tests of Conditional Coverage were introduced to deal with this issue by examining both the frequency and the time in which exceptions occur. Many tests have been developed but for the purpose of this study I going to focus on the most famous one: the **Christoffersen's Interval Forecast Test**.

Proposed by Christoffersen in 1998, the test applies the same log-likelihood testing framework of Kupiec but including at the same time a separate statistic for the independence of violations[35]. Thus, combined with the rate of coverage, the test examines whether the probability of a violation on any day depends on the outcome of the day before. The testing procedure described below is explained by many authors such as Jorion (2001), Campbell (2005), Dowd (2006) and in greater detail in Christoffersen (1998).

The test is carried out by first defining an indicator variable that gets a value of 1 if VaR is exceeded and value of 0 if VaR is not exceeded:

$$I_t \begin{cases} 1 & if\ violation\ occurs \\ 0 & if\ no\ violation\ occurs \end{cases}$$

Then define n_{ij} as the number of days when condition j occurred assuming that condition i occurred on the previous day. To illustrate, the outcome can be displayed in a 2 x 2 table:

[34] P. Christoffersen, D. Pelletier (2004), Backtesting Value at Risk: A Duration-Based Approach, Journal of Financial Econometrics, 2004, Volume 2, 84-108

[35] P. Christoffersen, (1998) Evaluating Interval Forecasts. International Economic Review, 39, 841-862.

	$I_{t-1} = 0$	$I_{t-1} = 1$	
$I_t = 0$	n_{00}	n_{10}	$n_{00} + n_{10}$
$I_t = 1$	n_{01}	n_{11}	$n_{01} + n_{11}$
	$n_{00} + n_{01}$	$n_{10} + n_{11}$	N

In addition, let π_i represent the probability of observing an exception conditional on state *i* on the previous day:

$$\pi_0 = \frac{n_{01}}{n_{00}+n_{01}} \ , \quad \pi_1 = \frac{n_{11}}{n_{10}+n_{11}} \quad and \quad \pi = \frac{n_{01}+n_{11}}{n_{00}+n_{01}+n_{10}+n_{11}}$$

If the model is sufficiently accurate, the violations today should not depend on whether an exception occurred on the previous day. In other words, under the null hypothesis the probabilities π_0 and π_1 should be equal. Thus, the test statistic for the independence of violations is a Likelihood-Ratio:

$$LR_{Ind} = -2ln\left(\frac{(1-\pi)^{n_{00}+n_{01}}\pi^{n_{01}+n_{11}}}{(1-\pi_0)^{n_{00}}\pi_0{}^{n_{01}}(1-\pi)^{n_{10}}\pi_1{}^{n_{11}}}\right) \qquad (3.19)$$

Then, combining the independence statistic with Kupiec's POF-test we can obtain a joint test the two properties of an effective VaR model, the correct failure rate and the independence of its violations:

$$LR_{CC} = LR_{POF} + LR_{Ind} \qquad (3.20)$$

Also LR_{CC} is Chi-Square distributed but in this case with two degrees of freedom since in this test two separate LR-statistics are present.

In the Christoffersen's framework it is possible to examine if the test is not passed because of the inaccurate coverage or the clustered violations or even both. This can be done by simply computing each statistic separately using a Chi-Square with just one degree of freedom. Campbell[36] underlines that in some cases the model may pass the joint test while failing one or both the separated tests. Therefore, it is usually recommended to run the separate tests even when the joint test yields a positive result.

[36] S. Campbell (2005), A Review of Backtesting and Backtesting Procedure, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington D.C.

3.5.4 Loss Function

Another possible measure applied in the verification of Value-at-Risk reliability is the so called "Loss Function".

Basel Committee on Banking Supervision (1996) indicates, together with the number of exceptions, also their magnitude as an element of regulatory concern. For this reason, J.A. Lopez[37] included this concern into a set of loss functions that can be generalized as:

$$L = \frac{1}{n}\sum_{i=1}^{n} C_{t+i} \qquad (3.21)$$

$$C_{t+1} = \begin{cases} f(R_{t+1} - VaR_t) & y_{t+1} < -VaR_t \\ g(R_{t+1} - VaR_t) & y_{t+1} \geq -VaR_t \end{cases} \qquad (3.22)$$

A common loss function is the Violation Ratio seen in the previous paragraphs, also known as Binomial Loss, but its limitation lies in the fact that it can only consider the number of exceptions without their magnitude.

A possible alternative applied also by Y. Liu[38] is the so called "Tick" Loss used in the quantile regressions and described as follows:

$$C_{t+1} = \begin{cases} (1-\alpha)|R_{t+1} - VaR_t| & y_{t+1} < -VaR_t \\ \alpha|R_{t+1} - VaR_t| & y_{t+1} \geq -VaR_t \end{cases} \qquad (3.23)$$

In this study I will apply this latter loss function in order to train the Neural Network and then compare the results with the other VaR models developed.

3.6 Considerations on VaR

Despite the highlighted flaws, Value-at-Risk remains the benchmark instrument used by risk managers in every financial institution such as banks, insurances mutual funds and pension funds.

The estimation can be based on parametric and nonparametric assumptions, both with their pros and cons. This is the reason why the concept of

[37] J. A. Lopez (1998), Methods for evaluating value-at-risk estimates. Federal Reserve Bank of NewYork research paper n. 9802.

[38] Y. Liu (2005), Value-at-Risk Model Combination Using Artificial Neural Networks

Combining Forecasts will become useful in the following pages. Once estimated the models, backtesting provides a fundamental feedback about their accuracy. In this chapter some of the most popular approaches to VaR model validation have been presented. A good model satisfies two equally important properties. First, it produces the 'correct' amount of violations with respect to the defined confidence level. At the same time, a validation process too focused only on the perspective of the unconditional coverage test could potentially lead to a situation where we accept a model unable to capture the changes in correlations and volatility, yielding violations close to each other. In the empirical sector we are going to see in detail the process that, starting from the time series of the selected ETFs, leads to the estimation of parametric and nonparametric VaR and then the assessment of their quality through the backtesting procedures analysed along this chapter.

Chapter 4

Artificial Neural Networks for Combining Forecasts

The purpose of this chapter is to introduce the topic of Artificial Neural Networks (ANN), an extremely powerful class of mathematical models. Since this term is very general and includes many concepts and approaches from mathematics, statistics and computer science, the aim will simply be to understand the general applications and functionalities before going into details for the specific application of this study.

The issue arise as there are many real-world problems that cannot be translated on an algorithm while the human brain can approximately compute. The question is then, *how do we learn to explore such problems?* The point is exactly about learning, since the human brain has not the same computational power of computers but it is more adaptive. A computer is static while human brains, as a biological neural network, can reorganize itself and hence learn compensating possible errors.

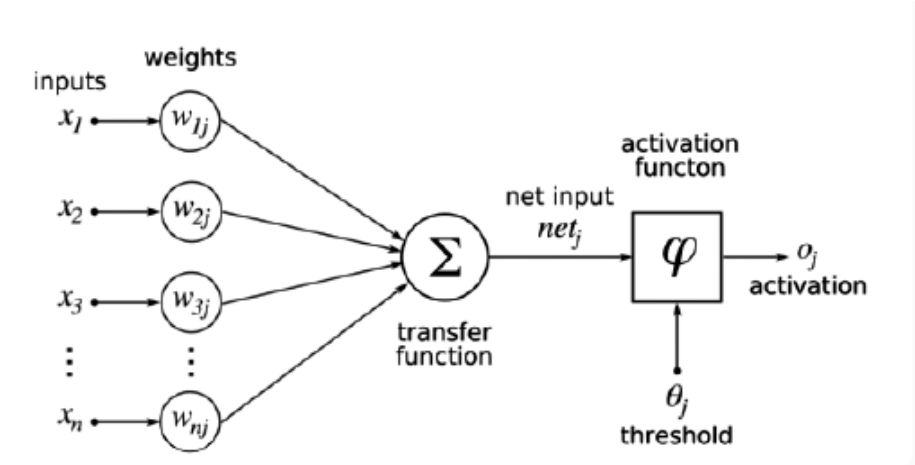
The study of artificial neural networks is indeed motivated by their similarity to biological systems having the capability to learn without an explicit programming phase: the result is an ability to generalize and associate data finding also reasonable solutions for similar problems of the same class that were not specifically trained.

The idea appeared for the first time when Warren McCulloch and Walter Pitts (1943) created a computational model for neural networks in the form of threshold logic unit that will be described in the following lines. Starting by a definition we can state that the term identifies an interconnected assembly of processing elements called *units* or *nodes,* whose functionality is based on the animal neurons. The processing ability of the network is located in the connections between these units (called *weights*) that are the result of an adaptation (or *learning*) process from a set of training patterns. The idea behind derives from the structure of human brains in which an estimated 100 billion nerve cells or *neurons* are present. Neurons communicate via electrical

signals that are spikes in the voltage of cell wall (called *membrane*). The interneuron connections are intermediated by conjunctions called *synapses*. Typically, each neuron receives thousands of connections from other neurons and, consequently, it is receiving a multitude of incoming signals which are integrated together resulting in an output response if the inputs exceed some threshold. In the determination of whether an impulse should be produced or not, some incoming signals results in an inhibitory effect preventing the neuron's "firing" while others are excitatory and promote the output generation. Thus, the processing ability of each neuron is supposed to reside in the kind and strength of the synaptic connections with other neurons.

The artificial equivalent of biological neurons follows the same rule as synapses are modelled through a value called weight so that each input is multiplied by this weight before being sent and summed with others in order to produce a *node activation* that can vary according to different rules. In *Figure 5.1* we have an example of the so-called *threshold logic unit* (TLU) in which if the activation exceeds a predetermined threshold, the unit produces a conventional output of 1 through a suitable transfer function, otherwise zero.

Figure 4.1: A simple artificial neuron



Source: Chrislb (2005)

Specific features of this network are:

- The parallel processing, due to the fact that neurons can simultaneously process the informations;

- The twofold function of neurons that can act both as signal processor and memory;
- The distributed nature of the data representation, i.e. knowledge is distributed throughout the network;
- The network's ability to learn from experience.

This last fundamental capacity allows neural networks to self-organize and adapt to incoming informations extracting connections between input and output from examples provided. The network is able to capture this attitude during an appropriate learning stage.

Along this chapter we are going to first describe the possible structures, features and applications of neural networks before focusing on the particular case analysed in the study. It is important to underline that neural networks (especially in the case of this thesis) can be considered a generalization of standard linear models, being able to capture non linearities and dependencies.

4.1 Structure of a Neural Network

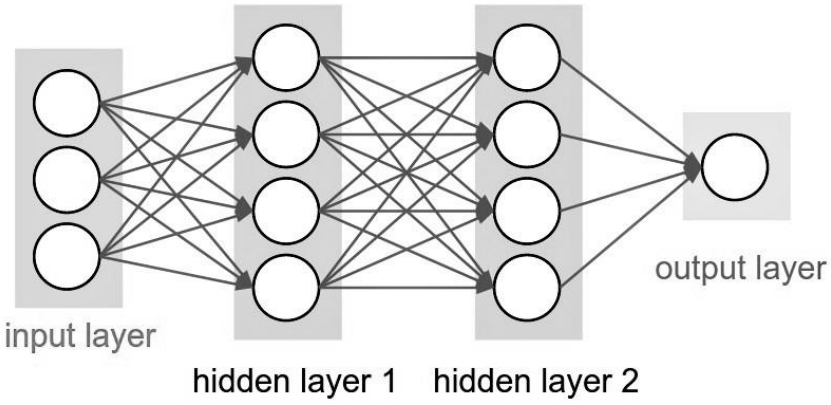
Neurons in the network can be combined according to several architectures: for instance, they may be arranged in layers (Multi-Layer Network) or they may have different connection topologies. Layered networks constituents are:

- Input layer, made of *n* neurons (in particular one for each input)
- Hidden layer, composed of one or more hidden (or intermediate) layers consisting of *m* neurons
- Output layer, consisting of *p* neurons (one for each output)

Considering the connections between layer's neurons we can distinguish two types of architectures:

1. The feedback architecture, with connections between neurons of the same or previous layer
2. The feedforward architecture that doesn't present feedback connections as the signals just go to the next layer's neurons

Figure 4.2: Multi-Layer Neural Network



Source: Stanford University (2019)

As we can see from *Figure 4.2*, each neuron receives *n* input signals x_i with connection weights w_i which sum together to an activation value of *y*. Then, a suitable transfer function transforms it to obtain the output F(y). The capability and efficiency of the network lie in the connection weights which are determined during the training phase.

The possible configurations are endless hence the choice of the optimal structure must be related to the type of searched target. Just as an example we can identify the main distinction in architectures as:

- **Perceptron:** it is the simplest model (Rosenblatt, 1958; Minsky & Papert, 1969) composed by a single neuron with *n* inputs and a single output. The basic learning algorithm of the perceptron analyses the configuration of the inputs (pattern) and, weighting variables through the synapses, decides which output is the best to be associated with the configuration. The main limitation of this architecture is of being able to solve only linearly separable problems.
- **Multi-Layer Perceptron (MLP) Network:** it is structured as a network with an input layer, one or more intermediate layers of neurons and an output layer. This network is a feedforward type and uses, in most cases, the backpropagation learning algorithm for the training phase. This method consists of starting from random values of weights and adjusting them with gradual and progressive changes after output's error until the learning algorithm converges to an acceptable error rate. There would be

many other types of complex structures but these lie beyond the purpose of this study as the feedforward supervised architecture is the most popular and widely used for the capacity of the model to generalize results for a large number of possible problems.

4.2 Activation Function

In this section we are going to summarize the main types of Activation Function (AF) developed over the years. The AF research and applications in deep architectures has always been a fundamental research field. These functions can be listed as follows:

A. Sigmoid Function

The Sigmoid Function can be considered in a group of three variants used in Deep Learning applications. It is a non-linear function mostly used in feedforward neural networks. It is a bounded differentiable real function, defined for real input values, with positive derivatives everywhere and some degree of smoothness. The Sigmoid Function is given by the relationship:

$$f(x) = \frac{1}{(1 + e^{-x})} \tag{4.1}$$

The function appears in the output layer of the Deep Learning architecture and it is used for predicting probability-based output. It has been successfully applied in different tasks such as binary classification problems, modeling logistic regressions as well as other neural network domains since its main advantage is being easy to understand and being applicable to shallow networks.[39]

[39] R. M. Neal (1992), Connectionist learning of belief networks. Artificial Intelligence, vol. 56, no. 1, pp. 71–113.

However, the main drawback are the sharp damp gradients during backpropagation from deeper hidden layers to input layer, gradient saturation, slow convergence and non-zero centred output hence causing the gradient updates to propagate in different directions.

B. Hyperbolic Tangent Function (Tanh)

The Hyperbolic Tangent Function is another type used in Deep Learning, it is zero-centred, smoother, whose range lies between -1 and 1. The output of the function is given by:

$$f(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})} \qquad (4.2)$$

Along the years, the tanh function became the preferred when compared to the sigmoid since the former gives better performances in multi-layer neural networks.[40] However, it may not solve the vanishing gradient problem typical of the sigmoid function as well.

The main advantage is given by the fact that the tanh produces zero centred outputs, aiding in this way the backpropagation process.

C. Softmax Function

The Softmax Function is another type used in neural computing, particularly in the estimation of probability distribution from a vector of real numbers. It produces an output in a range of values between 0 and 1, with them sum of probabilities being equal to 1. The Softmax Function is computed through the relationship:

$$f(x) = \frac{\exp\,(x_i)}{\sum_j \exp\,(x_j)} \qquad (4.3)$$

[40] B. Karlik and A. Vehbi (2011), Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 1, no. 4, pp. 111–122,

This function is mostly applied in multi-class models where it can return probabilities for each class, with the target class having the highest probability. The main difference between Softmax Function
and Sigmoid Function is that the latter is used in binary classification while the former for multivariate classification tasks.

D. Softsign

The Softsign was first introduced by Turian et al., 2009 and it is another non-linear activation function used in Deep Learning applications. It is a quadratic polynomial function given by:

$$f(x) = \frac{x}{(|x| + 1)} \qquad (4.4)$$

While the Softsign converges in polynomial form, the tanh function converges exponentially. This type of function has been used in regression computation problems but also applied to DL based test to speech systems with promising results. [41]

E. Rectified Linear Unit (ReLU) Function

The Rectified Linear Unit (ReLU) activation function was proposed by Nair and Hinton in 2010 and has been the most widely used activation function for Deep Learning applications. The ReLU is a faster learning AF offering better performances and a better generalization compared to Sigmoid and tanh functions.[42] It represents a nearly linear function and for this reason it preserves the properties of linear models that made them easy to optimize, with gradient descent methods.

The ReLU activation function performs a threshold operation to each input element where values lower than zero are hence the function is given by:

[41] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, and J. Miller, (2018) Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. International Conference on Learning Representations - ICLR, vol. 79, pp. 1094–1099.

[42] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, and G. E. Hinton (2013), On rectified linear units for speech processing. International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3517–3521

$$f(x) = \max(0, x) = \begin{cases} x_i & if\ x_i \geq 0 \\ 0 & if\ x_i \leq 0 \end{cases} \quad (4.5)$$

This function rectifies the values of the inputs lower than zero forcing them to zero and hence eliminating the vanishing gradient problem observed in the other types of activation function. The main advantage is that it guarantees faster computation since it does not compute exponentials and divisions, enhancing the overall speed of computation. Another remarkable property is that it introduces sparsity in the hidden units as it squishes the values the values between to maximum. Nevertheless, the ReLU presents the main limitation that it easily overfits if compared to sigmoid function although some dropout techniques are adopted to reduce this flaw. Another significant limitation is that sometimes during training it may cause some of the gradients to die, leading to some neurons to be dead as well and hence causing the weight updates not to activate in future data points. This can determine an obstacle to learning as dead neurons give zero activation.

These are just the most important activation functions; others are present and they can be summarized in the following table:

Figure 4.3: Activation Functions

Function	Formula
Linear	$f(x) = x$
Logistic (sigmoid)	$f(x) = 1/(1+e^{-x})$
Logistic symmetric	$f(x) = (1+e^{-x})/2$
Hyperbolic tangent	$f(x) = (e^{x}-e^{-x})/(e^{x}+e^{-x})$
Corrected tangent	$f(x) = \tanh(c \bullet x)$
Sinusoidal	$f(x) = \sin(x)$
Gaussian	$f(x) = e^{-x^2}$
Inverse Gaussian	$f(x) = 1-e^{-x^2}$

Source: C. Gallo (2015)

4.3 Applications of Neural Networks

The applications of Neural Networks can be group into three main areas:

1. Classification
2. Time Series Forecasting
3. Function Approximation

In the case of function approximation (or regression), networks are applied to all situations lacking a precise functional form describing input-output relations while in the area of time series prediction it aims to predict future values through past available periods of data. However, in each possible application, it is first necessary to divide time series into in-sample observations (training set) and out-of-sample observations (validation set). The network can also be trained with its own short-term forecast as input to provide a longer-term forecast.

For the effectiveness of this application it is important to assess some specific technical aspects such as the choice of input variables (since the relationships between them may change during time), an optimal learning level (a too short training process does not allow the network to capture the relationships while a too long one may cause overfitting) and the choice of the right time horizon for forecasting (as shorter forecast horizon results in a higher number of correct predictions while longer forecasting time horizon are on average less correct but the correct ones determine a higher average profit).

Neural Networks can also be used to classify data. Classification problems, unlike regression ones, require labelling each data point as belonging to of *n* classes available. Neural Networks can be trained to provide a discriminant function separating the classes. A classification problem can be learned without hidden units but sometimes a nonlinear function may be required to ensure the separation of classes and for this reason it can be solved only by a neural network presenting at least one hidden layer.

In this case it is necessary to first build an equivalent function approximation problem by the assignment of a target value to each class. For a binary problem of two classes we have a network of a single output *y* and a binary target value (0 for one class and 1 for the other) allowing to interpret the result as an estimated probability that a given sample belongs to one of two classes. In this kind of classification problems, the common activation function used is the Sigmoid Function, which saturates the two target values.

Typical classification problems are credit ratings, biological risk assessment or trust decisions and the network has the task of assigning the input to a corresponding output among several categories previously defined.

4.4 Training an Artificial Neural Network

The particular feature of Neural Networks is that they are not directly programmed but explicitly trained through the use of a learning algorithm for solving a given task. This process leads to "learning through experience" also known as *generalization*. The learning algorithm, helping in the definition of the network's configuration, in practice determines and conditions the ability of the network itself to provide correct answers to problems. Theoretically, a Neural Network can learn by:

1. Developing new connections
2. Deleting existing connections
3. Changing connecting weights
4. Changing the threshold values of neurons
5. Developing new neurons
6. Deleting existing neurons

Nevertheless, we can notice how the change in weight is the most common learning procedure.

In general, we can distinguish between three types of learning: *Supervised*, U*nsupervised* and by *Reinforcement*. In the first case, it is necessary to present to the network a set of examples to be used as inputs and corresponding outputs so that it can learn from them. The objective is to change weights not only to recognise the provided patterns but to achieve plausible results also with unknown but similar inputs (the network is hence supposed to be able to generalise). In the Unsupervised, instead, the network is trained only on the basis of a set of inputs without providing outputs. This is the biologically most plausible method but unfortunately it is not always suitable for all problems. Finally, learning by Reinforcement is used only in particular cases when it is not possible to specify input-output patterns; in this case reinforcement is

provided to the system which interprets it as a positive or negative signal about its behaviour and adjusts settings accordingly. Intuitively, this last procedure should be more effective than the Unsupervised learning since the networks is provided with specific criteria for problem solving.

As already stressed, a common procedure is to divide the inputs into a training set (70%) and a validation set (30%) and stopping the training when the network is able to provide good results on the training data as well as on the verification data.

4.4.1 Learning Curve and Learning Rate

The learning curve indicates the progress of the error which can be determined in various ways and it is fundamental to understand if the network is progressing or not. Furthermore, it is important to decide the rate at which the network changes weights compared to the error; this is called *learning rate* (often indicated as η). Thus, the selection of η is crucial for the behaviour of backpropagation and for learning procedure in general. If the chosen value is too large, jumps on the surface will be too large as well and, for instance, narrow valleys could be simply jumped over while a too small η can determine an extremely time-consuming training phase. Experience shows good learning rates are within the values of $0.01 \leq \eta \leq 0.9$.

The particular selection depends on the specific problem but it is a popular strategy to start from relatively large rate (e.g. 0.9) and slowly decrease to 0.01. This because in the beginning, a large training rate is able to lead to good results but later it often produces inaccurate learning while a smaller learning rate is more time-consuming but able to produce more accurate results.

First of all, we summarize the main indicators applied to quantitatively assess the error in the learning procedure. We point out:

1. Determination Index (R^2)

2. Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \quad (4.6)$$

3. Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - x_i}{y_i}\right| \quad (4.7)$$

4. Mean Square Error (MSE)

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n} = \frac{\sum_{i=1}^{n}(e_i)^2}{n} \quad (4.8)$$

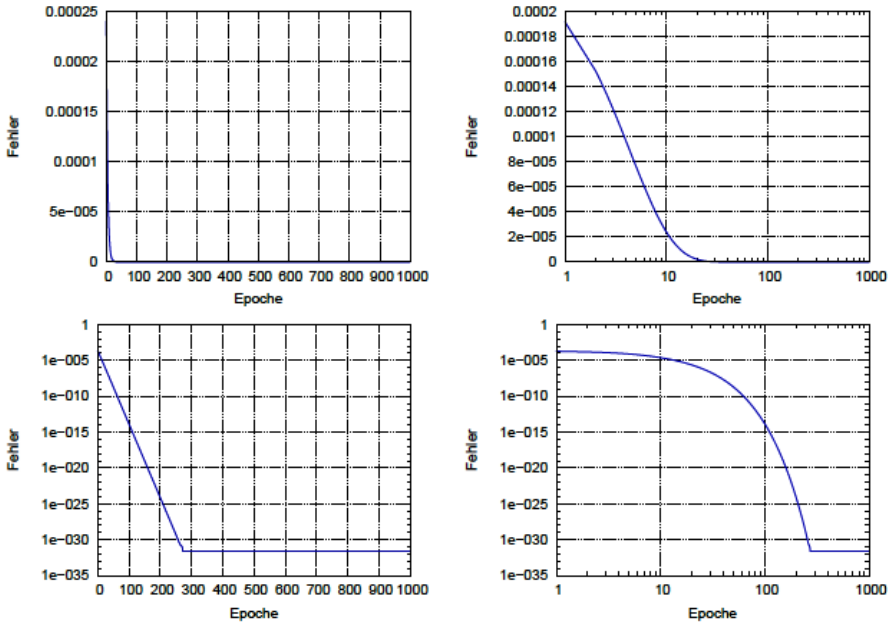
5. Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i)^2}{n}} \quad (4.9)$$

In which y_i is the actual output and x_i the predicted one.

All these indicators measure the spread between the original and the estimated output from the network so the learning procedure is in practice an optimization process aiming to minimize the above-mentioned error indicators. Also, it is important to set the number of epochs in which no improvement is made for stopping the process even though we can assume that a large number between 10,000 and 50,000 epochs is safe enough since a network will have difficulties in learning more than it has done in that point. The procedure should stop when, starting from different random initializing points, the network always reaches almost the same final error-rate. On the other hand, it might be possible that a curve descending fast in the beginning could be overtaken by another curve: this can indicate that either the learning rate of the worse curve was too high or the worse curve itself simply got stuck in a local minimum. It could be useful to analyse the learning curve when the network eventually begins to memorize the sample: indeed, if the learning curve of the training samples is suddenly rising while the one of the validation data is falling, this could indicate that the memorizing and generalization process is getting poorer. At this point a decision could be to stop the process as the network has already learned enough (the procedure is called *early stopping*). In the figures below we can see some learning curves along the various epochs even though idealized because too smoother than how they appear in real applications.

Figure 4.4: Learning Curve



Source: D. Kriesel (2007)

4.4.2 Gradient Descent

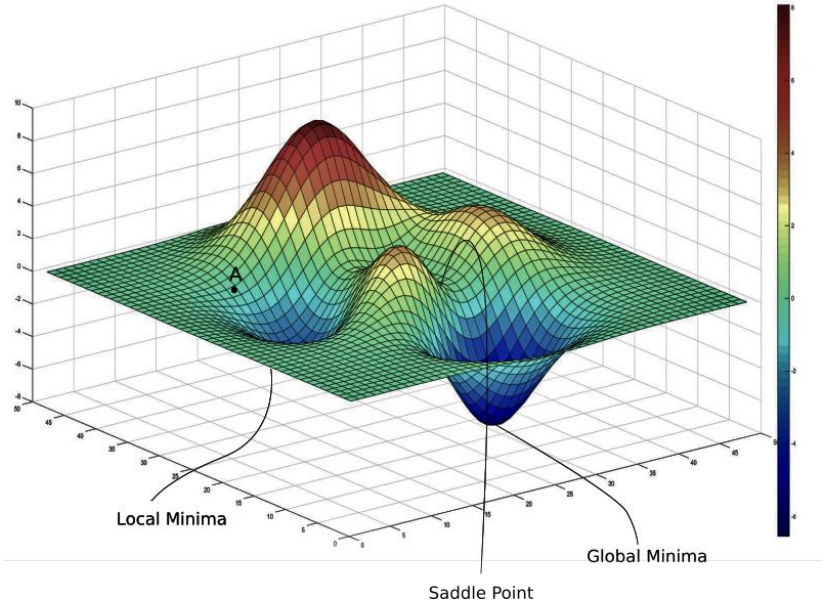
The entire learning process is based on the concept of *gradient descent*, it hence fundamental to first analyse what this term means.

Gradient descent procedures are usually applied when we want to maximize or minimize *n*-dimensional functions. The gradient is a vector g defined for any differentiable point, pointing exactly from that point towards the steepest ascent and indicates the gradient in this direction. Thus, the gradient can be considered a generalization of the derivative for multidimensional functions. Accordingly, the negative gradient exactly points towards the steepest descent.

In practice, the Gradient descent means going downhill through small steps from any starting point of the function towards the direction pointed by the gradient with the proportionality of step size given by $|g|$. Therefore, we move slowly on a flat plateau while we move rapidly downhill on a steep ascent. If we get into a valley, depending on the step's size, we could jump over or return into the valley across the opposite hillside in order to get closer and closer to the deepest point, similarly to a ball movement in a round bowl. These optimization procedures are not free from errors but they still work

well on many problems, which makes them an optimization paradigm frequently used.

Figure 4.5: Gradient Descent Optimization



Source: PaperspaceBlog (2018)

One of the major problem of gradient descent optimization is that it can get stuck in local minima and this issue usually increases proportionally to the size of the error surface and presenting no universal solution since, in reality, one cannot know if the global minimum has been reached and the training is successful. Also, when we are in the presence of flat plateaus, the gradient becomes negligibly small because there is almost no descent. In conclusion, it is possible to state that gradient descent requires several conditions to be applied and still it may not reach optimal results but due to the shortage of alternatives it remains the most applied procedure for training Neural Networks.

4.4.3 Backpropagation of Error

While the Gradient Descent is the mathematical base of the learning process, the Backpropagation of Error is the most applied method to train Multi-Layer Perceptron Neural Networks with semi-linear activation functions within the framework of Supervised Learning.

Backpropagation is a gradient descent procedure (presenting all strengths and weaknesses of that type of optimization) with an error function E receiving

all *n* weights as arguments and assigning them to the output error (i.e. being *n*-dimensional).

After choosing the weights of the network randomly, the backpropagation algorithm is used to compute the necessary corrections. The algorithm can be roughly decomposed in the following four steps:

i) Feed-forward computation

ii) Back propagation to the output layer

iii) Back propagation to the hidden layer

iv) Weight updates

The algorithm is stopped when the value of the error function has become sufficiently small.

The information fed into the network is $o_i w_{ij}$ where o_i is the stored output of unit i and w_{ij} the starting weight. The Backpropagation step computes the gradient of E with respect to this input and since o_i is treated as a constant we have:

$$\frac{\partial E}{\partial w_{ij}} = o_i \frac{\partial E}{\partial o_i w_{ij}} \qquad (4.10)$$

We can hence express the correction of weights Δw_{ij} by defining the so-called *delta rule* as:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \qquad (4.11)$$

Gradient-based Backpropagation is the standard method to train Artificial Neural Networks but has the condition of dealing with mean forecasts and with symmetric differentiable cost functions.

For the purpose of this study, since we are going to deal with VaR models and being hence interested in quantile forecast, we have to face an asymmetric non-differentiable cost function. Thus, it is necessary to resort to an alternative method to the standard one, which is the use of a Genetic Algorithm optimization.

In the next paragraph this alternative method of optimization of biases and weights is going to be presented.

4.4.3 Genetic Algorithm (GA)

Over many generations natural populations evolve according to the principles of natural selection and "survival of the fittest" clearly stated by Charles Darwin in *The Origin of Species*. By mimicking this process algorithms are able to evolve solutions to real world problems if they have been suitably encoded. As in nature, individual within a population compete with each other for resources, the most successful will have a have a relatively large number of offspring whilst poorly performing ones will produce a few or even no offspring at all.

This means that the genes from the highly adapted individuals will spread to an increasing number of individuals in each successive generation. The combination of good characteristics from different ancestors can sometimes produce "super fit" offspring whose fitness is greater than that of either parent. This is the same process through which species evolve and become more and more suited to environment in which they live.

Genetic Algorithm is an optimization solver based a reproduction of the natural selection, the process that drives biological evolution. GA's principles were first developed by Holland.[43]

The algorithm, starting from an initial population of solution candidates to the given problem, evaluates the quality of each of them according to a specific cost function. Then, it repeatedly modifies the population of individual solutions selecting at each step random individuals from the current population to be parents and combining them to produce the children of the following generation.

In this way, the population "evolves" towards the optimal solution.

It is hence possible to summarize the Genetic Algorithm's process:

[43] J.H. Holland (1975), Adaptation in Natural and Artificial Systems. University of Michigan Press

1. Create an initial population that can be random (the most common approach) or specified by the designers.
2. The algorithm creates a sequence of new populations, called *generations*. The individual present in each generation are used to create the next generation following these steps:

- Score each member of the current population by computing its fitness according to the cost function.
- Scale the raw fitness scores to convert them into a more usable range of values.
- Select parents based on their fitness
- Produce children from the parents.
- Replace the current population with the children of the following generation.

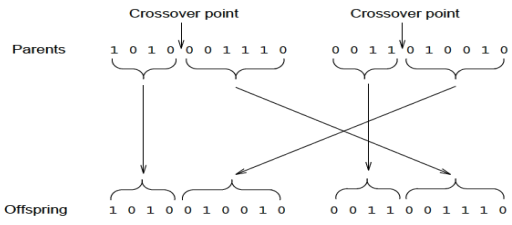
3. The procedure is repeated until one of the stopping criteria is met.

In this process it is fundamental to stress the features of some terms.

With the term *Reproduction* is specified the act of selecting the members of the population to move to the next population with probabilities proportional to their fitness. In this way the "better" chromosomes will have a higher probability of reproduce whereas the "bad" ones of being eliminated.

Crossover indicates when pairs of chromosomes in the new population are chosen at random to exchange genetic materials (their bits) in a mating operation indeed called crossover. A crossover operator is applied, dividing each parent into two parts, an hypothetical child 1 will be made by the first part of the first parent and the second part of the second parent when child 2 by the remaining parts. This produces two new chromosomes that replace the parents.

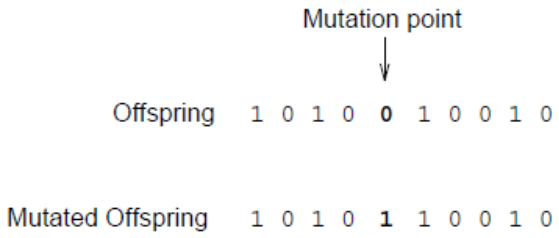
Figure 4.6: Single-point Crossover



Source: D. Beasley (1993)

The *Mutation* is when randomly chosen bits in the offspring are flipped. This operation gives to the GA the property of ergodicity which indicates that it will be likely to reach all parts of the state-owned space, without the travel in the resolution process.

Figure 4.7: A single mutation

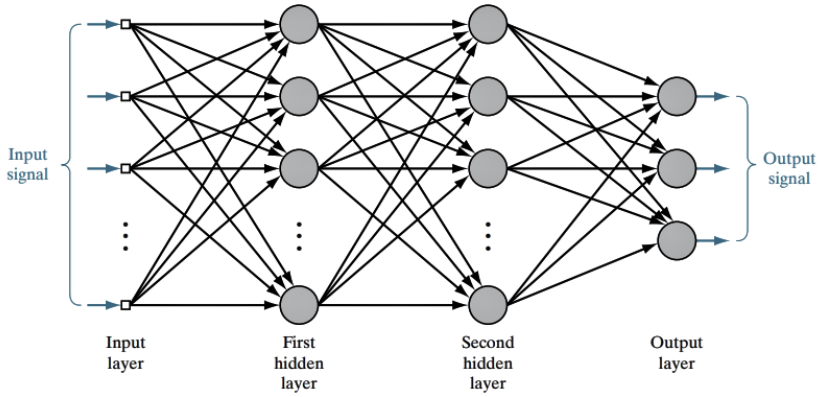


Source: D. Beasley (1993)

GA can be used to solve several optimization problems that cannot be solved by standard Gradient-Based optimization algorithms. Since this includes cases in which the objective function is discontinuous, non-differentiable, stochastic or highly nonlinear, the GA presents the perfect features to be used in the training of our Artificial Neural Network.

4.4 Multi-Layer Perceptron (MLP) Neural Network

Figure 5.8



Source: S. Haykin (2009)

In this paragraph we are going to deepen the features of the particular Neural Network that will be applied in the practical part of the study: the Multi-Layer Perceptron Neural Network.

As previously shown, this structure distinguish itself for the peculiarities of having one or more hidden layers, a nonlinear differentiable activation function and showing high degree of connectivity determined by synaptic weights of the network.

However, these same characteristics might determine some deficiencies in understanding the behaviour of the network: the presence of distributed form of nonlinearity and the high connectivity makes the theoretical analysis of MLP more difficult and when we are in the presence of several hidden neurons, the visualization of the learning process is made harder.

As already underlined, MLP Networks are often trained through Back-Propagation Algorithm dividing the process in two phases:

1. Forward Phase: the synaptic weights of the network are fixed and the input signal is propagated layer by layer until reaching the output.
2. Backward Phase: an error signal is produced by comparing the network's output with the desired response. The resulting signal is propagated through the network in a backward direction and successive adjustments of the synaptic weights are made.

The *Function Signal* propagates forward, emerging as an output signal while the *Error Signal* is originated at the output of the network and propagated backward. Each hidden or output neuron of a MLP performs two computations: a function signal appearing at the output of each neuron and expressed as a continuous nonlinear function of the input signal and synaptic weights related to that neuron. Also, the gradient vector is estimated since it is needed for the backward process.

In the case of classification tasks, hidden neurons play a critical role of feature detectors as they gradually discover the most important features characterizing the training data along the learning process. This is done by performing a nonlinear transformation on input data into a space called *feature space* where the classes of interest in a pattern-classification task may be more easily separated.

In conclusion, thanks to their flexibility MLP can be applied for several tasks and their nonlinearity allow them to theoretically approximate any function if

provided with enough neurons and data. As showed by K. Hornik et al. MLP can be considered universal approximators. [44]

4.4 Combining Forecasts

As we have seen in the previous chapter, there are many approaches in estimating VaR, ranging from parametric to non-parametric. In practice, one of the main issues is how to choose the "best" model among the candidates since the different methods might lead to different measures and significant errors. The risk of choosing the inappropriate model is called "model risk" and it is an important question left to risk managers. This issue results in an abundant literature concerning about model comparison.[45] Nevertheless, VaR models selection presents some difficulties: First, many studies indicate that individual models may be differently affected by non-stationarities such as volatility clustering and structural breaks.[46] For instance, in the presence of structural changes and a sudden increase of volatility, the Historical Simulation would not produce a tomorrow's VaR prediction much different from today's one since HS is based on empirical quantiles. On the other hand, GARCH model is able to capture volatility clustering property of financial time series but it would be only affected by today's increasing volatility temporarily and then the VaR prediction would revert to the previous level when the volatility decreases.

Thus, we could expect GARCH to have better performances in the short run while HS to have better chances to win in the long run being slower in changes but more stable and more precise in parameter estimation.

Due to the difficulty related to model selection, one possible solution is to find the best forecast by combination instead of selection. The theory of

[44] K. Hornik, M. Stinchcombe and H. White (1989), Multilayer Feedforward Networks are Universal Approximators, Neural Networks, 2, 359-366,

[45] Christoffersen etc. (1998, 2001, 2004), Sarma etc. (2003), Lopez (1998).

[46] M. Aiolfi - A. Timmermann (2004), Persistence in Forecasting Performance and Conditional Combination Strategies. Forthcoming in Journal of Econometrics
Hendry, D.F. and M.P. Clements (2002). Pooling of Forecasts. Econometrics Journal 5, 1-26

combining forecasts was originally developed by Bates and Granger.[47] From a theoretical point of view, forecast combination seen as a way to pool the informations contained in the individual forecast models into a third separate model. On average, combination can absorb the different adaptability of VaR models diversifying in this way the forecast error uncertainty. Since the main objective is not to find the correct model but the best one, this approach is accepted by practitioners. Tons of empirical literature supports the forecast combinations in different fields such as forecasting GDP, inflation, stock price, etc. More recent empirical work has further confirmed the accuracy gains by forecast combination.[48] Timmermann in 2004 provided a survey paper about forecast combination[49]. However, little empirical work has been done for the conditional quantile forecasting. Giacomini and Komunjer constructed a Conditional Quantile Forecast Encompassing test for the evaluation and combination of the Conditional Quantile Forecast.[50] The first to employ a MLP Neural Network to combine VaR forecasts was Y. Liu in 2005 obtaining promising results. [51]

In this study I am going in particular to follow this last method, using the two VaR models as inputs, training the Network with the use of a Genetic Algorithm in order to optimize the weights and obtaining as an output a more robust forecast.

[47] J.M. Bates, and C.W.J. Granger (1969), The combination of forecasts. Operations Research Quarterly, 20, 451-468

[48] J.H. Stock, M. W. Watson (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set

[49] A. Timmermann, (2004). Forecast Combinations. Forthcoming in Handbook of Economic Forecasting (Edited by Elliott, Granger and Timmermann (North Holland)

[50] R. Giacomini, I. Komunjer (2005). Evaluation and Combination of Conditional Quantile Forecasts Journal of Business & Economic Statistics Vol. 23, No. 4, pp. 416-431

[51] Y. Liu (2005), Value-at-Risk Model Combination Using Artificial Neural Networks

Chapter 5

Empirical Application

In this chapter we are going to practically apply the theoretical concepts analysed all along the previous chapters. We are hence studying first the statistical properties of the three selected ETFs, focusing on their log-returns. Then we are estimating their VaR both applying the nonparametric Historical Simulation and the parametric method assuming normal-distributed returns and modeling the volatility with a GARCH model backtesting the results in order to assess the reliability of the two methods. Finally, we are combining trough the Multi-Layer Perceptron Artificial Neural Network the two models according to the principle of Combining Forecasts to obtain a third model that could be able to capture the nonlinear dependencies showed by the HS and with a higher degree of adaptability to breaks typical of GARCH models. We are then assessing the results through the process of backtesting, comparing the HS, GARCH and Combined model performances.

5.1 Statistical Analysis of ETFs

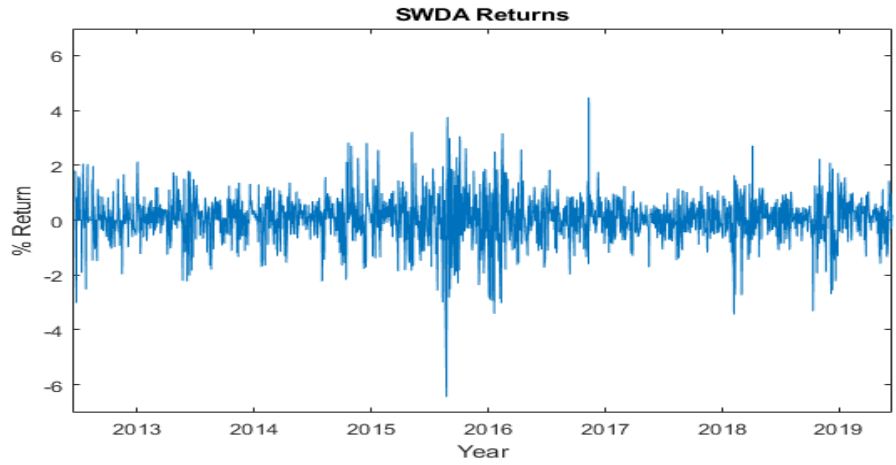
As shown in Chapter 2, three different ETFs are analysed retrieving 1750 observations (the equivalent of 7 years of trading days) between 2012 and 2019 from Bloomberg Terminal ©: iShares Core MSCI World UCITS ETF (SWDA), iShares Core MSCI Europe UCITS ETF (IMEU) and SPDR S&P 500 ETF (SPY). Returns are then computed as 100 times the difference in log-prices. The results and statistics are summarized in the following table and figures:

Figure 5.1: Summary Statistics

	SWDA	IMEU	SPY
Mean	0,0466	0,0231	0,0428
St.Dev	0,8342	0,8365	0,8059
Skewness	-0,4699	-0,7901	-0,4435
Kurtosis	7,3886	8,7621	6,6632

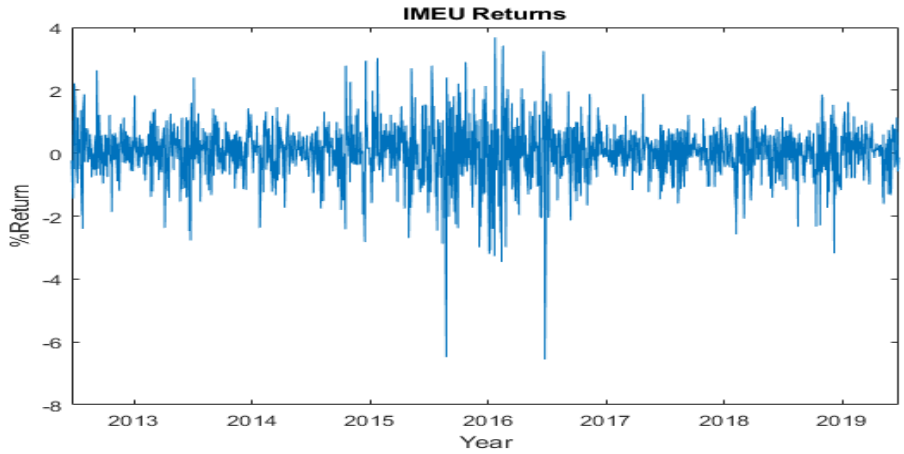
Source: Personal Elaboration

Figure 5.2



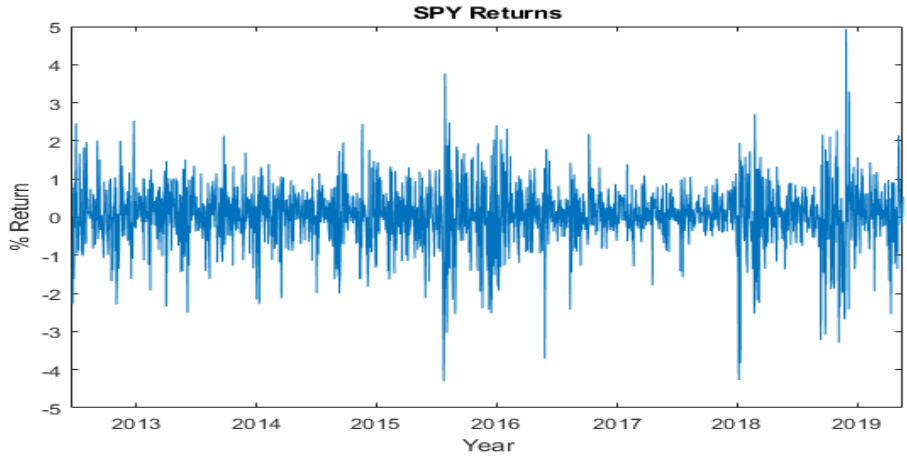
Source: Personal Elaboration

Figure 5.3



Source: Personal Elaboration

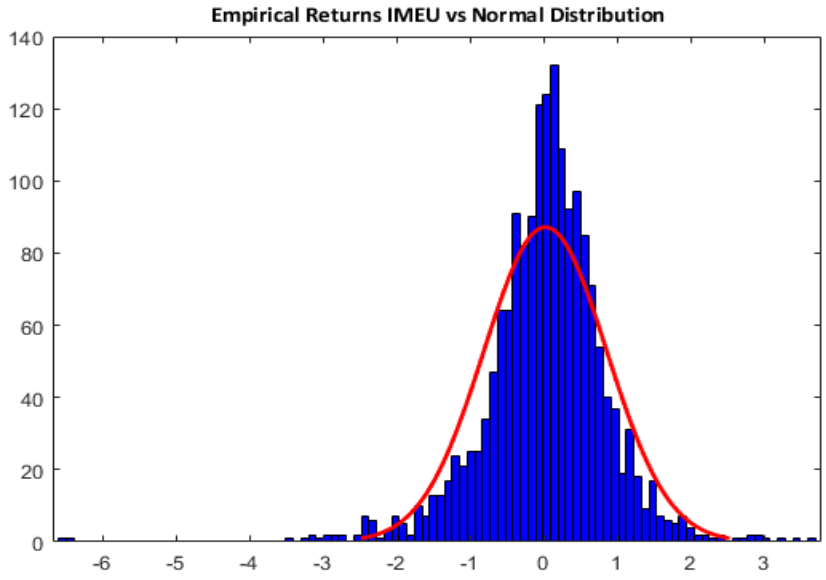
Figure 5.4



Source: Personal Elaboration

We can immediately notice how the three funds' returns present negative skewness (particularly pronounced for IMEU) and heavy tails since the level of kurtosis is much higher than the benchmark of 3 of the normal distribution. This last feature in particular is important because the presence of "fat tails" is one of the foundations of using GARCH models in this study. In fact, the Jarque-Bera Test and the Kolmogorov-Smirnov Test reject the null hypothesis of normality. As an example, we can see in Figure 6.5 a comparison between the empirical distribution of returns of IMEU and the normal distribution.

Figure 5.5



Source: Personal Elaboration

5.2 Value-at-Risk Estimation

At this point we can de-mean the returns and begin the VaR estimation according to the two different methods selected: the nonparametric in the form of Historical Simulation and the parametric assuming normality of returns and hence modeling the volatility according to a GARCH model. It is already possible to notice how the assumption of normality is particularly strong compared to the returns' empirical statistics seen before.

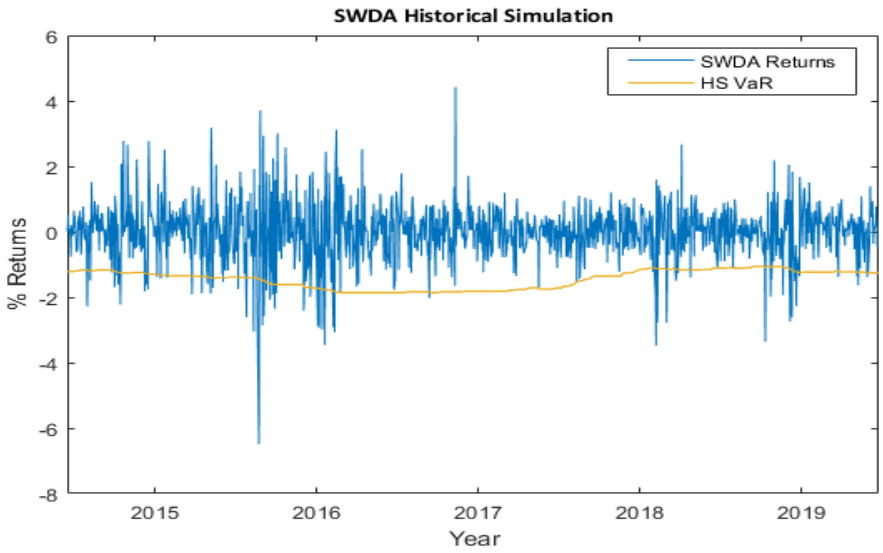
In both methods we divide the dataset into an estimation window of 500 observation (used for the estimation of the model) and a testing window

employed for assessing the adaptability of the model to out-of-sample observations. Then, we estimate the quality of the model through the process of backtesting, using the Violation Ratio, Mean Loss, POF Test, TUFF Test, Basel Traffic Lights, Christoffersen's Interval Forecast Test as quantitative metrics for compare the results.

5.2.1 Historical Simulation

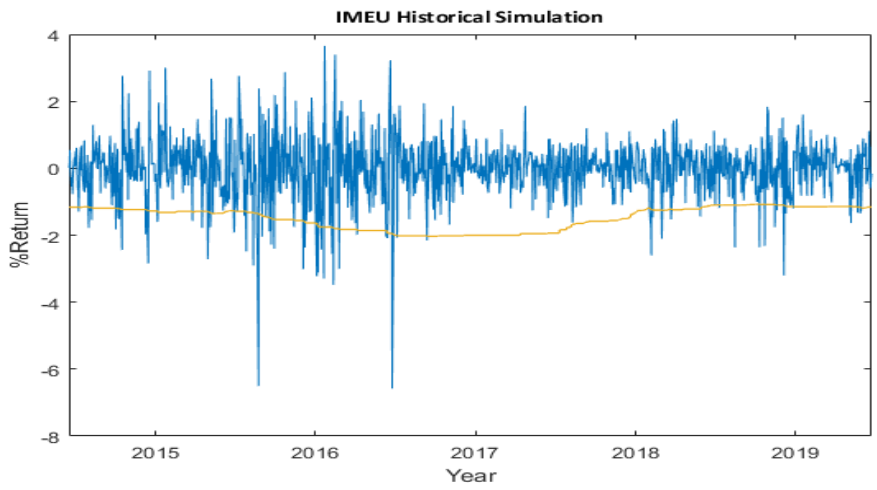
Setting the rolling estimation to a common level of 500 days we can compute the Value-at-Risk according to the Historical Simulation. The results for the three ETFs with a confidence level of 95% are as follows:

Figure 5.6



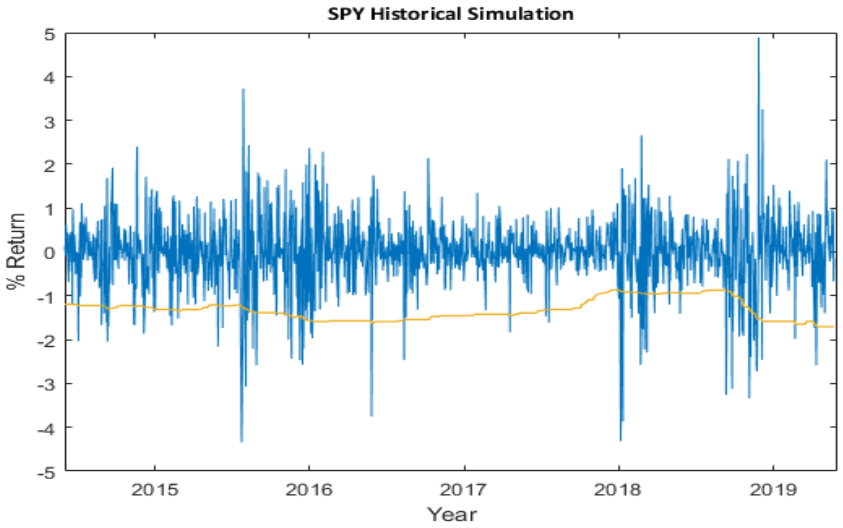
Source: Personal Elaboration

Figure 5.7



Source: Personal Elaboration

Figure 5.8



Source: Personal Elaboration

While in the following table the backtesting metrics results are summarized:

Figure 5.9

	SWDA	IMEU	SPY
VR	1,1617	1,1303	1,1639
Loss	0,1108	0,1155	0,1121
TL	green	green	green
POF	accepted	accepted	accepted
LRatio	1,6702	1,0944	1,7572
TUFF	accepted	accepted	accepted
LRatio	0,35381	0,026435	0,19779
CCI	rejected	rejected	rejected
LRatio	8,6295	14,851	18,855

Source: Personal Elaboration

From the reported values we can consider the HS a notable method in terms of stability and number of violations since in particular the VR and the POF Test shows acceptable results. The problems arise for the Conditional Coverage Test as the Likelihood Ratio is well higher than the target and indeed also from a visual analysis it is possible to notice how the model is stable but too little responsive to changes in volatility. This was foreseen given that the typical box-shaped behaviour of HS does not allow to capture volatility clustering and structural breaks since the forecast adapt much slowly to possible changes. We can hence conclude that HS provide good and stable results but its lack of responsiveness determines the failure of Conditional Coverage Test.

5.2.2 GARCH Model

For the parameter estimation I employed the Econometric Modeler in MATLAB's Econometric Toolbox Version 5.1 (R2018b).

Estimating several combinations, I selected four different models among which the GARCH (1,1) stands out as the best one.

Figure 5.10: SWDA GARCH Parameters

- ARCH (1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.46341	0.025028	18.5159	1.5371e-76
ARCH{1}	0.049736	0.02538	1.9597	0.050034

AIC	1062.0459
BIC	1070.4711

- ARCH (3)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.38646	0.029358	13.1637	1.4196e-39
ARCH{1}	0.047178	0.025461	1.8529	0.063895
ARCH{2}	0.074511	0.038121	1.9546	0.050634
ARCH{3}	0.08395	0.046322	1.8123	0.069941

AIC	1055.7438
BIC	1072.5782

- GARCH (1,1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.029019	0.013543	2.1427	0.032137
GARCH{1}	0.87597	0.041696	21.0083	5.5055e-98
ARCH{1}	0.060362	0.020022	3.0149	0.0025709

AIC	1042.5165
BIC	1055.1543

- GARCH (1,2)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.031219	0.016029	1.9476	0.051458
GARCH{1}	0.86715	0.052423	16.5416	1.8405e-61
ARCH{1}	0.044512	0.024802	1.7947	0.072698
ARCH{2}	0.019663	0.032625	0.6027	0.54671

AIC	1044.2627
BIC	1061.1051

Source: Personal Elaboration

Figure 5.11: IMEU GARCH Parameters

- ARCH (1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.46336	0.025309	18.3083	7.109e-75
ARCH{1}	0.020504	0.040086	0.51149	0.60901

AIC	1048.3301
BIC	1056.7553

- ARCH (3)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.34045	0.029304	11.618	3.3402e-31
ARCH{1}	0.0013976	0.0337	0.041472	0.96692
ARCH{2}	0.16949	0.051501	3.2909	0.00099857
ARCH{3}	0.10597	0.053438	1.983	0.047368

AIC	1026.1807
BIC	1043.0151

- GARCH (1,1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.052472	0.022478	2.3344	0.019577
GARCH{1}	0.79929	0.074002	10.8009	3.4094e-27
ARCH{1}	0.087756	0.032992	2.66	0.0078152

AIC	1026.5515
BIC	1039.1894

- GARCH (1,2)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.090433	0.025044	3.611	0.00030508
GARCH{1}	0.66525	0.078249	8.5017	1.8686e-17
ARCH{1}	0.0029883	0.030575	0.097737	0.92214
ARCH{2}	0.13851	0.044904	3.0845	0.0020386

AIC	1021.2639
BIC	1038.1063

Source: Personal Elaboration

Figure 5.12: SPY GARCH Parameters

- ARCH (1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.46929	0.032257	14.5485	5.9689e-48
ARCH{1}	0.13069	0.051769	2.5245	0.011587

AIC	1104.6343
BIC	1113.0595

- ARCH (3)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.30305	0.040097	7.5579	4.0964e-14
ARCH{1}	0.13148	0.047872	2.7465	0.0060238
ARCH{2}	0.030027	0.04112	0.73022	0.46525
ARCH{3}	0.22549	0.064439	3.4993	0.00046645
ARCH{4}	0.090652	0.046886	1.9335	0.05318

AIC	1093.7981
BIC	1114.8309

- GARCH (1,1)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.097587	0.043739	2.2311	0.025672
GARCH{1}	0.6902	0.11225	6.1485	7.8204e-10
ARCH{1}	0.12756	0.048057	2.6543	0.0079472

AIC	1095.2853
BIC	1107.9231

- GARCH (1,2)

Parameter	Value	StandardError	TStatistic	PValue
Constant	0.10302	0.057072	1.805	0.071072
GARCH{1}	0.67246	0.15719	4.2779	1.8867e-05
ARCH{1}	0.10861	0.051396	2.1132	0.034585
ARCH{2}	0.026509	0.071135	0.37266	0.7094

AIC	1097.1336
BIC	1113.976

Source: Personal Elaboration

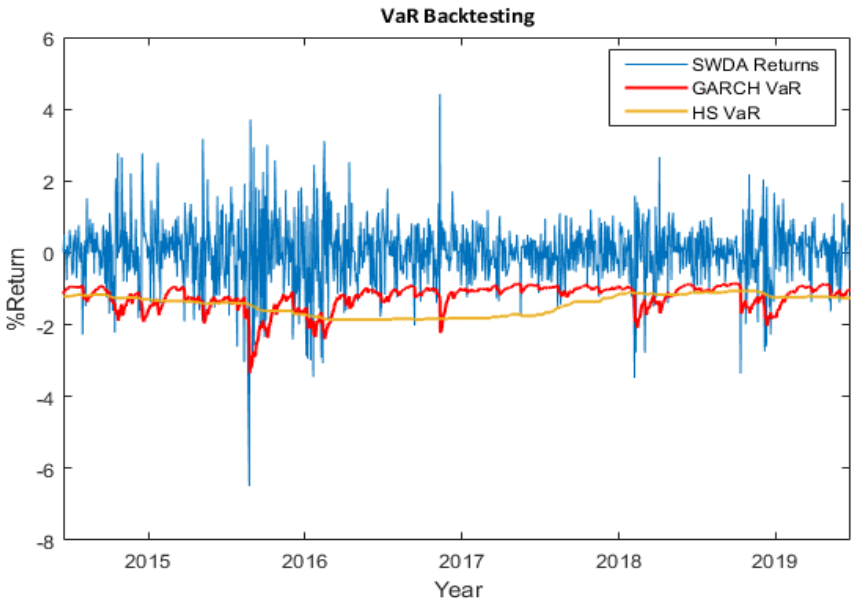
From the analysis of parameters' statistical significance and comparing the models through the Akaike Information Criteria and Bayesian Information Criteria, we can see how the GARCH (1,1) appears to be the more suitable model in all the three datasets.

Thus, we can write the three GARCH models, employing the Lag Operators, as:

$$\sigma_t^2 = \omega + \alpha L y_t^2 + \beta L \sigma_t^2 \qquad (5.1)$$

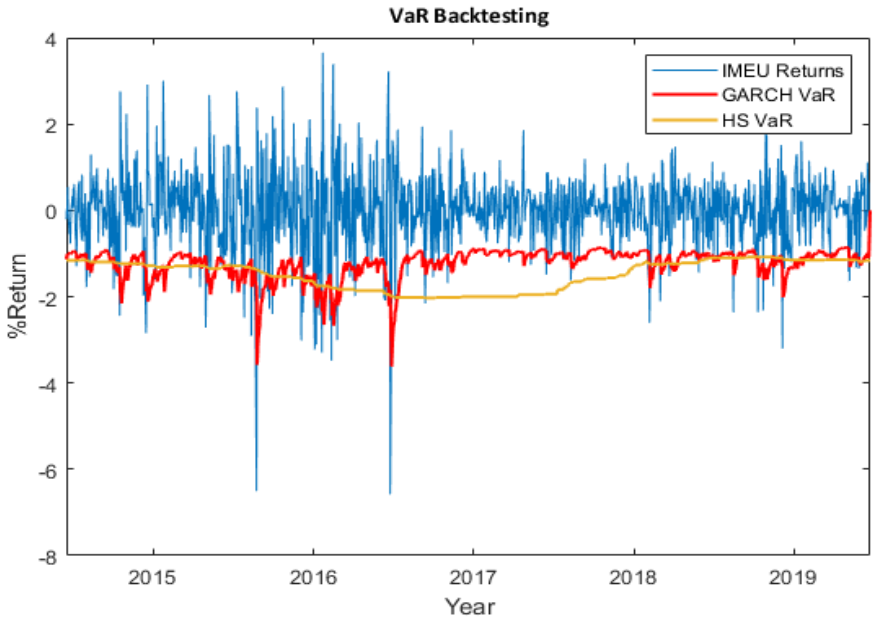
The graphical results (compared with the Historical Simulation) are:

Figure 5.13: SWDA Comparison



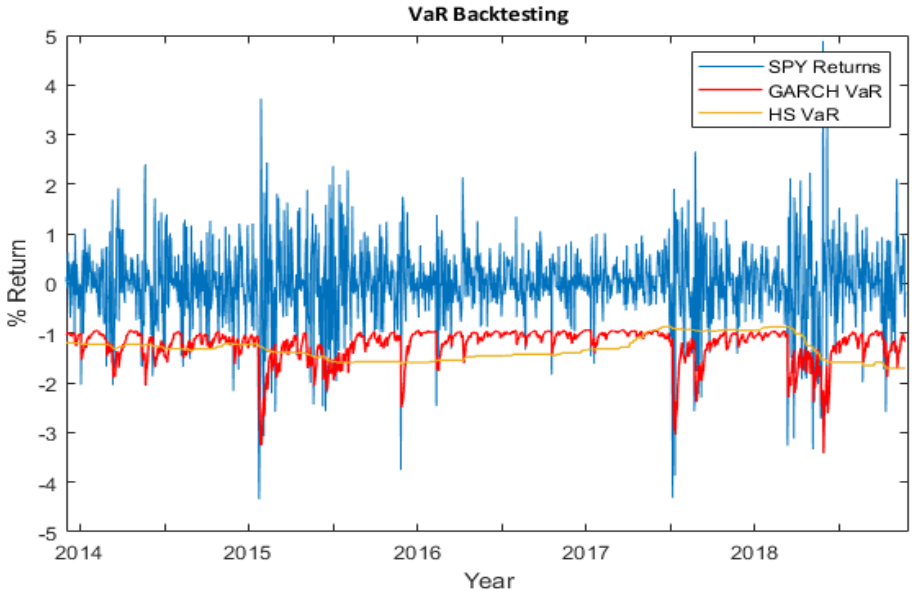
Source: Personal Elaboration

Figure 5.14: IMEU Comparison



Source: Personal Elaboration

Figure 5.15: SPY Comparison



Source: Personal Elaboration

Whilst in the table, the statistical test results are summarized:

Figure 5.16

	SWDA	IMEU	SPY
VR	1,4129	1,4275	1,1792
Loss	0,1082	0,1085	0,1003
TL	yellow	yellow	green
POF	rejected	rejected	accepted
LRatio	10,188	10,89	2,0919
TUFF	accepted	accepted	accepted
LRatio	0,35381	0,011307	0
CCI	rejected	rejected	rejected
LRatio	8,2979	6,2607	5,7343

Source: Personal Elaboration

Although in the chart it is possible to notice how the GARCH VaR is much more responsive to changes in volatility, we can also see how the number of violations is higher leading to the rejection of tests (for SWDA and IMEU). In the Conditional Coverage Test the results in the Likelihood Ratio are better if compared to HS VaR but still not sufficient to accept the test. In terms of Mean Tick Loss, the GARCH model provides better results since the number of exceptions is higher but their magnitude is much more limited.

In conclusion, the Historical Simulation is much more unresponsive but in the long run produces on average better results while GARCH is able to adapt rapidly to changes in volatility (hence providing better performances for the CCI test even if still not acceptable) but in the long run is less precise even though the magnitude of exceptions is lower. This might be particularly due to the very strong assumption of normal distribution of returns. Indeed, we have seen how the hypothesis of heteroskedasticity appears to be realistic determining the validity of GARCH in modeling volatility but the normality of returns results in an inability to effectively capture extreme variations (especially for the case of SWDA and IMEU where the kurtosis of returns is more pronounced).

It is now evident how the problem of model selection arises and the difficulty of choice between the two VaR since both presents advantages and flaws. Thus, this alternative approach to Risk Management based on the use of Neural Networks appears to have the capability to overcome the problem.

5.3 Artificial Neural Network for Model Combination

Once estimated the two inputs (the individual models), it is now the moment to employ them in the design of the Neural Network. I have chosen an Historical Simulation and GARCH model since they use only partially overlapping informations: HS is based on the empirical distribution of returns while GARCH applies conditional volatility forecast model. Thus, it is likely that the combination of the two will provide better results thanks to the more informations employed as demonstrated by Liu (2005)[52].

I hence divided the individual forecasts into two subsamples: the first 70% of data (a common value for the in-sample Training Set) is used to select the optimal specification of ANN models, in particular the value of weights and biases. Then, I used the remaining subsample to compare the performances between individual VaR forecasts and ANN combination. The Network was

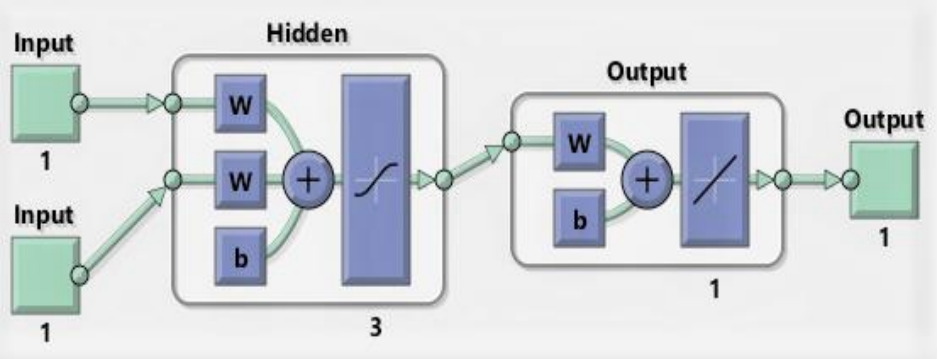
[52] Y.Liu (2005), Value-at-Risk Model Combination Using Artificial Neural Networks

trained minimizing the “Tick Loss” function and then the results are measured on the basis of the Backtesting Criteria previously described: Violation Ratio, Tick Loss, POF-test, TUFF-test and CCI-test.

While common ANNs are employed in forecasting minimizing symmetric loss functions, such as the Mean Squared Error, through the Backpropagation of Error based on Gradient Descent, in this study I had to deal with a conditional quantile forecast hence with an asymmetric non-differentiable 'tick' loss function for which the use of a Genetic Algorithm is more suitable. Thus, some parameters must be set for the initialization of the GA, they are either derived from previous studies and the result of empirical tests. The Population Size is established as 10 times the number of neurons. The number of Generations is set as 200 since in the training experience showed to achieve a good convergence. The Crossover Parameter is set as a common 0.5 and the Mutation Parameter as 0.08. These last two must not be too high since Crossover and Mutation are two infrequent events that tend to recombine the genes, destroying in this way the possible good results obtained and reducing the probabilities of achieving global minima. The choice of parameters is made considering the trad-off between training efficiency and computing time.

After several long training sessions I concluded that the best network structure for this task would have been a Shallow Neural Network with three layers in total (a network with a higher number of layers would be inserted in the category of Deep Neural Networks): an input layer with two inputs, a hidden layer made up of three neurons an output layer and a single output employing a tan-sigmoid activation function. The choice of the number of neurons follows the same rationale of the parameters in a common regression problem: too many parameters may determine extremely good performances with in-sample data but result in poor out-of-sample performances due to the so-called “Overfitting”. The network was hence designed in MATLAB as follows:

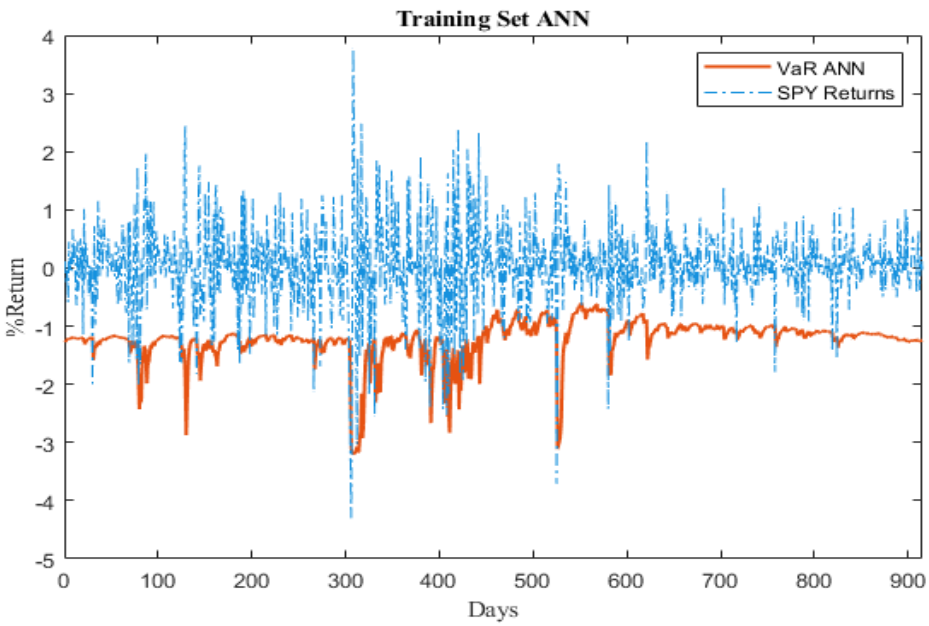
Figure 5.17: ANN Optimal Structure



Source: Personal Elaboration

The Training procedure had the main objective to find an optimal balance between in-sample and out-of-sample performances. This difficulty can be showed in the example below in *Figure 5.18*, where the Neural Network "learned almost by heart" the Training Set of SPY perfectly minimizing the loss function and resulting in a remarkable in-sample performance but then finding many difficulties with new data.

Figure 5.18: A case of Overfitting



Source: Personal Elaboration

Thus, the aim is to find an optimal structure (a combination of weights and biases) that is able to minimize the target loss function but at the same time

without resulting in overfitted parameters that would be useless if applied to new data sets.

The results of the training procedure are showed in *Figure 6.19* where the table summarizes the in-sample loss comparison between HS, GARCH and ANN Value-at-Risk.

As expected, the ANN is able to minimize the loss function by designing an optimal structure that adapts to the provided training set, obtaining better results if compared to the corresponding performances of the inputs (in this case the VaR forecast with HS and GARCH model).

Figure 5.19: In-Sample Loss

	SWDA	IMEU	SPY
HS	0,1144	0,1255	0,0964
GARCH	0,1139	0,0955	0,0896
ANN	0,1083	0,0939	0,0891

Source: Personal Elaboration

Thus, if compared to HS models the ANN is able to reduce the in-sample loss by 5,33% for SWDA, 25,17 % for IMEU and 7,57% for SPY.

For GARCH models the improvements are more contained as the loss was already rather low but it resulted in lower losses of 4,92% for SWDA, 1,67% for IMEU and 0,56% for SPY.

Now, the important issue become the assessment of the out-of-sample performances of the Neural Network.

The second subsample is then applied as inputs for the ANN, maintaining the same architecture, in order to obtain a second output-forecast to be compared to the corresponding input performances.

Thus, applying the same backtesting criteria we can see he comparison of results in terms of Violation Ratio, Mean Tick Loss, POF Test, TUFF Test and CCI Test summarized in the following table.

Figure 5.20: Out-of-Sample Performance

	SWDA			IMEU			SPY		
	ANN	HS	GARCH	**ANN**	HS	GARCH	**ANN**	HS	GARCH
VR	1,0937	1,3021	1,25	1,4062	1,1458	1,1458	1,3811	1,8414	1,5345
Loss	0,1019	0,097	0,1022	0,091	0,0884	0,0891	0,1254	0,1253	0,1489
POF	accepted	accepted	accepted	accepted	accepted	accepted	accepted	rejected	rejected
LRatio	0,17261	1,691	1,1743	2,978	0,4114	0,41136	2,685	11,799	5,0901
TUFF	accepted	accepted	accepted	accepted	accepted	accepted	accepted	accepted	accepted
LRatio	0,443	0,27176	0,27176	3,3215	0,39756	0,39756	0,19779	0,10455	0,19779
CCI	accepted	rejected	rejected	accepted	accepted	rejected	accepted	rejected	rejected
LRatio	0,58714	5,516	6,2054	2,1218	2,0229	4,5235	2,21	11,754	5,1539

Source: Personal Elaboration

The table shows how the ANN VaR was able to pass all the statistical tests, in particular the Conditional Coverage Test that, except for the case of HS used with IMEU, was not acceptable for any input VaR.

In terms of Violation Ratio, the ANN in the case of SWDA provides a remarkable value of 1,0937, also for SPY the Neural Network performs better while in the IMEU case there is no particular improvement even though the out-of-sample Mean Tick Loss remains contained highlighting how there still might be violations but their magnitude is limited. The out-of-sample Tick Loss is slightly higher for the ANN VaR if compared to the inputs but this was obviously expected and supported by theory since the Neural Network has been trained to minimize the loss function of another dataset.

We can hence conclude that the combination of the two models brings considerable quantitative improvement in almost every component of the backtesting procedure, giving legitimacy to this experiment of model combination as an alternative to model selection.

Once considered the most important quantitative aspect of the study, a graphical analysis can be useful to more deeply understand the rationale behind the Neural Network's output.

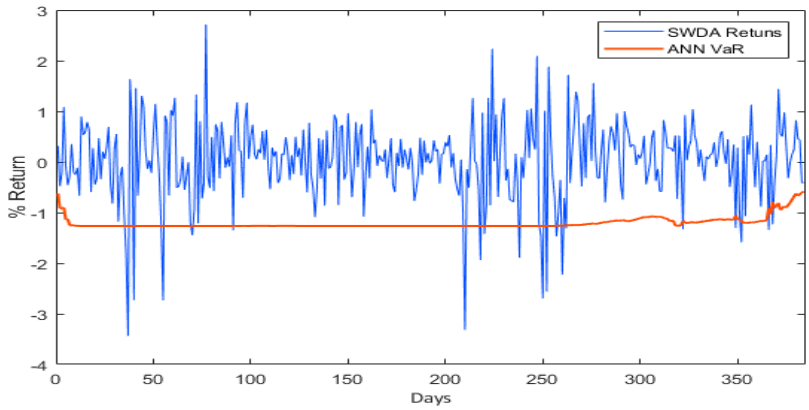
Indeed, one of the main problems of Machine Learning and Artificial Intelligence in general, that is still limiting their application in some fields, is the so called "black-box effect".

More in detail, as the Network become more and more "deep" (inserting hidden layers and neurons), the outputs might be more complex and consequently more difficult to be explained. The problem is no more obtaining the result but understanding the reasoning behind it.

Fortunately, for the purpose of this study, I employed a rather simple structure with only one hidden layer and a limited number of neurons, allowing to more easily understand the output results.

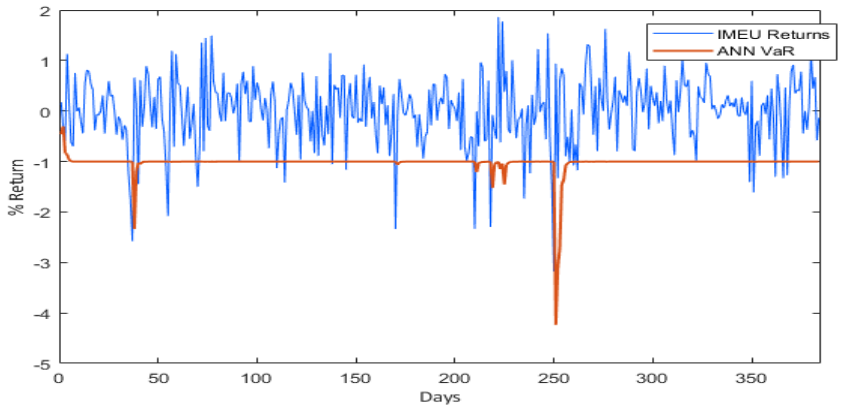
Therefore, analysing the charts, it is possible to see how for instance the weights related to HS model for SWDA and IMEU are higher if compared to the GARCH model. This is also intuitive since we have seen how the HS performed better in the first two ETFs. This is why the ANN VaR appears less responsive to changes in volatility both for SWDA and IMEU (even though in this case some spikes are present).

Figure 5.21: ANN VaR for SWDA



Source: Personal Elaboration

Figure 5.22: ANN VaR for IMEU

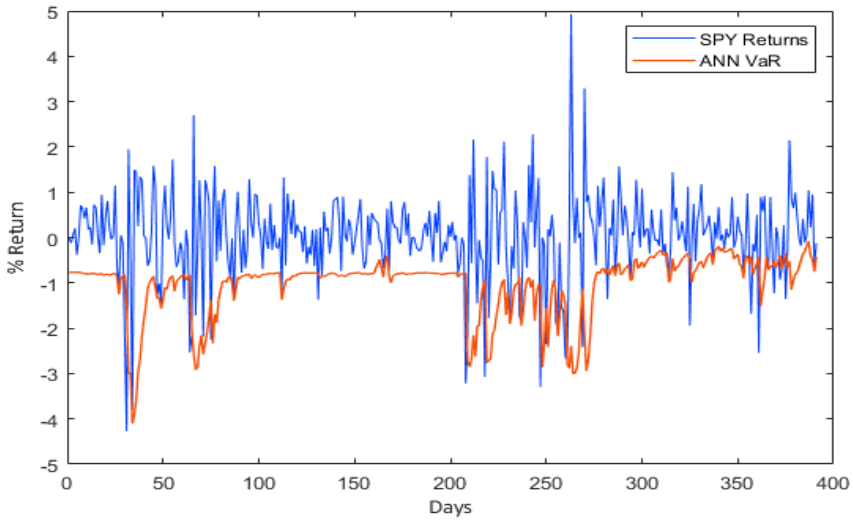


Source: Personal Elaboration

The results for the ETF SPY are indeed different because of the opposite nature of the inputs used.

In this case, the GARCH model performed quite well compared to the HS and both for the POF and CCI the Likelihood Ratio was close to the limit of acceptance. These inputs resulted in a "preference" of the Neural Network for the GARCH, determining higher weights and producing a combined Value at Risk much more dynamic and responsive as we can see from the figure below:

Figure 5.23: ANN VaR for SPY



Source: Personal Elaboration

In conclusion, the model combination through the use of Artificial Neural Networks showed encouraging results since the architectures were able to combine the two initial individual forecasts, minimizing the in-sample loss, into a third model that produced notable results for all the Backtesting Tests in all the three ETFs (even though in the case of IMEU there was not a particular improvement as the out-of-sample HS had already passed all the tests). The ANN VaR resulted in violations that were lower in amount, magnitude and which did not show any significant correlation.

We can hence conclude that despite the long fine-tuning process, the ANN combination model is able to improve the individual models' performances representing a valid alternative to common model selection.

Concluding Remarks

In this study we have proposed an alternative approach to Market Risk Management for Pension Funds investing in ETFs. The choice was particularly affected by two main reasons: this type of instrument is growing in importance for the asset allocation of Pension Fund Managers (thanks to its stability and low commissions) and at the same time allows to reproduce a well-diversified portfolio, easing the computation and the understanding of the Neural Networks' results.

After having illustrated the main characteristics of Pension Funds' investments and ETFs (including their development and valuation) we focused on the most important index of risk applied at the moment in the financial industry: the Value-at-Risk. We have first described the most common models and then dedicated the remaining part of the study to Historical Simulation and GARCH.

In Chapter 4 we have introduced the Artificial Neural Network's characteristics, based on structures made up by neurons (weights), biases and activation functions through which the input informations are passed, showing the common training procedure based on Gradient Descent but underlining how this was not applicable because of the presence of an asymmetric and non-differentiable loss function. We hence showed how a Genetic Algorithm, reproducing the evolutionary process, could be used in training the Network (in other words the optimization of weights and biases).

HS and GARCH models both present pros and cons in their application for Value-at-Risk estimation since they are based on various assumptions. HS model is more stable and in the long run is able to provide better results but it has problems in adapting to rapid changes in volatility. On the contrary, the GARCH model is much more responsive (producing also violations of lower magnitude) but the normality assumption is too strong at least for the returns of the ETFs analysed.

Since the former is an empirical quantile while the latter is a theoretical quantile based on heteroskedasticity, the two models presents non-

overlapping informations that can be correctly combined. Indeed, in order to avoid the so-called "model risk", the idea was to combine through an Artificial Neural Network the HS and GARCH model instead of selecting just one of the two. The results show how the combined output is able to capture the advantages of both models and result acceptable in the main backtesting criteria such as Violation Ratio, Conditional and Unconditional Coverage and "Tick" Loss Function.

The performances of the input VaR affected the output features since, obviously, the Neural Network gave higher weights to the better-performing model. Thus, the VaR for SWDA and IMEU showed dynamics more stable and less responsive (being more similar to Historical Simulation) while for SPY the better-performing GARCH resulted in a model more responsive to structural breaks.

We can conclude underlining that this approach is not limited to Pension Funds but can be obviously applied to other financial institutions facing the same Market Risk and considering stocks or indexes instead of ETFs. Also, the application can be implemented through the use of more complex models such as for instance Student-T distributions or Extreme Value Theory and with larger datasets in which Neural Networks can be trained more efficiently.

Appendix A.

In this Appendix we are presenting the codes applied in MATLAB for the estimation of Value-at-Risk models of Historical Simulation, GARCH and their consequent Backtesting.

GARCH parameters have been estimated through the use of the Econometric Modeler in MATLAB's Econometric Toolbox Version 5.1 (R2018b).

```
%Historical Simulation%
S=SPY;
t=dates;
R=diff(log(S))*100;
R=R-mean(R);
M = 500; % Estimation Window
t1 = length(R);
for t = 501:1774
      VaR_HS(t) = Historical(R((t-M+1):t),0.05);
end
VaR_HS=VaR_HS(501:end);
plot(dates(502:end),[R_test,-VaR_HS'])

%GARCH%
%Estimation of omega, alpha, beta through
Econometric Modeler
R_test=R(501:end);
Sigma2= zeros(length(R_test),1);
Sigma2(1)= var(R_est);
for i = 2: 1274
     Sigma2(i) =omega+beta*R_test(i-
1)^2+alpha*Sigma2(i-1);
end
vol_SPY=sqrt(Sigma2);
VaR=-norminv(0.05)*vol_SWDA;
plot(dates(502:end),[R_test,-VaR,-VaR_HS'])

%BACKTESTING%
vbt = varbacktest(R_test,VaR);
vbt = varbacktest(R_test,VaR_HS');
runtests(vbt);
summary(vbt);
pof(vbt)
tuff(vbt)
cci(vbt)
```

Appendix B.

In this Appendix we are presenting the codes applied in MATLAB in designing the Neural Networks and their Training procedure through the use of a Genetic Algorithm. The procedure is done only for SPY dataset.

```
%Training set%
input1=-VaR_HS(1:890);
input2=-VaR(1:890)';
targets=R_test(1:890)';
inputs=[input1;input2];
n=3; %neurons
net = feedforwardnet(n); %set the type of network
hidlaytransfcn = net.layers{1}.transferFcn;
hidlaytransfcn ='tansig';
net =configure(net, inputs, targets);
h = @(x) myloss2(x, net, inputs, targets);
%function to handle
options =
optimoptions('ga','CrossoverFraction',0.5,'Mutation
Fcn', {@mutationuniform, 0.08},
'Generations',200,'PopulationSize',
10*n,'display','iter','PlotFcn','gaplotrange');
[x, err_ga] = ga(h, 2*3*n+1, options);
% with n neurons, 3n+1 variables are required in
the weights and biases column vector.
% n for the input weights
% n for the input biases
% n for the output weights
% 1 for the output bias
net = setwb(net, x');
out=net(inputs);
t=1:length(out);
plot(t',[out',R_test(1:890)])
myloss(R_test(1:890),out,0.05)
myloss(R_test(1:890),-VaR(1:890),0.05)
myloss(R_test(1:890),-VaR_HS(1:890),0.05)
%Validation set%
input1=-VaR_HS(891:end);
input2=-VaR(891:end)';
inputs=[input1;input2];
targets=R_test(891:end)';
net = configure(net, inputs, targets);
net = setwb(net, x');
out=net(inputs);
```

References:

1. A. Damodaran (2007), Strategic Risk Taking: A Framework for Risk Management. Pearson Education, New Jersey.
2. A. Olivieri, E. Pitacco (2010), Introduction to Insurance Mathematics. Springer
3. A. Timmermann, (2004). Forecast Combinations. Forthcoming in Handbook of Economic Forecasting (Edited by Elliott, Granger and Timmermann (North Holland)
4. A.Seddik (2006), Exchange Traded Fund as an Investment Option. Palgrave MacMillan.
5. Artzner P., Delbaen F., Eber J. and Heath D. (1999), Coherent measures of risk. Mathematical Finance, Vol. 9, pp. 203-228
6. Autorité de Marché Fianciers (2017), ETFs: Characteristics, Overview and Risk Analysis-The case of French Market
7. B. Karlik and A. Vehbi (2011), Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. International Journal of Artificial Intelligence and Expert Systems (IJAE), vol. 1, no. 4, pp. 111–122
8. B. Mandelbrot (1963), The Variation of Certain Speculative Prices. The Journal of Business, Vol. 36, No. 4, pp. 394-419
9. Basel Committee of Banking Supervision (1996), Supervisory Framework for The Use of “Backtesting” in Conjunction with The Internal Models Approach to Market Risk Capital Requirements
10. C.S. Hsieh, J. H. Chou (2009), Forecasting Value at Risk (VAR) in the Shanghai StockMarket Using the Hybrid Method
11. Chia-Lin Chang, Michael McAleer, Chien-Hsun Wang (2017), An Econometric Analysis of ETF and ETF Futures in Financial and Energy Markets Using Generated Regressors
12. D. B. Loeper (2009), The four pillar of retirement plans. John Wiley & Sons, Inc.

13. D. Beasley, D. R Bull, R.R. Martin (1993), An Overview of Genetic Algorithms. University Computing 15(2), 58-68 Inter
University Committee on Computing
14. D. Franzen (2010), Managing Investment Risk in Defined Benefit Pension Plan. OECD Working Paper No. 38
15. D. J. Abner (2010), The ETF. John Wiley & Sons, Inc.
16. D. Kriesel (2007), A Brief Introduction to Neural Networks. (ZETA2-EN)
17. D.B. Loeper (2009), The four pillar of retirement plans. John Wiley & Sons, Inc.
18. Delcoure, N., & Zhong, M. (2007) On the premiums of iShares. Journal of Empirical Finance, 14, 168-195.
19. DWS (2018), Passive Investing: reshaping the global investment landscape
20. E. F. Fama (1965), The Behaviour of Stock-Market Prices. The Journal of Business, Vol. 38, No. 1. pp. 34-105.
21. E. Fama (1965), The Behavior of Stock Market Prices, Journal of Business 38, 34-105
22. E. Hehn (2005), Exchange Traded Fund, Springer
23. E. Hentov, A. Petrov, S. Odedra (2018), How do Public Pension Funds invest? From Local to Global Assets. State Street Global Advisor
24. El Hachloufi Mostafa, El Haddad Mohammed and El Attar Abderrahim (2016), Minimization of Value at Risk of Financial Assets Portfolio using Genetic Algorithms and Neural Networks. Journal of Applied Finance & Banking, vol. 6, no. 2, 2016, 39-52
25. ESMA (2014), ESMA/2014/937EN, Guidelines for competent authorities and UCITS management companies
26. F. A. Sortino (2001), Managing downside risk in financial markets, Quantitative Finance Series
27. F. Stewart (2010), Pension Funds' Risk Management Framework: Regulation and Supervisory Oversight. OECD Working Paper No. 40

28. Hendry, D.F. and M.P. Clements (2002). Pooling of Forecasts. Econometrics Journal 5, 1-26
29. Hung-Chun Liu, Yu-Ju, Cheng Yi-Pin Tzou (2017), Value-at-Risk-based risk management on exchange traded funds: the Taiwanese experience
30. J. A. Lopez (1998), Methods for evaluating value-at-risk estimates. Federal Reserve Bank of NewYork research paper n. 9802.
31. J. Danielsson (2011) Financial Risk Forecasting. Wiley Finance
32. J.H. Holland (1975), Adaptation in Natural and Artificial Systems. University of Michigan Press
33. J.H. Stock, M. W. Watson (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set
34. J. Robbins (2014), Essentials retirement planning. Business expert
35. J.M. Bates, and C.W.J. Granger (1969), The combination of forecasts. Operations Research Quarterly, 20, 451-468
36. James Chen (2019), Market Risk. Investopedia
37. John C. Hull (2009), Options, Futures, and Other Derivatives
38. K. Dowd (2006), Retrospective Assessment of Value-at-Risk. Risk Management: A Modern Perspective, pp. 183-202, San Diego, Elsevier.
39. K. Dowd, (1998) Beyond Value at Risk: The New Science of Risk Management. Wiley, New York.
40. K. Hornik, M. Stinchcombe and H. White (1989), Multilayer Feedforward Networks are Universal Approximators, Neural Networks, 2, 359-366
41. K. K. Lai, J. Yen (2009), A Statistical Neural Network Approach for Value-at-Risk Analysis
42. K. Lannoo, M. Barslund, A. Chmelar, M. von Werder (2014), Pension Schemes Study. European Parliament-Directorate General For Internal Policies
43. K.Gurney (1997), An introduction to neural networks. UCL Press
44. K.T. Tsai (2004), Risk Management Via Value at Risk, ICSA Bulletin, January 2004.

45. Kuester, K., Mittnik, S., and Paolella, (2006), Value-at-Risk prediction: a comparison for alternative strategies. Journal of Financial Econometrics, Vol. 4(1), pp. 53-89
46. M. Aiolfi - A. Timmermann (2004), Persistence in Forecasting Performance and Conditional Combination Strategies. Forthcoming in Journal of Econometrics
47. M. Caporin (2003), Evaluating value-at-risk measures in presence of long memory conditional volatility. GRETA, working paper n. 05.03
48. M. Choudhry (2006), An introduction to Value at Risk, John Wiley & Sons Inc.
49. M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, and G. E. Hinton (2013), On rectified linear units for speech processing. International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3517–3521
50. M. Dorocáková, (2017). Comparison of ETF´s performance related to the tracking error. Journal of International Studies, 10(4), 154-165
51. M. Haas (2001), New Methods in Backtesting, Financial Engineering, Research Center Caesar, Bonn.
52. P. Christoffersen, (1998) Evaluating Interval Forecasts. International Economic Review, 39, 841-862.
53. P. Christoffersen, D. Pelletier (2004), Backtesting Value at Risk: A Duration-Based Approach, Journal of Financial Econometrics, 2004, Volume 2, 84-108
54. P. Christoffersen, J. Hahn and A. Inoue (2001). Testing and Comparing Value-at-Risk Measures, Journal of Empirical Finance, 2001, Volume 8, 325-342.
55. P. H. Kupiec (1995), Techniques for Verifying the Accuracy of Risk Measurement Models. The Journal of Derivatives Winter, 3 (2) 73-84;
56. P. Jorion (2006), Value at Risk: The New Benchmark for Managing Financial Risk (3rd ed.). McGraw-Hill
57. P.Wang, M. Zhang, R.Shand, K.E. Howel (2014),Retirement, Pension Systems and Models of Pension Systems

58. Petronio, F., Lando, T., Biglova, A., & Ortobelli, S. (2014). Optimal portfolio performance with exchange traded funds. - Central European Review of Economic Issues, 17(1), 5-12
59. R. A DeFusco, S. I Ivanov, & G. V. Karels, (2011), The exchange traded funds' pricing deviation: analysis and forecasts. Journal of Economics and Finance, 35(2), 181-197.
60. R. M. Neal (1992), Connectionist learning of belief networks. Artificial Intelligence, vol. 56, no. 1, pp. 71–113.
61. R. Prudencio and T. Ludermir (2004), Using Machine Learning Techniques to Combine Forecasting Methods
62. R. Urwin, T. Hodgson, B. Collie, L. Yin, M. Hall (2019), Global Pension Assets Study. Thinking Ahead Institute - Willis Tower Watson
63. R.F Engle (1982), Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, Econometrica 50: 987-1007
64. R.Ferri (2012), All you need to know about ETFs. Don Phillips
65. R.G. Donaldson and M. Kamstra, (1996). Using Artificial Neural Networks to Combine Forecasts, Journal of Forecasting, 15, 49-61
66. R. Giacomini, I. Komunjer (2005). Evaluation and Combination of Conditional Quantile Forecasts Journal of Business & Economic Statistics Vol. 23, No. 4, pp. 416-431
67. Rachev, S., C. Menn and F. Fabozzi (2005), Fat-Tailed and Skewed Asset Return Distributions. New Jersey: Wiley
68. Riskmetrics (1996). Technical Document. Technical report. J.P. Morgan
69. S. Campbell (2005), A Review of Backtesting and Backtesting Procedure, Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington D.C.
70. S. Haykin (2009), Neural Networks and Learning Machines. McMaster University Hamilton, Ontario, Canada
71. S. Manganelli, R. F. Engle (2001), Value at Risk Models in Finance. Working Paper n. 75. European Central Bank

72. T. Bollerslev (1986), ‘‘Generalized Autoregressive Conditional Heteroskedasticity,’’ Journal of Econometrics,31: 307–27.
73. T.S. Beder (1995), VaR: Seductive but Dangerous, Financial Analyst Journal, Sep-Oct, 12-24
74. U.S. Securities and Exchange Commission-Mutual Funds and ETFs, A Guide for Investors
75. Vanguard Asset Management, Limited (2015), Understanding ETF liquidity and trading
76. W. Enders (2014), Applied Econometric Time Series. Wiley
77. W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, and J. Miller, (2018) Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning. International Conference on Learning Representations - ICLR, vol. 79, pp. 1094–1099.
78. W. Sun, S. Rachev, Y. Chen, F.J. Fabozzi (2009), Measuring Intra-Daily Market Risk: A Neural Network Approach
79. Y. Maluf, Otávio Ribeiro de Medeiros (2014), Value-at-Risk of Brazilian ETFs with Extreme Value Theory Approach
80. Y.Liu (2005), Value-at-Risk Model Combination Using Artificial Neural Networks