Politecnico di Torino

Master's Degree in Physics of Complex Systems

Master's Degree Thesis
developed at the Department of Mathematics, King's College London

Complex Systems Approaches to Financial Markets: Scaling Relations, Correlations and Physics Inspired Modeling

Supervisors

Prof. Luca Dall'Asta

Prof. Tiziana Di Matteo

Prof. Luciano Pietronero

Dr. Andrea Zaccaria

Candidate

Matteo D'Alessandro

Academic Year 2022-2023

Abstract

Complex systems are usually composed of many interacting parts which together give rise to the emergence of collective behaviours and properties that need proper tools to be analyzed. One of the main examples of complex systems, characterized by heterogeneous interconnections between their single components, are human societies and a paradigmatic case study of certain aspects of these systems, particularly in terms of economic interactions, are financial markets. The price of an asset is indeed just the consequence of the interplay between various agents that act at different time scales and are influenced by the external environment. Studying the statistical properties of financial data is therefore an opportunity to figure out how and why some particular features arise in the context of complex systems, and which are some effective tools to quantify these peculiarities.

This Master's Thesis project is about various fundamental tools coming from statistical physics and complexity science in general, that are used to analyse economical and financial systems. In particular the dissertation focuses on the empirical study of financial data, employing different methods and ideas inspired by a physical approach to the analysis of complex systems: scaling relations, correlations and physics-based modeling. Investigating the scaling properties of financial time series can give crucial insights about the underlying processes generating the empirical observations and can provide useful tools to detect consistent patterns across different scales. The analysis of the statistical relation between the price of different assets can offer the possibility to quantify their interactions and to extract the important information contained in the correlation structure of financial markets. A physics-based modeling approach to financial data may offer a unique perspective which can provide interesting insights into the underlying mechanisms and dynamics of the system.

The data used consist of a set containing all the daily closing prices from 1990 to 2022 of the stocks comprised, as of November 2022, in the S&P 500 index.

In the first part of the project, some of the main empirical statistical properties of financial time series found in literature are retrieved on the data: heavy tails, aggregational Gaussianity, absence of autocorrelations and volatility clustering. Particular attention is placed on the estimate of the tails exponents of the distribution of the returns, hallmark of the non-Gaussianity of financial data.

The second part of the work is about the main ideas behind the emergence of scaling laws in complex systems and their study. Particular focus is placed on the multifractal analysis of financial markets, its theoretical foundation and the methods to apply it. The Generalized Hurst Exponent method, as well as some of its extensions, is presented and applied to the data set to extract the so called multiscaling proxy and the Hurst exponents of the time series.

In the third part of the thesis different kind of correlation measures are presented and exploited to study the statistical relations between the time series of the stocks in the data set, both statically and dynamically. Moreover, the correlation structure of the market is represented with a complete graph and an information filtering technique taken from network theory (Minimum Spanning Tree) is employed to highlight peculiar clustering properties of the data.

The last part of the project is devoted to the study of a relatively new random walk model by Takayasu et al. (2010), named the PUCK model, in which the random walker is subjected to a potential centered at its moving average position. The model is presented and its application as a novel type of time series data analysis tool, characterizing the time-dependent stability of markets, is shown. Finally, its scaling properties are investigated through the previously illustrated methodology and a relation between its parameters and the scaling exponents is devised.

Table of Contents

List of Tables

List of Figures

Acronyms

GICS

Global Industry Classification Standard

GHE

Generalized Hurst exponent

RNSGHE

Relative normalized and standardized generalized Hurst exponent

ACSR

Autocorrelation segmented regression

MST

Minimum spanning tree

PMFG

Planar maximally filtered graph

Chapter 1

Introduction

Modern science uses the expression "complex systems" referring to systems typically made up of multiple interacting components that together give rise to collective behaviors and properties which need proper tools to be analyzed. One of the main examples of complex systems are financial systems, being composed of many agents that interact heterogeneously in a complicated way and the agent themselves being complex individuals or groups who behave based on both rational decision-making and emotions [1].

In particular, financial markets are open systems where many subunits interact nonlinearly in the presence of feedback, and are characterized by many participants interacting among each others with various strategies at different time scales and frequencies. The price of an asset is indeed the consequence of this complicated interplay between the constituent components of the systems and the external environment; and to provide a more complete picture one also needs to take into account the structure of dependency between different financial assets. Studying this kind of systems can therefore be extremely challenging but it could be also an opportunity to figure out how and why some particular features arise in the context of complex systems, and which are some effective tools to quantify these peculiarities [2, 3].

Following these ideas, around 1990, some physicists started to gain interest in exploring the complexity of financial systems and one of the first works belonging to this stream was «Lévy walks and enhanced diffusion in Milan Stock-Exchange» by Rosario Nunzio Mantegna, who published this innovative paper by showing the violation of the central limit theorem on the stock market [2, 4]. From this point on, the physics community started to understand the importance of non-Gaussian processes in financial markets along with their multiscale and scale-free properties, and many researchers began to work on economics problems to test a variety of new conceptual approaches deriving from the physical sciences [2, 5].

To be fair, one of the first demonstration of interest in social and economic systems from a physicist point of view came from Ettore Majorana, who wrote a pioneering article in which he pointed out the essential analogy between statistical laws in physics and in social sciences, in a period in which for the first time the determinism of classical physics was being questioned by the advent of quantum mechanics [6]. It is however thanks to the growing digitalization of society happened in the Nineties, that a very large number of data began to become available in various fields and particularly in financial markets, for which every single transaction or changes in financial prices was recorded [2].

According to Bikas Chakrabarti, this new interdisciplinary field was named with the term "Econophysics" in 1995, at the second Statphys-Kolkata conference in Kolkata (India), by the physicist H. Eugene Stanley, who was also the first to use it in print [7, 8]. Econophysics can indeed be defined as an interdisciplinary field that applies the methods of statistical physics, non linear dynamics, network theory to macro-micro/economic modeling, financial market analysis and social problems [5, 9].

Many may wonder how two disciplines that appear to be so different can be related; however making a deeper comparison between economical and physical systems, one can highlight many analogies that make clear why tools from the physical sciences are believed to be useful for an economic system. Many tools of statistical mechanics or statistical physics, are built to extract the average properties of a macroscopic system from the microscopic dynamics, and a lot of economical systems are characterized by various agents competing in a dynamically changing environment. Moreover, economic systems may be investigated on various size scales and in order to understand their global behavior concepts such as stochastic dynamics, correlation effects, self-organization, self-similarity and scaling are needed [8, 10].

One of the biggest contribution of econophysics up to now has been in the data analysis, thanks to the identification of empirical regularities and stylized facts of the distributions of the returns of financial assets, and the design of mathematical models and tools for dealing with such a vast amount of data [2, 8, 11]. It is indeed extremely important to always take into account, when studying these systems, the real nature of the underlying processes, that can only be extracted and understood by beginning the analysis from the statistical properties of real world observations.

From the early onset of this relatively new field, it appeared clearly how real financial data, in particular in the form of financial time series, are characterized by non-normal statistical properties and large fluctuations, which must be taken into account by investigating the so-called "tails of the distribution". The underlying processes in financial markets are indeed frequently characterized by infinite variance, and for this reason the central limit theorem becomes unsuitable for the analysis of complex systems especially belonging to this particular framework [4, 2].

If one focuses on the statistical analysis of price fluctuations, it can be very useful to investigate the so-called *scaling* relations between their different probability distributions at different time scales. This features of the price fluctuations can in fact give useful insights about the aggregate statistics of the underlying process and its diffusive properties [2, 9].

As the violation of the central limit theorem must necessarily be taken into account, also the validity of the random walk as the stochastic process describing price fluctuations has to be questioned, mainly for the fact that most of the real processes are correlated and not uniscaling. By uniscaling it is meant that, if one describes the (log) price of an asset with a random process, its scaling properties, as the scaling of the distribution of the q-moments of the price variation, can be simply described by a single parameter which is directly related to the model's fractal dimension [2, 9].

It is instead widely accepted nowadays that the kinds of processes encountered in finance, and in many complex systems, are characterized by scaling relations which are very complicated and not simply fractal. To describe real world data are thus needed models with multifractal (or multiscaling) properties where, for instance, the scaling of the distribution of the q-moments of the fluctuations of the prices needs a proper function to be characterized, which is simply a spectrum of scaling exponents. This particular feature is peculiar of complex systems and it derives from the fact that the systems' properties at a given scale are not preserved at different scales [3, 2, 9].

A further peculiarity, which concerns the collective dynamics of financial systems, is the network of interactions and interdependencies between the elements that constitute the complex structure that is being analyzed. For instance in financial markets, different assets display a high cross-dependence due to common flows of information and similar investment strategies; getting information about this framework can be challenging but crucial to understand deeply how the systems evolves as well as how each element can impact on the others in periods of high instability such as crises or crashes [2, 3, 9].

This thesis project is aimed at reviewing some important results obtained in the field of econophysics in the last decades, with particular emphasis on the concept of multiscaling from both a theoretical and an empirical point of view.

In the first chapter it is presented a theoretical review of the already mentioned "stylized facts", and their retrieval on real financial data. The analysis are completed with an interpretation of the results in the context of the Global Industry Classification Standard (GICS), with the purpose of showing the potentiality of these measures for practical applications.

The second chapter is devoted to the mathematical introduction of the concept of multiscaling (or multifractality) and the presentation of some stochastic models which include this feature. Different methods to measure this property on financial time series are described, and they are tested on real data. The results are reviewed and employed to point out important traits of financial assets derivable from their scaling relations.

The third chapter focuses on the measures of correlations between financial time series belonging to the same market, both dynamically to study their evolution, and statically in order to highlight the network of interdependencies between different financial assets.

The last chapter of the project is dedicated to the presentation of a relatively novel stochastic model by Takayasu et al. (2010) for the price of financial assets and to its analysis in term of the concepts shown in the preceding chapters. In particular, its scaling properties are investigated with the objective to associate the model's parameters with the well-established scaling measures for financial time series.

Chapter 2

Stylized Facts

Since its birth in the nineties, Econophysics has strongly focused on the study of great amount of financial data both to devise their distinguishing features and to design effective models. In particular, many studies about the statistical properties of financial time series have been carried on, revealing a set of features common across different instruments, markets and time periods which are acknowledged as *stylized facts*. Financial time series have indeed always been of great interest both to practitioners and researchers, but it was thanks to the advent of stock exchange computerization, happened in the Eighties, that all transactions in financial markets started to be recorded with great detail [8, 2]. Thanks to this huge data collection opportunity, it became possible to analyze the sequence of prices of any asset at different time frequencies and nowadays the so-called high frequency data are recorded every millisecond [8, 11]. It is therefore important to understand which are these *stylized facts* as they are fundamental to describe the empirical properties of financial systems as well as to figure out the main traits which characterize the stochastic processes underlying the price movements. These *facts* are indeed usually used as a benchmark to test the effectiveness of financial models, which clearly need to reproduce the features of real data in order to be considered potentially effective [12, 13].

2.1 The Data Set

In this work all the analyses are performed on the time series of the daily last traded prices from 02-Jan-1990 to 30-Nov-2022 of the stocks comprised, as of the last date, in the S&P 500 index. The data is provided by the Bloomberg Inc. platform and the specific price employed is identified on the database by the label "PX_LAST". It represents the most recent price at which the security traded before the market closed and it is commonly used to track the closing prices of various financial instruments, such as stocks, bonds, commodities, and currencies, on a daily basis. The choice of using daily data comes from the fact that they are easier to analyze and it's simpler to find errors in recorded prices. Moreover, the algorithms used in this work have been presented and tested in literature using this specific frequency and there is no particular reason for which a change is needed.

Regarding the decision to study stocks comprised in the S&P 500 index there are several reasons:

- The S&P 500 index is widely considered as a benchmark for the overall performance of the U.S. stock market, as it includes 503 large-cap companies across various sectors, providing a *representative* sample of the U.S. economy [14].
- Stocks in the S&P 500 tend to have *higher liquidity and trading volume* compared to stocks outside the index, making it easier to obtain accurate price data [14].
- The use of S&P 500 stocks as a standard allows for better *comparability* with previous existing literature [11, 8].
- The S&P 500 index is widely *followed* by market participants, analysts, and investors and by studying the stocks which belong to it, one can gain insights into market behavior, price dynamics, risk factors, and investor sentiment that are relevant to investors and policymakers [14, 15].

When dealing with large real world data sets it is necessary to check if the data are reliable, complete and consistent. In particular, in the case of financial time series it is important to check if there are missing data or if some of them are wrong due to some errors performed by the computerized collection system. For the data set analyzed no missing values are present because automatically the Bloomberg software can be programmed to fill a gap with the preceding value, avoiding the extraction of time series with "holes" [1]. On the other hand, for some time series there are long consecutive periods (weeks, months) over which the stock retains the same exact price, that is probably due to the fact that some portion of these time series are missing and are filled by the software always with the same value.

To detect which stocks are characterized by this issue a simple algorithmic procedure is designed and reported in Appendix A.1. The original data set utilized consists of a 8587×503 matrix whose columns are the time series of all the stocks. For each of them the date of the first recorded price can be different while the ending date is the same (30-Nov-2022). This is due to the fact that not all stocks comprised in the index, as of 30-Nov-2022, were traded on the market from the starting date of the data set (02-Jan-1990). On the Matlab software missing values are stored as NaNs (not a number), and in the following pages this symbol will be used to indicate when a value in the data set is not recorded.

Using the code in Appendix A.1 on the initial data matrix and setting `dt` $= 50$ days, 9 stocks have been removed: Garmin Ltd, Incyte Corp, Monster Beverage Corp, Phillips 66, Schwab (Charles) Corp, Skyworks Solutions Inc, Bio-techne Corp, Take-two Interactive Software, United Airlines Holdings Inc. The validity of the choice is also confirmed by an accurate visual inspection of the data set, and given the fact that only a tiny fraction of the total set faces this issue, it has been chosen to simply exclude these stocks form the analysis.

As an example the time series of United Airlines Holdings Inc is shown:

[1] When a market holiday is present the reported price is the previous day price.

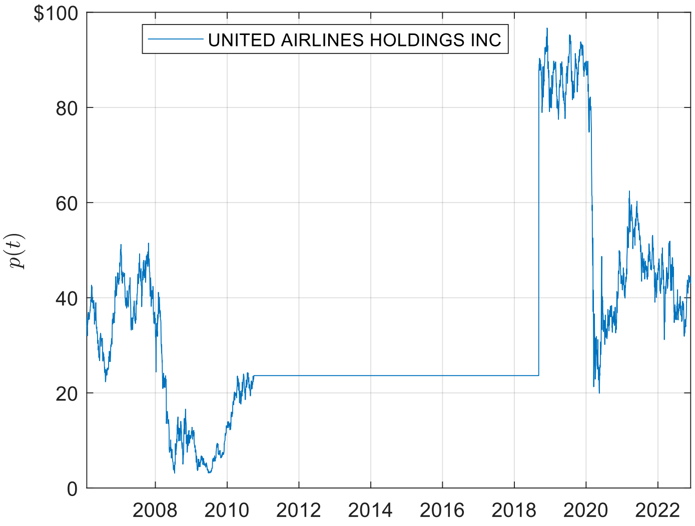

Figure 2.1: Time series of United Airlines Holdings Inc in which is clearly visible an error made by the data recording software.

It appears clear how a section of the time series has faced problems in the recording and thanks to the devised procedure the issue is detected and thus the stock is removed.

After the cleaning procedure the data set comprises 494 stocks, which can be visualized using a pie chart in which the various companies in the data set are classified according to the sectors of the GICS Global Industry Classification Standard) [16]. This classification is a widely used framework for categorizing companies into industry sectors and sub-industries. It was developed by MSCI (formerly Morgan Stanley Capital International), S&P and Dow Jones Indices to provide a standardized and globally recognized classification system for investors and financial professionals [16].

The GICS consists of four hierarchical levels: 11 *Sectors*, which represent broad segments of the economy: Communication Services, Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Real Estate, Utilities; each sector is divided into *Industry Groups*, for a total of 25 of them; industry groups are further divided into specific *Industries*, for a total of 74 of them; at the lowest level, industries are further segmented into *Sub-industries*, for a total of 163 of them, providing a more detailed categorization [16].

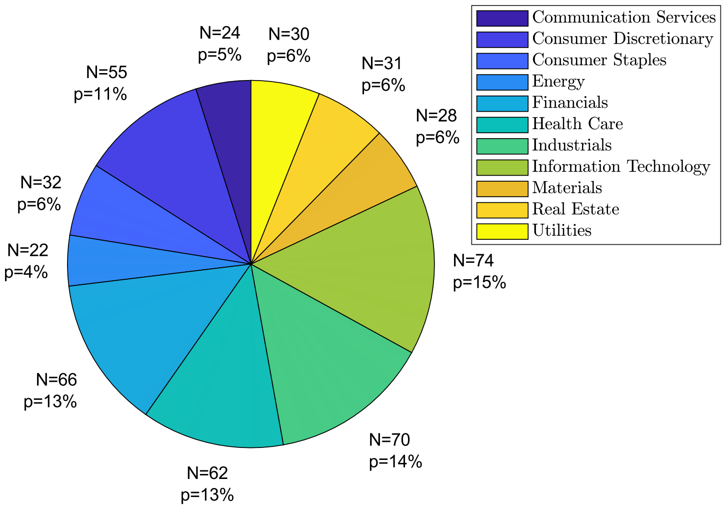

Figure 2.2: Pie chart of the data set employed in the project in which the number and the fraction of stocks belonging to each GICS *Sector* is reported.

Given that the S&P 500 index is built in such a way to be a *representative* sample of the U.S. economy [14], the data set analyzed includes all the *Sectors* present in the market.

2.2 Empirical Stylized Facts

In this section four of the most important empirical stylized facts, which can be found in literature [8, 11], are reproduced on the data set and also an investigation of some results within the GICS Sector classification is proposed. Most of the following analysis are performed on the distribution of the asset returns and therefore a proper definition of these quantities is needed. Let $p(t)$ be the price[2] of a financial asset at time t, the (log) return over a period of time τ is defined as:

$$r_\tau(t) = \log p(t+\tau) - \log p(t) \tag{2.1}$$

This quantity is nothing but the logarithm of the relative price variation, and it can assume both positive and negative values depending on the movement of $p(t)$ over the period τ. If $\tau = 1$ day the quantity $r_1(t)$ is referred to as daily return.

[2] Daily last traded price in the analyzed data set.

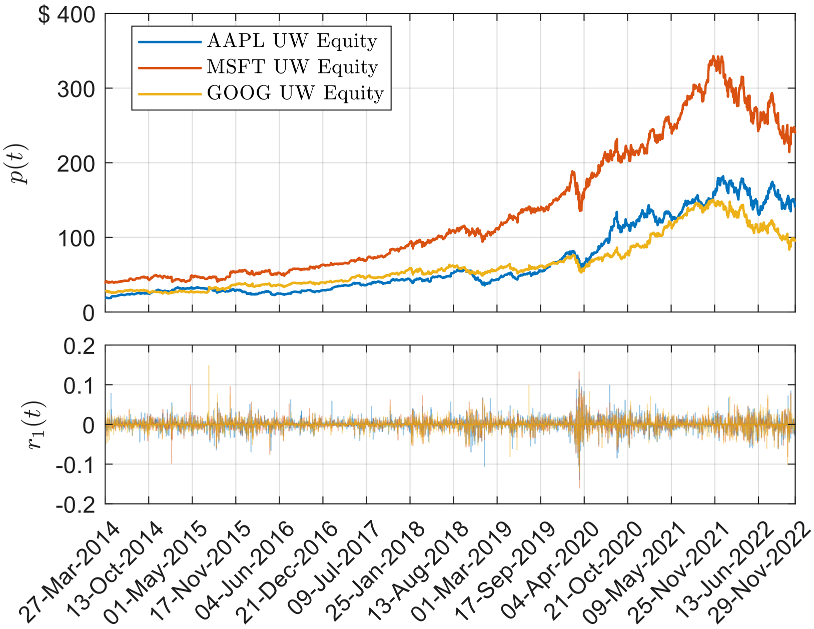

Figure 2.3: Plot of the top 3 most capitalized stocks of the data set: Apple Inc, Microsoft Corp, Alphabet Inc. On the top the price time series, on the bottom the daily return time series. The data is shown from Mar-2014, and not from the beginning of the data set, just to make the plots more clear.

When performing statistical analysis of returns it is important to pay attention to the presence of trends in the data that could affect the validity of the investigation. In order to remove this possibility, the best-fit line (in the least-squares sense) is removed from the $\log p(t)$ time series of each stock, and only after this operation the returns are computed. The idea of this transformation is taken from the works by Giuseppe Brandi, Tiziana Di Matteo et al. [17, 18, 19]. In this way the properties of the distribution of the returns of each stock are not affected by the presence of price trends and their study can give important insights about the empirical features of the analyzed market.

2.2.1 Heavy Tails

Since the first studies on the empirical distribution of asset returns [20], it appeared clear that the normal distribution wasn't valid for modelling and describing financial processes. Non-normal fluctuations are indeed common in financial systems and in other quantities relevant to economics, and therefore implementing methods which can help to estimate the behavior of the probability distributions in the region of large and rare variations, known as the "tail" of the distribution, is crucial [9, 8, 11, 21]. For instance, within the context of risk management, the study of these tails can provide a better understanding of the potential magnitude and likelihood of extreme losses, enabling to make more accurate risk assessments and enhancing the robustness and accuracy of models [9].

One way to measure the deviation of a distribution from a gaussian is to compute the so-called excess kurtosis:

$$\gamma = \frac{\langle (r_\tau(t) - \langle r_\tau(t) \rangle)^4 \rangle}{\sigma_\tau^4} - 3 \tag{2.2}$$

where for the real measures the average values and the variance σ_τ^2 are replaced with the empirical averages. Knowing that $\gamma = 0$ for a gaussian, it follows that empirical distributions which show a value higher than 0 are characterized by the so-called "fat tails" [1, 8, 11].

The results of the excess kurtosis γ for the distributions of the daily return of the data set are reported in the subsequent plot:

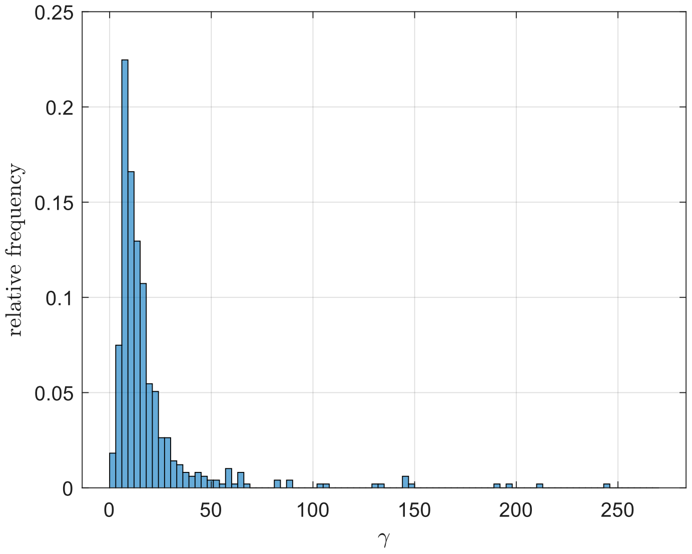

Figure 2.4: Histogram of the excess kurtosis measured on the time series of the daily returns of the stocks in the data set. Large kurtosis characterize the whole set meaning that most of the returns distributions are fat tailed.

The bin with the highest relative frequency corresponds to values of γ between 6 and 9, and it can be observed that almost every time series in the data set is characterized by a large kurtosis compared to the gaussian. Distributions with this feature are called *leptokurtic* and in financial systems they imply a larger probability to have extreme returns, both positive or negative [1].

One of the reason why in these systems empirical distributions deviate from a gaussian is the violation of the central limit theorem (CLT) [4, 1, 9]. Indeed, the underlying processes in complex systems frequently do not satisfy the assumptions of the CLT:

- these processes are usually not purely additive and therefore one cannot assume that they are a simple sum of variables [12, 22];
- the variables in complex systems are nearly always correlated, making impossible to consider them independent [8, 11];

- there's no guarantee that the variance of these processes is finite [1];
- the variables are often not even identically distributed [1].

To get a more quantitative perspective about the tails of these distributions it is common to study the so-called complementary cumulative distribution, defined as:

$$P_{>}(s;\tau) = \int_{s}^{+\infty} p(r_\tau)dr_\tau \tag{2.3}$$

and in different works [11, 23, 24] it has been shown that this quantity can be assumed to asymptotically follow a power law:

$$P(r_\tau > x) = P_{>}(x;\tau) \sim x^{-\alpha} \tag{2.4}$$

It means that by quantifying the parameter α one can get important insights about the statistics of extreme values. For instance, if the cumulative distribution function of a process satisfies equation (2.4), it follows that only the first $n < \alpha$ moments are finite [1]. It is also significant to underline that the distribution of the returns depends on the time scale τ over which they are calculated, and in the next subsection more details on this feature will be presented.

In this work, in order to reproduce the results obtained in literature, it is chosen to estimate the tail exponents α of the daily returns of all the stocks in the data set. Instead of simply applying a least square linear fitting in log-log scale, it is preferred to employ a maximum-likelihood fitting method, which has been proved to provide more reliable results [24]. Assuming that the daily returns are drawn from a distribution that follows a power law of parameter α for $x > x_{min}$, by maximizing the likelihood of the observations given the model, the following estimates are obtained for the exponent of the complementary cumulative distribution functions [24]:

$$\hat{\alpha} = n\left[\sum_{i=1}^{n} \ln \frac{x_i}{x_{min}}\right] \tag{2.5}$$

$$\sigma = \frac{\hat{\alpha} - 1}{\sqrt{n}} + O(1/n) \tag{2.6}$$

n indicates the number of empirical points used for the fit, and it has been proved that the estimate is valid for $n > 50$ [24]. To follow this prescriptions, stocks for which there are less than 50 points are removed from this analysis (31 stocks). Regarding x_{min} it is chosen x such that $P(r_1 > x) = 10^{-1}$, and thus the computations are performed using all the points falling in the range $P(r_1 > x) \in [10^{-3}, 10^{-1}]$. In this way 2 decades of points are used to estimate the parameters and the large values which usually deviate from the power law are also taken out. The same quantities and assumptions are also employed for the negative tails.

As an example, it is displayed the complementary cumulative distribution for the most capitalized (as of 30-Nov-2022) stock of the data set, namely Apple Inc. The plot is realized in log-log scale to highlight the power law behaviour, and a comparison is made with a Gaussian with same average and standard deviation of real data.

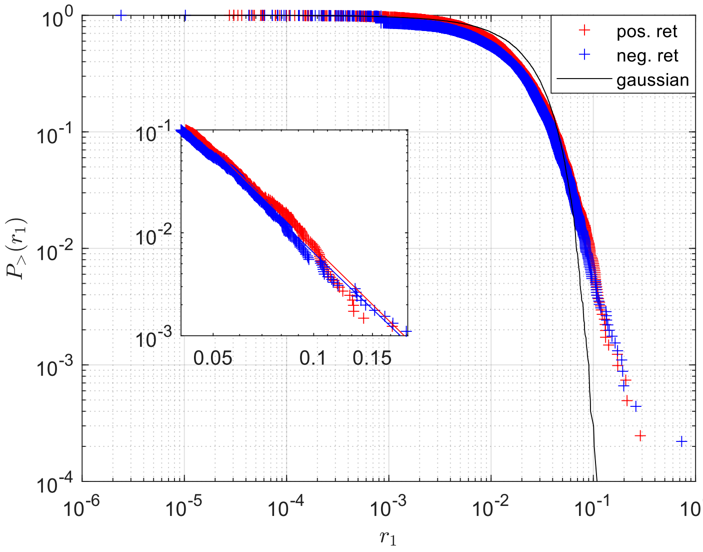

Figure 2.5: Apple Inc empirical complementary cumulative distribution for daily returns from 02-Jan-1990 to 30-Nov-2022. In the small figure, over the data points, a straight line whose slope is $-\hat{\alpha}$ computed with equation (2.5) is showed both for positive and negative tails.

It is clear how extreme returns are much more probable for a power law and why the Gaussian cannot be a valid model to assess risk when dealing with these kind of systems.

The results for all the considered time series are reported in the following plots:

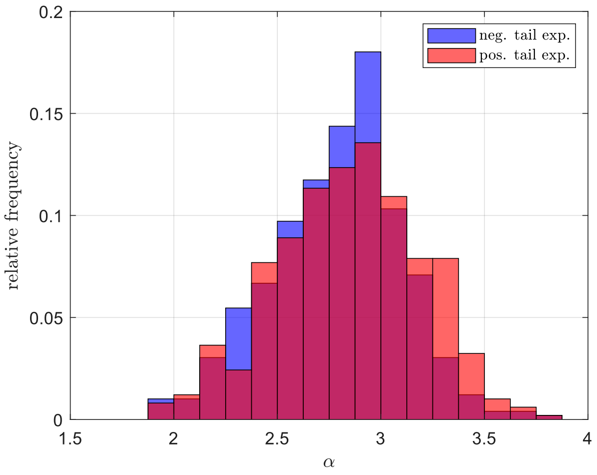

Figure 2.6: Histogram of the results of the maximum likelihood estimates (2.5) of the positive and negative power law tail exponents for the daily returns distributions of all the selected stocks in the data set.

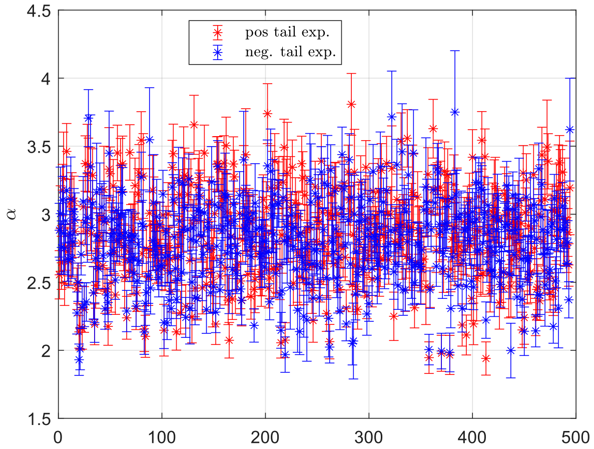

Figure 2.7: Results of the maximum likelihood estimates (2.5) of the positive and negative power law tail exponents for the daily returns distributions of all the considered stocks in the data set; on the x-axis each stocks is identified with a number between 1 and 494 (alphabetical order). The error bars are the σ in (2.6).

For financial time series the measured values of α approximately fall within the interval (2,5) [8, 9, 11] and the obtained results are in line with these findings. The positive and negative tail exponents of the same stock are in some cases also different, meaning that the empirical analyzed distributions can also be asymmetrical, having different probabilities for very large or very small daily returns. Just a few values are slightly smaller than 2 but considering their error bars the measures correctly exceed this value.

Tail Exponents by Sector

To complete the description about the heavy tails, the results are also presented within the GICS Sector classification [16], in order to highlight if there are some groups of stocks characterized by peculiar properties. To take into account the estimate errors on the exponents a weighted average is performed for each sector [25]:

$$\hat{\alpha}_{sec} = \frac{\sum_i \hat{\alpha}_i/\sigma_i^2}{\sum_i 1/\sigma_i^2} \tag{2.7}$$

$$\sigma^2(\hat{\alpha}_{sec}) = \frac{1}{\sum_i 1/\sigma_i^2} \tag{2.8}$$

having used for α_i the simple average between the two tail exponent and as the σ_i the squared sum of the two related estimate errors. The subsequent plot displays the results:

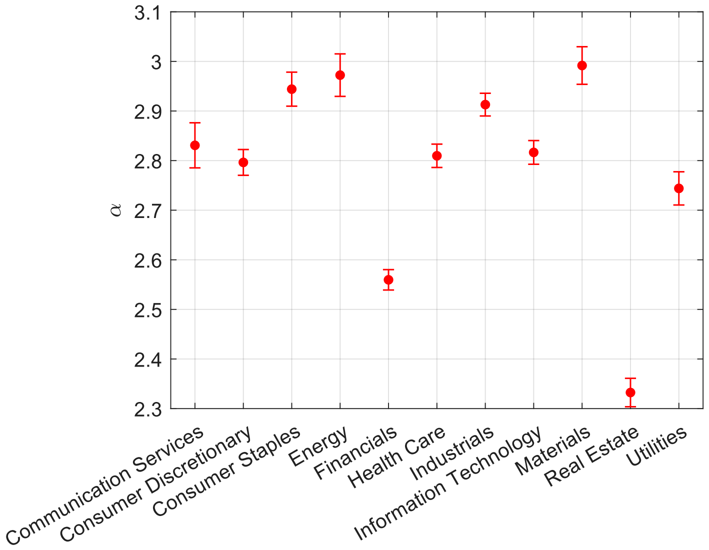

Figure 2.8: Weighted average tail exponent for each sector in which the companies in the data set are divided. For each stock it is used the average between positive and negative tail exponents and the squared sum of the two related errors. The error bar is the standard error of the mean.

An higher α implies that the corresponding sector contains stocks whose daily return distribution tails decays on average faster than the one of a sector with a lower α. In the investigated case, the Real Estate and the Financials sectors show a lower average α with respect to the others, which instead display similar values. To make more clear the implications of this property, the results on the two aforementioned sectors are shown, and for each one of them the daily return time series of the two stocks with the lowest $\hat{\alpha}$ and the one with the highest are presented:

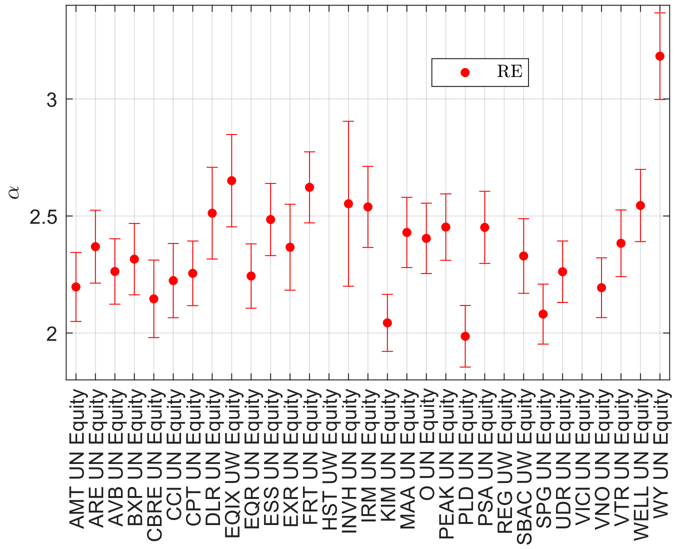

Figure 2.9: Simple average between positive and negative tail exponent for each stock considered in the Real Estate sector. The error bar is the squared sum of the two estimates σ. In correspondence of stocks for which there are not enough points to estimate α, no measure is reported.

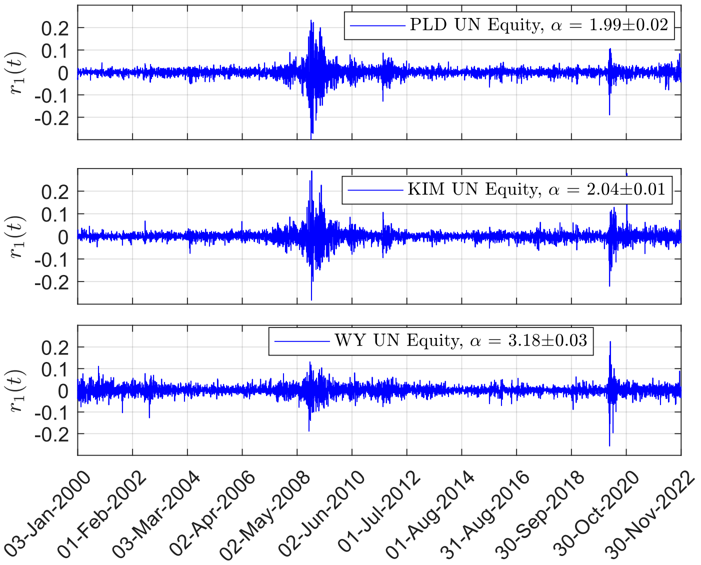

Figure 2.10: Daily returns time series for 3 stocks of the Real Estate sector. It is clear that when α is smaller (first two plots), the returns can assume more extreme values compared to cases when α is larger (third plot).

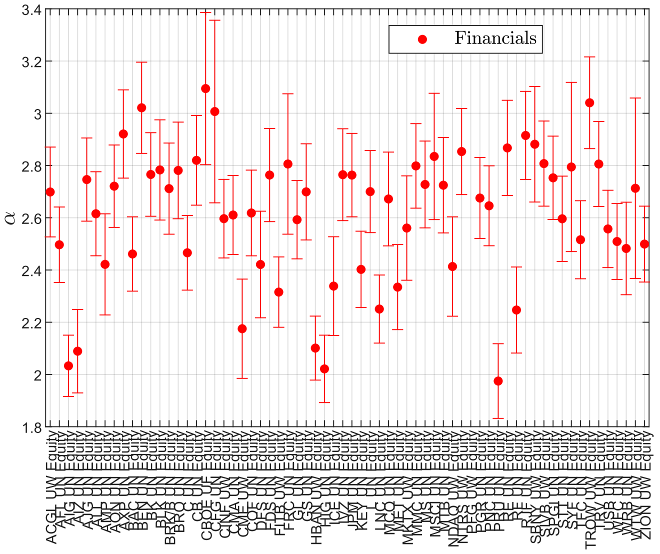

Figure 2.11: Simple average between positive and negative tail exponent for each stock considered in the Financials sector. The error bar is the squared sum of the two estimates. In correspondence of stocks for which there are not enough points to estimate α no measure is reported.

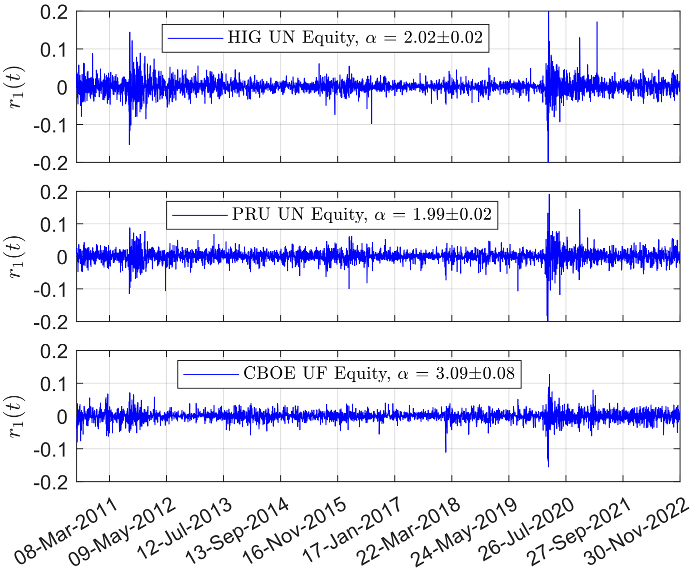

Figure 2.12: Daily returns time series for 3 stocks of the Financials sector. It is clear that when α is smaller (first two plots), the returns can assume more extreme values compared to cases when α is larger (third plot).

It is interesting to observe that potentially these tail exponents could be also exploited to gather companies, for instance in figure (2.12) both Hartford Financial Services Group Inc (HIG UN Equity) and Prudential Financial Inc (PRU UN

Equity) offer services related to Insurances and have a very similar α.

2.2.2 Aggregational Gaussianity

From previous analysis it is clear that studying the distribution of the returns of financial assets can give crucial insights to take into account extreme events in an effective way. Since now, the main focus has been placed on daily returns but it is widely recognized that the time scale τ over which returns are computed strongly affects the properties of their distribution [8, 11, 1]. Indeed, it has been observed that when τ increases, the heavy tail property of the distribution weakens, and it approaches a Gaussian form [8, 11, 26, 27]. This second stylized facts is called *aggregational Gaussianity*, and it is somehow the recovery of the validity of central limit theorem for large time scales. On the top of that, the notable dependence of (2.3) on τ means that the process underlying prices is not trivial for small time scales as there are peculiar effects, like the fat tails, which are less evident when τ increases. In the subsequent plots we can observe a clear example of this tendency and an overview over the whole data set.

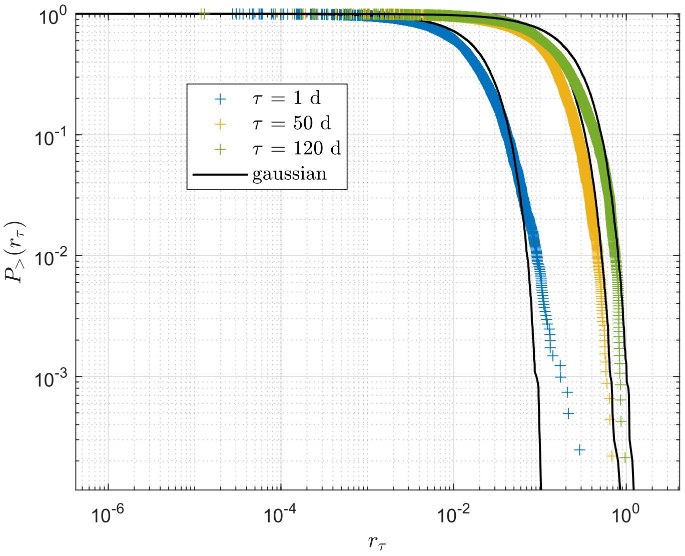

Figure 2.13: Empirical complementary cumulative distribution of Apple Inc returns from 02-Jan-1990 to 30-Nov-2022. It is evident that increasing the time scale τ over which returns are computed the distributions get closer to a Gaussian with average and variance computed on the empirical returns.

It is evident in this example that increasing τ the tails of the return distribution do not deviate in a consistent way from the Gaussian, and thus the normal assumption should not be discarded completely to model these quantities. On the other hand, when the time scales are small, $\tau = 1$ day in the example, the data strongly deviate from the Gaussian. To get an overview of the data set, the excess kurtosis γ of the returns distribution for various τ is computed for all the stocks in the data set.

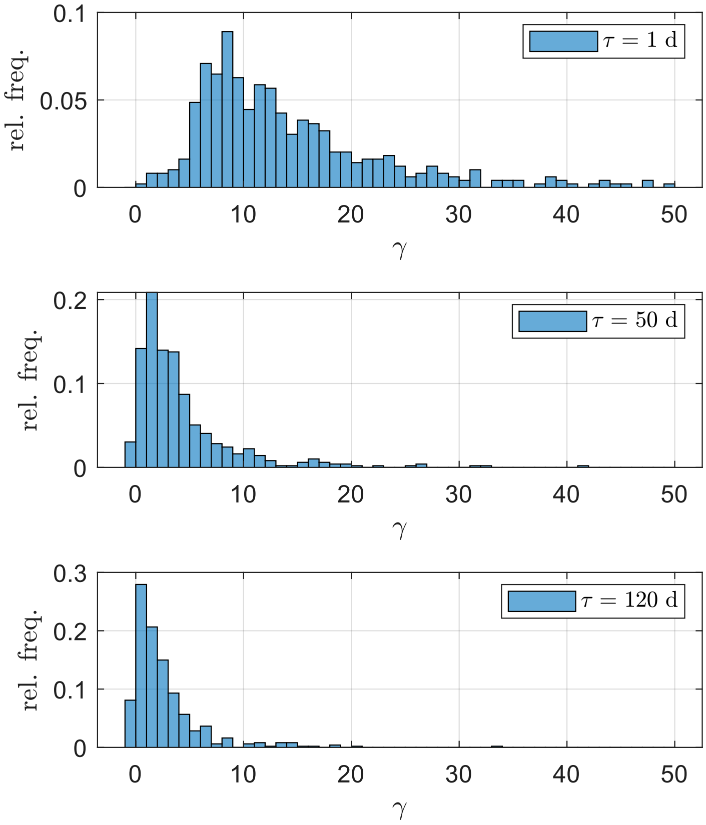

Figure 2.14: Excess kurtosis γ for the returns distribution at various time scales $\tau = [1, 50, 120]$ days of all stocks in the data set.

As τ grows the peak of the histogram get closer to 0, meaning that the empirical distributions get closer to the normal one.

2.2.3 Absence of Autocorrelations

When studying financial data it is fundamental to examine if there exists a correlation between a given time series and a delayed copy of itself. In this way one can detect if the system has memory and therefore if there are any systematic relationships, such as trends or patterns, between past and future observations. Usually the measure that is used to quantify this property is the Pearson correlation coefficient defined in this way:

$$\begin{aligned} C_\tau(\Delta t) &= corr(r_\tau(t+\Delta t), r_\tau(t)) = \\ &= \rho(r_\tau(t+\Delta t), r_\tau(t)) = \\ &= \frac{cov(r_\tau(t+\Delta t), r_\tau(t))}{\sigma_{r_\tau(t+\Delta t)}\sigma_{r_\tau(t)}} \end{aligned} \tag{2.9}$$

that on real data simply becomes the sample correlation coefficient:

$$\rho(r_\tau(t+\Delta t), r_\tau(t)) = \\ = \frac{1}{N-1}\sum_{i=1}^{T-\Delta t}\left(\frac{r_\tau(t+\Delta t) - \langle r_\tau(t+\Delta t)\rangle}{\sigma_{r_\tau(t+\Delta t)}}\right)\left(\frac{r_\tau(t) - \langle r_\tau(t)\rangle}{\sigma_{r_\tau(t)}}\right) \tag{2.10}$$

where the averages and the standard deviations are also computed on data [28, 29]. This quantity measures the linear correlation between two variables, that in this case are $r_\tau(t+\Delta t)$ and $r_\tau(t)$, and it is practically their normalized covariance. It can assume values in the interval $[-1{,}1]$ where 0 means no correlation, +1 positive linear correlation and -1 negative linear correlation [28, 29].

Regarding financial data, it is well known that returns do no exhibit significant autocorrelation [8, 11, 30]. The main reason is that if linear correlations between price variations existed, these could be simply exploited to devise arbitrage strategies which would immediately reduce them [11]. Therefore the third stylized facts is the *absence of autocorrelations* of returns and it is also acknowledged as an evidence of the validity of the efficient market hypothesis [11, 31]. In fact, this theory states that it is nearly impossible to consistently outperform the market as all relevant information about a security is quickly reflected in its traded price [32].

To showcase this property on the data set, the correlation of daily returns of all the available stocks is computed. The average over the whole data set is reported in the subsequent plot.

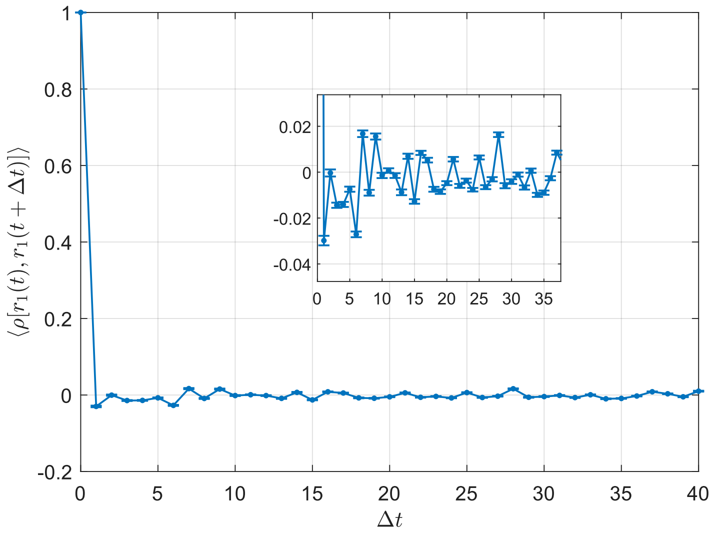

Figure 2.15: Autocorrelation for the daily returns of every stock in the data set averaged over the whole data set. The error bars are the standard error of the mean of each correlation measure.

From the plot it is evident that the autocorrelation immediately decays to zero and oscillates around this value for any time lag Δt employed for the calculations.

2.2.4 Volatility Clustering

Having observed absence of autocorrelations for returns one would argue that also their non-linear functions follow the same property and therefore that price increments are totally independent [11]. However, if one computes, for instance, the absolute value or the square of the returns, they exhibit a positive long-range autocorrelation that slowly decays to zero [8, 11]. This well known property is widely acknowledged as *volatility clustering* and was first described as the tendency of financial markets to experience extended periods of high volatility followed by extended periods of low volatility [20]. It suggests that large price movements tend to occur in clusters, rather than being randomly distributed over time, and together with the heavy tails, described in Section (2.2.1), it is a further evidence that prices cannot be simply modelled as random walks [11].

As an example the squared daily returns of Apple Inc are plotted and it is evident that large returns are often part of a cluster of high volatility while flat regions belong to groups of low price variations.

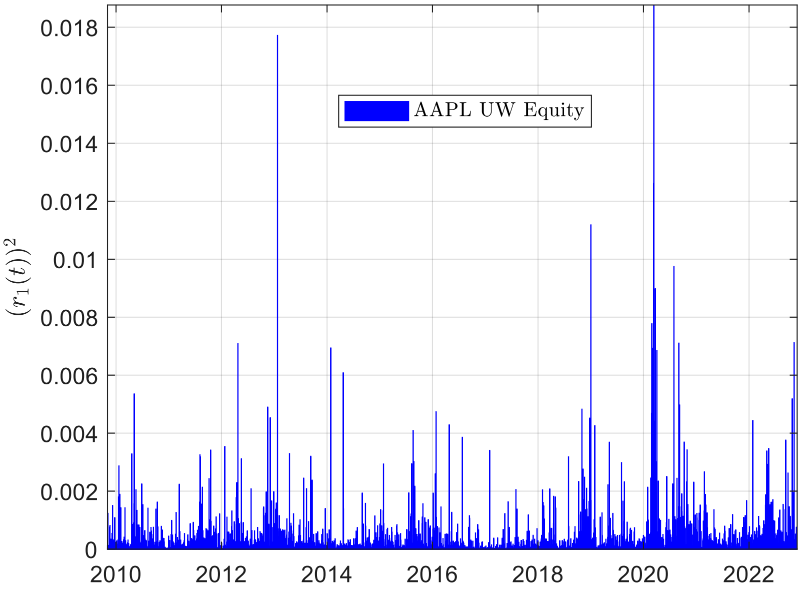

Figure 2.16: Squared daily returns of Apple Inc's time series in the period between 2010 and 2022. Many clusters of high or low volatility can be observed along the period.

To showcase this fourth *stylized facts* over the whole data set the average autocorrelation for the absolute and for the squared daily returns is computed for every stock and averaged over all of them. The results are also compared to the autocorrelations of the simple returns, already displayed in figure (2.15), in order to emphasize the described persistence.

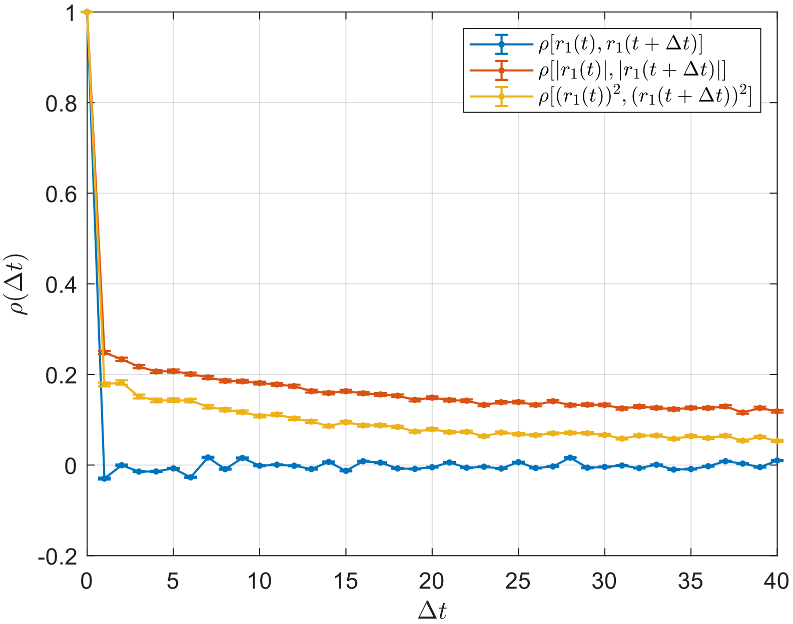

Figure 2.17: Autocorrelation for different functions of the daily returns of every stock in the data set averaged over the whole data set. The error bars are the standard error of the mean (standard deviation divided by the square root of the number of data point) of each correlation measure.

In many past works [8, 11] the autocorrelation of different powers υ of the absolute returns has been proven to approximately decay as a power law:

$$\rho(|r_1(t)|^{\upsilon}, |r_1(t+\Delta t)|^{\upsilon}) \sim (\Delta t)^{-\beta} \tag{2.11}$$

with β roughly in the interval [0.1,0.4] for absolute and squared returns [8, 11, 33], even if this range varies quite a lot across the literature.

With the aim of extracting these values from the analyzed data set, the autocorrelation of the absolute daily returns of each time series is fitted with a power law of the form $A(\Delta t)^{-\beta}$ and for each estimation Δt belongs to the set $[1, min(\Delta t_0^i$,1000)] where Δt_0^i is the last time lag for which the autocorrelation function is still positive. In order to avoid spurious results due to the scarcity of data points in a single time series, only stocks for which the maximum employed time lag $(\Delta t_{max}) < 0.3$ times the single time series' length are considered (93 stocks removed).

By way of illustration the power law fit is showcased again for the most capitalized stocks of the data set, namely Apple Inc. In this particular case the autocorrelation remains positive even for $\Delta t \sim 1000$ days and therefore all the available points are used for the fit. The standard error reported is simply computed from the 95 % confidence interval of the fit coefficients' results.

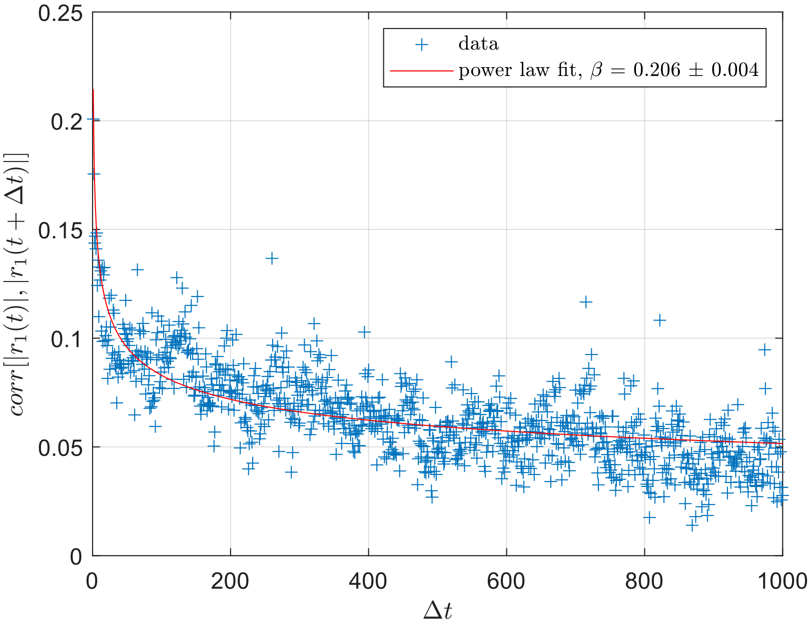

Figure 2.18: Example of the power law fit of the autocorrelation of the absolute daily returns from 02-Jan-1990 to 30-Nov-202 of Apple Inc stock. The red line is the resulting power law.

The obtained results for all the data set are reported in the following histogram:

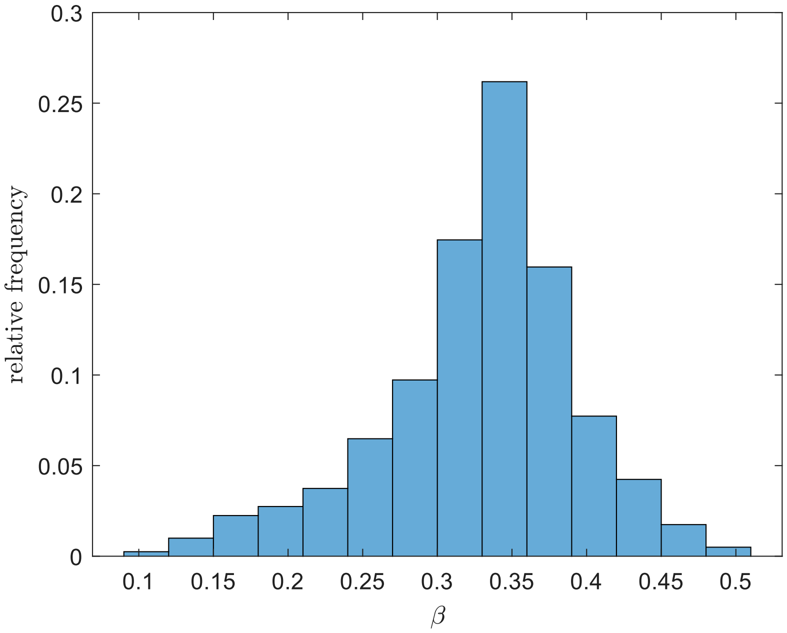

Figure 2.19: Histogram of the exponents β of the power law decay fit of the absolute daily returns autocorrelation of the stocks considered in the analysis.

The obtained values present a distribution which is sharply peaked around $\beta \simeq 0.35$ while the 2.5 and the 97.5 percentiles are respectively 0.169 and 0.4465. As a result, the 95% of the obtained values fall approximately in the interval reported in literature [8, 11]

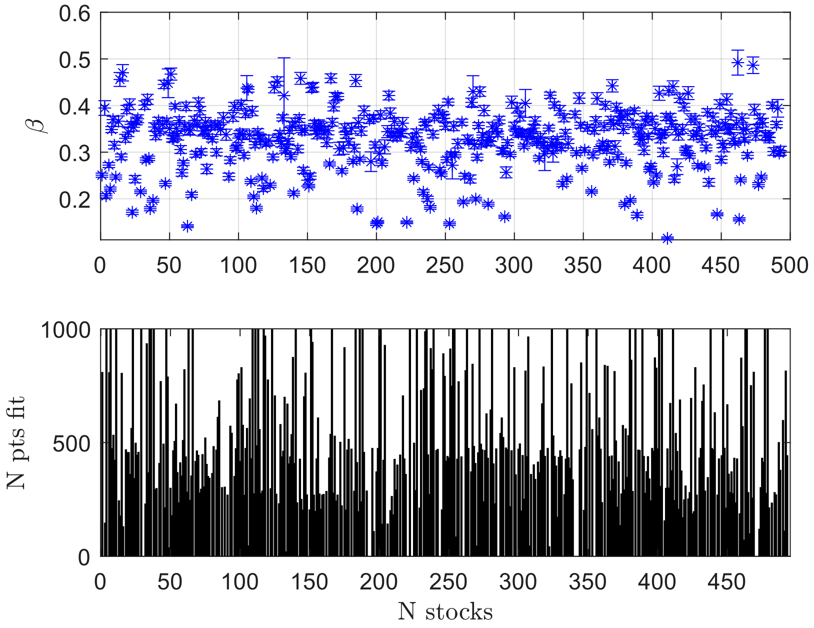

Figure 2.20: Results of the power law fits exponent for all the stocks considered in the analysis. Estimates with less points are, as expected, characterized by a larger standard error.

The results obtained must be taken with care as it is clear that there are also some estimates which are realized using few hundreds of points and in fact in the plot above these values are the one characterized by the highest standard errors. The purpose of these computations is however to give a quantitative estimate of the phenomenon of volatility clustering and therefore these exponents can in any case give an idea about how fast the analyzed autocorrelation decays.

Exponents by Sector

Just to complete this part, as done in section (2.2.1), the weighted average of β for each sector is computed and reported in the subsequent plot:

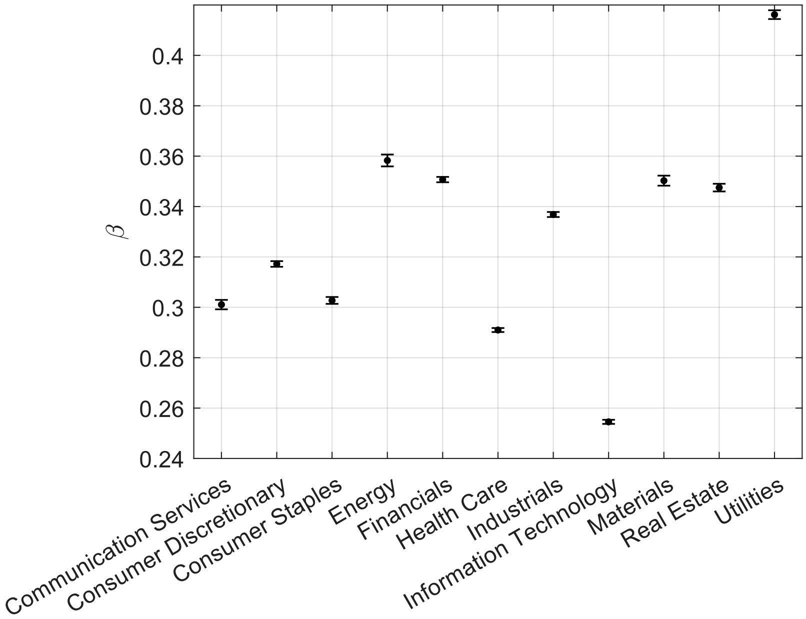

Figure 2.21: Weighted average (Equation 2.7) of the exponent β for each sector in which the companies in the data set are divided. The error bar is the standard error of the mean according to Equation (2.8).

An higher β implies that the corresponding sector contains stocks whose autocorrelation of the absolute daily return distribution decays on average faster than the one of a sector with a lower β. It means that stocks characterized by low values of the exponent β show longer and more evident clusters of large or small volatility. It follows that when one models the process underlying volatility this dependence cannot be neglected and these empirically observed features must necessarily be included [11].

In the investigated case, the Utilities and the Information Technology sectors show an average β which is further from the other, respectively higher for the former and lower for the latter. To make more clear the implications of this property, the results on the two aforementioned sectors are shown, and for each one of them the daily absolute return time series of three stocks are shown: the 2 with the highest (lowest) β and the other with the lowest (highest).

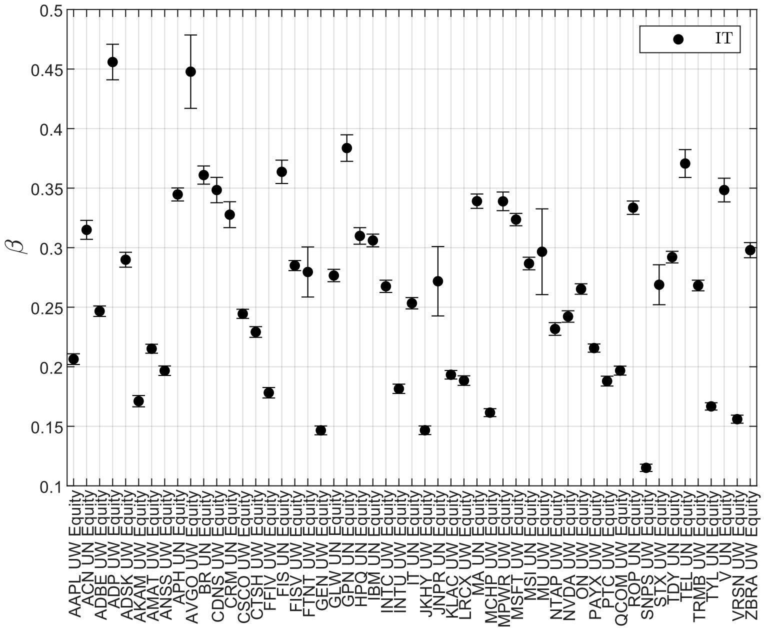

Figure 2.22: Results for the exponent β in the Information Technology sector. The larges standard errors are observed when a low number of points is used in the fit.

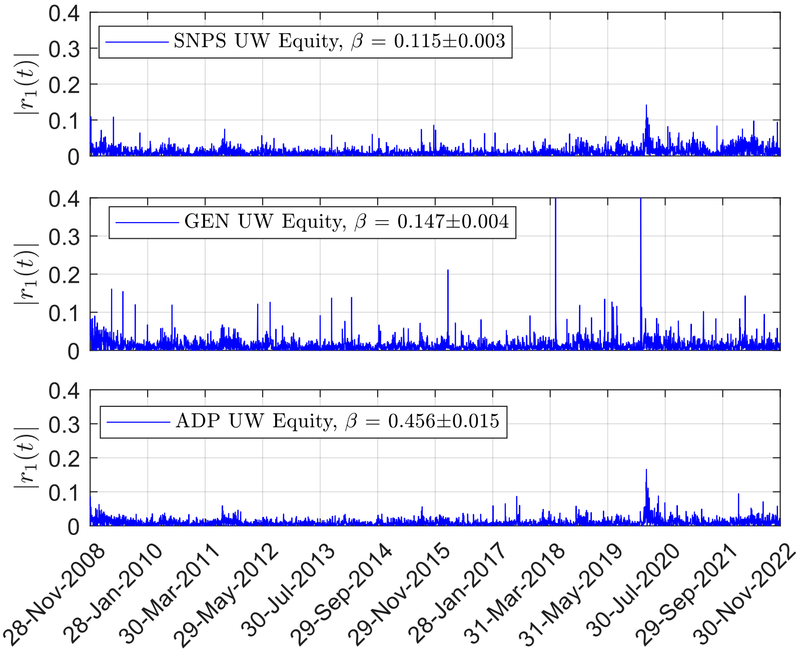

Figure 2.23: Absolute daily returns time series for 3 stocks of the Information Technology sector. It is clear that when β is smaller (first two plots), the volatility clusters are more evident compared to cases when β is larger (third plot).

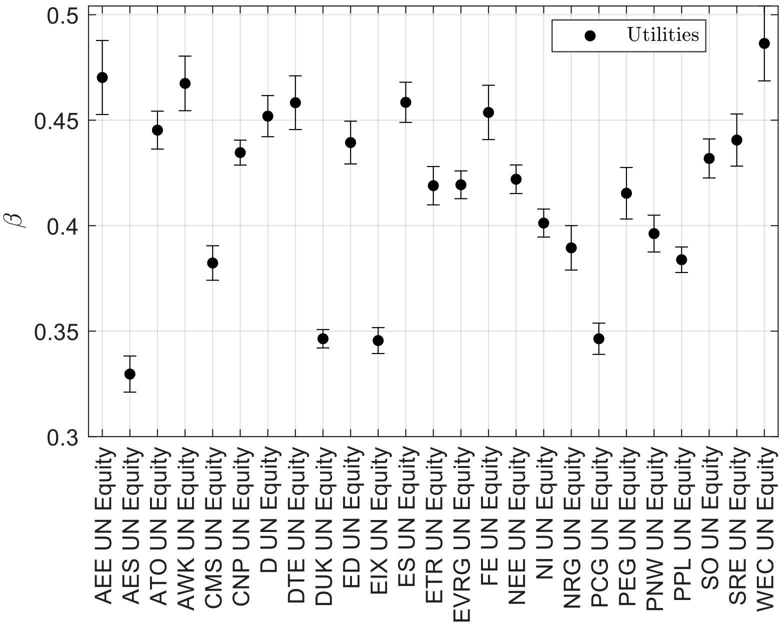

Figure 2.24: Results for the exponent β in the Utilities sector. The lower the number of points used in the fit, the larger the standard errors observed.

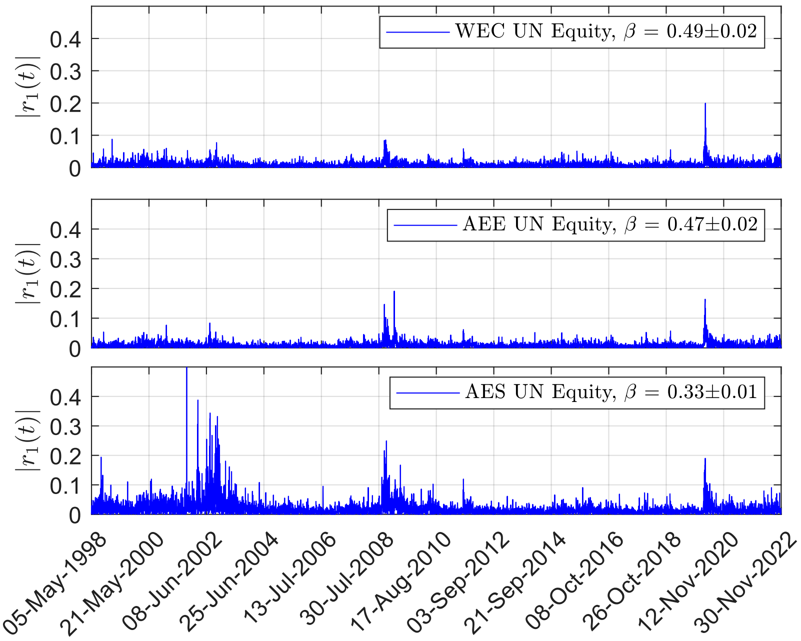

Figure 2.25: Absolute daily returns time series for 3 stocks of the Information Utilities sector. It is clear that when β is smaller (third plot), the volatility clusters are more evident compared to cases when β is larger (first two plots).

It is interesting to observe that stocks with $\beta \sim 0.1 - 0.2$ show a lot of clusters along all the analyzed period and therefore can be considered to be riskier as the price is more volatile. On the other hand, time series which give back a value of $\beta \sim 0.4 - 0.5$ show smaller absolute returns and clusters which persist for shorter periods. These last type of stocks are thus featured by more contained price variations and in this sense can be seen as safer assets.

One may wonder if the measures of β are somehow related to the measures of α previously described. Computing the Pearson correlation coefficient [28]:

$$\rho_{\bar{\alpha},\beta} = \frac{cov(\bar{\alpha},\beta)}{\sigma_{\bar{\alpha}}\sigma_{\beta}} \tag{2.12}$$

between the average among the positive and negative tail α and the β for each stock for which both the estimates have been computed (401 stocks), one gets that there's no statistically meaningful correlation. This implies that the two exponents describe different properties of the assets analyzed and thus must be taken into account simultaneously when making statistical analysis of financial assets.

2.2.5 Other Important Empirical Properties

Other than the empirical stylized facts presented, there are many other properties which are shared by financial data [11] and on some of them the next chapters will be focused. Mainly two empirical aspects of financial time series will be analyzed: scaling properties and correlations.

Scaling properties are well known in physics, statistics and mathematics and refer to how a system, a phenomenon, a law or a process behaves when its characteristic scale or size is modified. In the specific case of financial time series one looks for patterns or properties that are repeated at different time scales [34] and it will be shown how to quantify them and the consequence of these measures on the understanding of financial markets. A great focus will be placed on the concept of *multiscaling* and, other than presenting the cutting-edge methods to estimate it, it will be quantified on the considered data set to further confirm the fact that financial time series are indeed multiscaling [34, 35]. In a sense, this feature could have been also presented as the fifth stylized fact in this chapter but, given its importance in the setup of this work, an entire chapter will be devoted to its description.

Correlations are instead a measure of the interdependency between financial assets and can be useful to acquire important information on how much the various stocks are more or less influenced by the the movement of the others. Moreover, the study of the correlations of asset returns has been a valid instrument to study the hierarchical structure of financial markets and the presence of clusters of companies who share similar traits [8, 11, 36, 37]. In the following chapters this dependency measures will be presented and employed to get interesting insights about the market structure and its evolution.

Chapter 3

Scaling Properties

Scaling properties (or laws) refer to how a system or a phenomenon behaves when its characteristic scale or size is modified.

These are well known in statistical physics as are vastly used to empirically study the behavior near critical points of phase transitions. Moreover, their combination with the concept of universality and with the renormalization group technique has led to a powerful approach to study critical phenomena and dynamical systems, enabling researchers to uncover peculiar properties previously unknown [38].

In general, the study of scaling laws is fundamental in the analysis of complex systems for many reasons: it provides insights into the essential properties and dynamics of the phenomenon or of the process under examination; it is useful to make predictions about the behavior of the system at different scales; it may reveal universal principles that apply across different domains [39].

The study of scaling properties must be introduced giving a rapid overview about the concepts of self similarity and scale invariance, which are crucial to understand how profoundly scaling laws can tell about the nature of a process or a phenomenon under analysis.

3.1 Self-similarity, Fractals, Scale Invariance

This section is a rapid summary of the main concepts related to self-similarity, fractals and scale invariance exposed in the first chapter of [40], which can be useful for the next topics presented.

3.1.1 From Geometrical Fractals to Scale Invariance

Definition 3.1.1 *A similarity $\boldsymbol{S}$ is a linear transformation obtained as a composition of translation, rotation and uniform scaling (enlarging or reducing).*

It follows that two geometrical objects A and B are similar ($\sim$) if it exists a couple of similarity transformation $\boldsymbol{S}, \boldsymbol{S}'$ such that $\boldsymbol{S}(A) \cong B$ and $A \cong \boldsymbol{S}'(B)$.

Definition 3.1.2 *A generalized similarity transformation $\hat{\boldsymbol{S}}$ is the union of n similarity transformations.*

Definition 3.1.3 *A geometrical object O is self-similar if it exists at least a generalized similarity transformation $\hat{\boldsymbol{S}}$ such that $\hat{\boldsymbol{S}}(O) \cong O$*

A peculiar example of self-similar geometrical objects are the so called *fractals* defined for the first time by Benoît Mandelbrot in [41]. These objects are characterized by the property that their parts, when magnified by some suitable scale factor, look similar to the whole, and a fractal can be built recursively applying a generalized similarity transformation an infinite amount of times to a geometrical structure.

A typical example is the so called Cantor Set which can be built applying recursively the following generalized similarity transformation:

$$\hat{\boldsymbol{S}} = \boldsymbol{S}_1 \cup \boldsymbol{S}_2, \text{ with } \boldsymbol{S}_1 = \frac{1}{3}x \text{ and } \boldsymbol{S}_1 = \frac{1}{3}x + \frac{2}{3}$$

to the unitary one dimensional interval $I = [0{,}1]$. The fractal object is therefore defined by $\hat{\boldsymbol{S}}^{\infty}(I)$.

Definition 3.1.4 *The fractal (or Hausdorff) dimension d_F of an object O embedded in $\mathbb{R}^d$ is defined as:*

$$d_F = -\lim_{r \to 0} \frac{\log N(r)}{\log r} \tag{3.1}$$

where $N(r)$ is the minimum number of d-dimensional balls of radius r ($B_{x_0}(r) = \{x \in \mathbb{R}^d : \|x - x_0\| \leq r\}$) necessary to cover the object O (minimum with respect to the position of ball centers at fixed r);

Definition 3.1.5 *A fractal is a subset of an Euclidean space such that*

$$d_T < d_F \leq d$$

where d is its embedding dimension (i.e the dimension of the space in which the object lives), d_T its topological dimension (i.e. the local dimension of each point of the object).

For instance the Cantor Set is formed by point-like objects ($d_T = 0$) and is embedded on the real line ($d = 1$). Its fractal dimension can be simply computed by analytically defining the quantity $N(r)$ present in Definition 3.1.4. At every step of the construction of the Cantor Set the geometrical object is covered by 2^n balls of diameter $2r = 3^{-n}$. It follows that :

$$\begin{aligned} d_F &= -\lim_{r \to 0} \frac{\log N(r)}{\log r} = && (3.2)\\ &= -\lim_{n \to \infty} \frac{\log(2^n)}{\log(3^{-n}2^{-1})} = && (3.3)\\ &= -\lim_{n \to \infty} \frac{n \log(2)}{-n \log 3 - \log 2} = && (3.4)\\ &= \frac{\log 2}{\log 3} \simeq 0.63 && (3.5) \end{aligned}$$

It is interesting to observe that a scaling law natural arises:

$$N(r) = 2^{-1}N(r/3) = 3^{-d_F}N(r/3) \tag{3.6}$$

as at each iteration the number of balls to cover the object increases with the decrease of the size.

Definition 3.1.6 *In general a differentiable function is said to be scale invariant if:*

$$f(x) = \lambda^{-\alpha} f(x/\lambda) \tag{3.7}$$

for every choice of λ and for some choice of α.

In the case of fractals, the *self-similarity* of the object arises in a relation of this type that is however valid only for a finite set of scale factors λ that is related to its fractal dimension(s).

However, it must be underlined that *self-similarity* leads to power laws of the type in Equation (3.7) but the opposite is not guaranteed. Nevertheless, it is already clear that studying the scaling properties of a system can give important insights about its profound nature [39].

3.1.2 Random Fractals and Statistical Scale Invariance

To complete this introduction, it is important to also describe the so-called *random fractals*, which are geometric objects with approximately fractal properties (i.e. self-similarity) that can be obtained as realizations of stochastic processes with peculiar features. Indeed, the probabilistic laws governing these processes are self-similar and these processes are characterized by the so-called *statistical scale invariance.*

The simplest example of a random fractal is the realization of a Brownian motion in $\mathbb{R}^d$ which can be approximated by a discrete random walk with a step of length r_0 and directions randomly drawn from a rotationally invariant distribution with zero average. Let $p(\vec{x})$ be the distribution of the individual displacement $\vec{x}$, the net displacement of the walk after m steps is $\vec{X}_m = \sum_{i=1}^{m} \vec{x}_i$. Having i.i.d. steps and exploiting the linearity of the expectation, the first and second moment of the net displacement are simply:

$$\left\langle \vec{X}_m \right\rangle_m = 0 \tag{3.8}$$

and

$$\left\langle |\vec{X}_m|^2 \right\rangle_m = \left\langle \sum_{i=1}^{m}\sum_{j=1}^{m} \vec{x}_i \cdot \vec{x}_j \right\rangle_m = \tag{3.9}$$

$$= \sum_{i=1}^{m} \langle |\vec{x}_i|^2 \rangle = m r_0^2 \tag{3.10}$$

where the subscript m represent the average computed with respect to the distribution of the the $m-$step displacement.

If one wants to compute the fractal dimension d_F of the realizations of this process, it is necessary to compute the minimum number $N(r)$ of 1-balls of length r needed to cover these paths. Choosing $r = r_0$ one obtains:

$$N(r_0) = nN(\sqrt{n}r_0) \tag{3.11}$$

as the first and second moment of the net displacement after m steps are exactly the same for both a random walk of m steps of length r_0 and a random walk of m/n steps of length $\sqrt{n}r_0$. Assuming $N(r_0) \sim r_0^{-d_F}$ one gets $d_F = 2$ and, knowing that the topological dimension of a random walk path embedded in $\mathbb{R}^d$ is $d_T = 1$, it follows that the realizations of the process are *random fractals* for $d \geq 2$. It means that the generating process itself is self-similar and this also implies the so-called statistical scale invariance of the underlying statistical laws.

Supposing for instance that $p(\vec{x})$ is a Gaussian distribution:

$$p(\vec{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{|\vec{x}|^2}{2\sigma^2}} \tag{3.12}$$

the probability distribution of the displacement $\vec{X}_n = \sum_{i=1}^{n} \vec{x}_i$ after n steps is:

$$p_n(\vec{X}_n) = \int \prod_{i=1}^{n} d\vec{x}_i p(\vec{x}_i) \delta^{(d)}(\vec{X}_n - \sum_{i=1}^{n} \vec{x}_i) = \tag{3.13}$$

$$\ldots \tag{3.14}$$

$$= \frac{1}{(2\pi n\sigma^2)^{d/2}} e^{-\frac{|\vec{X}_n|^2}{2n\sigma^2}} \tag{3.15}$$

which is again a Gaussian with a standard deviation $\sqrt{n}$ times larger than the one of the single step displacement. Performing a rescaling of $\vec{X}_n$ by the quantity $\sqrt{n}$, in order to recover the original single step scale, one gets $\vec{x}' = \vec{X}_n/(n)^{d/2}$ which , due to probability conservation under change of variables, implies:

$$p(\vec{x}')d\vec{x}' = p_n(\vec{X}_n)d\vec{X}_n \quad \Longrightarrow \quad p(\vec{x}') = p(\vec{x}) \tag{3.16}$$

It means that the distribution of the new rescaled random variable $\vec{x}'$ obtained from the n steps displacement $\vec{X}_n$ is the same as the individual step one. It follows that the process is *statistically scale invariant*.

To conclude, if one assumes a scaling law for the root mean square displacement of the process after n steps:

$$\bar{R}(n, \sigma) = \left(\left\langle |\vec{X}_n|^2 \right\rangle\right)^{1/2} \sim n^{\nu} \tag{3.17}$$

a power law with exponent $\nu = 1/2$ is obtained.

Scaling Properties of a Simple Random Walk

If one considers a simple random walk but expressed by the difference equation:

$$x(t+1) = x(t) + f(t) \tag{3.18}$$

with $f(t) \sim N(0, \sigma^2)$, it is simple to extract the scaling properties of the process from the moments of the distribution of the increments. Defining $y_\tau(t) = x(t+\tau) - x(t)$, due to the independence of the $f(t)$ variables it is simply distributed as:

$$y_\tau(t) \sim N(0, \tau\sigma^2) \tag{3.19}$$

which obviously does not depend on t. It follows that:

$$\mathbb{E}(|y_\tau(t)|^q) = \frac{(\sigma^2\tau)^{\frac{1+q}{2}}}{2^{\frac{q}{2}}\sqrt{\pi\sigma^2\tau}}\Gamma\left(\frac{1+q}{2}\right) \sim \tau^{\frac{q}{2}} \tag{3.20}$$

which reveals how the process of the increments scales with τ for a simple random walk. In the next sections it will be shown the importance of this scaling relation.

3.2 Self-affine Processes and the Hurst Exponent

In order to generalize the previously presented concepts it is important to define *self-affine* processes, that are a wider class of processes which also include the *self-similar* one. Moreover, one of the most known measure of long-term memory of time series, called the Hurst exponent, is introduced.

3.2.1 Self-affine Processes

Definition 3.2.1 *Granted $X(0) = 0$, a random process $\{X(t)\}$ that satisfies:*

$$\{X(ct_1), ..., X(ct_k)\} \overset{d}{=} \{c^H X(t_1), ..., c^H X(t_k)\} \tag{3.21}$$

for some $H > 0$ and all $c > 0$, is called self-affine.

H is called the self-affinity index or scaling exponent of $X(t)$. To be more clear one can say that self-affinity allows for different rescalings along the directions of an orthonormal basis, while self-similarity requires the same rescalings along each direction. It follows that self-similar objects are invariant under both dilations and rotations [22, 34]. Typical examples of self-affine processes used in finance are the L-stable processes and the Fractional Brownian Motions (FBM), $B_H(t)$. While the first are characterized by stable and independent increments, the FBM has dependent increments with negative autocorrelation for $0 < H < 1/2$ and positive for $1/2 < H < 1$. When $H = 1/2$ the FBM becomes a simple Brownian Motion [34, 22].

3.2.2 Hurst Exponent

The Hurst exponent is a statistical measure of the long-term memory of time series and can be obtained from the asymptotic behavior of the autocorrelation function [34, 42]. For instance, if one considers a stationary standard Gaussian function $X(t)$ with $\mathbb{E}[X(t)] = 0$ and $\mathbb{E}[X^2(t)] = 1$ the autocorrelation function:

$$C(\Delta t) = \mathbb{E}[X(t)X(t + \Delta t)] \tag{3.22}$$

can be useful to measure the roughness of the profile $X(t)$ in the Euclidean plane [34]. Indeed, if the correlation function behaves as

$$C(\Delta t) \sim 1 - |\Delta t|^{\alpha}, \quad \text{for} \quad \Delta t \to 0 \tag{3.23}$$

for $\alpha \in (0{,}2]$ one can extract the fractal dimension of the random path of $X(t)$ with the relation [22, 34]:

$$d_F = 2 - \frac{\alpha}{2} \tag{3.24}$$

On the other hand, analyzing the asymptotic behaviour at large time lags of $C(\Delta t)$, one can quantify the long-range dependence of the process:

$$C(\Delta t) \sim |\Delta t|^{-\beta}, \quad \text{for} \quad \Delta t \to +\infty \tag{3.25}$$

for $\beta \in (0{,}1)$. The Hurst exponent is defined from this quantity by the equation [34]:

$$H = 1 - \frac{\beta}{2} \tag{3.26}$$

and gives information about the long-memory dependence of the process. It is important to underline that the fractal dimension is a local property while the long-memory dependence is a global one.

For self-similar processes (in a n-dimensional space), the local properties are preserved also at larger scales and it follows that:

$$d_F = n + 1 - H \tag{3.27}$$

In general processes characterized by $H \in (0.5{,}1)$ are characterized by long-memory dependence or persistence, while when $H \in (0{,}0.5)$ the process is said to be anti-persistent [34].

In general, as it will be shown, the Hurst exponent can be measured on time series regardless the statistical properties of the underlying process generating them, nevertheless it must be always checked that the assumed scaling relations are empirically valid [34].

Re-scaled Range Statistical Analysis

Between the Fifties and the Sixties Hurst and his colleagues introduced a statistical analysis technique to describe the long-term dependence of water levels in various kind of channels and reservoirs [43, 44]. The methodology is called the re-scaled range statistical analysis and is a practical tool to extract the Hurst exponent of a given time series.

Given a time series $X(t)$ defined for $t = n{,}2n{,}3n, ...$, the average and the standard deviation over a time window are:

$$\langle X \rangle_T = \frac{n}{T} \sum_{k=1}^{T/n} X(kn) \tag{3.28}$$

$$S(T) = \left(\frac{n}{T} \sum_{k=1}^{T/n} [X(kn) - \langle X \rangle_T]^2 \right)^{1/2} \tag{3.29}$$

The range R of the time series is defined as the difference between the maximum and the minimum values of $X(t)$ in the interval $[n, T]$ [34]:

$$R(T) = \max_{n \leq t \leq T} [X(t)] - \min_{n \leq t \leq T} [X(t)] \tag{3.30}$$

The Hurst exponent can be defined from the scaling relation of the following quantity [34]:

$$\frac{R(T)}{S(T)} \sim \left(\frac{T}{n} \right)^{H} \tag{3.31}$$

This measure can therefore detect the long-range dependence of a signal and, given the fact that for an independent random process $H = 0.5$ [34, 19] (see also Equation (3.20) in the perspective of multiscaling spectrum of self-affine processes as $H(q)$ and H are related [34]), when the measured $H \neq 0.5$ the underlying process present non-trivial correlation properties. Obviously this is just the simplest version of the technique and several extensions and upgrade have been proposed in the years [34].

3.3 Multifractality and Multiscaling Properties

Multifractality was initially observed in the context of turbulence in fluid mechanics [45, 46]) and was then theoretically defined for random measures and consequently for random processes [22, 47]. In general, systems which exhibit multifractality are characterized by scaling laws defined by a spectrum of exponents and not by a single fractal dimension.

3.3.1 Self-similar Random Measures

A random measure μ defined on an interval $X \in \mathbb{R}$ is a mapping defined on a probability space and valued in the class of all measures on X. It is a transition kernel that assigns random variable to each subset of X.

Definition 3.3.1 *A random measure μ satisfying these properties:*

- *for any affine transformation S on the real line, for any interval $I_1 \subseteq I_2$*

$$\frac{\mu(SI_1)}{\mu(SI_2)} \text{ and } \frac{\mu(I_1)}{\mu(I_2)} \tag{3.32}$$

are identically distributed whenever $I_1, I_2, SI_1, SI_2 \subseteq X$;

- *for all non-decreasing sequence of compact intervals $I_1 \subseteq ... \subseteq I_n$ contained in X, the random variables*

$$\frac{\mu(I_1)}{\mu(I_2)}, ..., \frac{\mu(I_{n-1})}{\mu(I_n)} \tag{3.33}$$

are statistically independent;

is self-similar [22].

If the interval X is of the form $[0, T]$, $0 < T \leq \infty$, the first property implies the existence of a positive random process $M(c)$ independent of μ such that [22]:

$$\mu[0, ct] \overset{d}{=} M(c)\mu[0, t] \text{ for } 0 < t \leq T,\ 0 < c \leq 1 \tag{3.34}$$

Given two constants $a, b \leq 1$, thanks to the second property it can be written that:

$$\frac{\mu[0, abt]}{\mu[0, t]} = \frac{\mu[0, abt]}{\mu[0, at]}\frac{\mu[0, at]}{\mu[0, t]} \tag{3.35}$$

that implies the following property for process M [22]:

$$M(ab) \overset{d}{=} M_1(a)M_2(b) \tag{3.36}$$

where M_1 and M_2 are two independent copies of M.

To complete, assuming, without loss of generality, $X = [0,1]$, and considering $\mu(a) \equiv \mu[0, a]$, Relation (3.36) implies [22]:

$$\mathbb{E}[M(a)^q] = a^{\tau(q)+1} \tag{3.37}$$

and therefore:

$$\mathbb{E}[\mu(a)^q] = \mathbb{E}[\mu(1)^q]a^{\tau(q)+1} \tag{3.38}$$

which is the scaling-relation which characterize multifractals. It follows that the multifractality is a direct consequence of the statistical self-similarity of the random measure μ [22].

The function $\tau(q)$ is the so-called scaling function and satisfies the following properties [22]: $\tau(0) = -1$, $\tau(1) = 0$, concavity due to Hölder's inequality [22].

3.3.2 Multifractal Processes

From random measures the multifractality can simply be extended to stochastic processes.

Definition 3.3.2 *A stochastic process $\{X(t)\}$ is called multifractal if it is stationary and satisfies:*

$$\mathbb{E}(|X(t)|^q) = c(q)t^{\tau(q)+1},\ \textit{for all } t \in \mathcal{T}, q \in \mathcal{Q} \tag{3.39}$$

where $\mathcal{T}$ and $\mathcal{Q}$ are intervals on the real line with positive length such that $0 \in \mathcal{T}, [0,1] \in \mathcal{Q}$, and $\tau(q)$ and $c(q)$ are functions with domain $\mathcal{Q}$ [22].

The function $\tau(q)$ is the so-called scaling function already defined in the previous subsection. The concavity implies that the multifractal scaling relation can only hold for bounded time intervals and therefore multifractal processes must contain a *crossover*.

It is interesting to show that self-affine processes are multifractal. Given a self-affine process $\{X(t), t \geq 0\}$, with self-affinity index H it is known from definition (3.21) that $X(t) \overset{d}{=} t^H X(1)$ and thus:

$$\mathbb{E}[|X(t)|^q] = t^{Hq}\mathbb{E}[|X(1)|^q] \tag{3.40}$$

The scaling defined in relation (3.39) holds and in particular:

$$\tau(q) = Hq - 1 \text{ and } c(q) = \mathbb{E}[|X(1)|^q] \tag{3.41}$$

In this case the scaling function is linear in q and the scaling behaviour only needs H to be characterized [22]. From this observation comes the subsequent classification [34, 22]:

- multifractal processes with linear $\tau(q)$ are called *uniscaling* or *unifractal*;
- multifractal processes with non-linear $\tau(q)$ are called *multiscaling* or *multifractal*;

Usually, when one retrieves the scaling (3.39) property empirically it is claimed that the data are multiscaling [34, 18].

3.4 Multifractal Models in Finance

When studying the scaling with time of the process describing the price variation of a financial asset, the multiscaling properties defined in the preceding sections were widely observed in different markets[48, 49, 34]. Many researchers thus decided to devise models which could explain these properties with the help of the theory of multifractal processes [22].

In the following two of the main models within this framework are presented: the Multifractal Model of Asset Returns by Mandelbrot et al. [22] and the Multifractal Random Walk by Bacry et al. [12, 50]. In particular the second one is usually used as a benchmark to test the effectiveness of multiscaling estimation methods as it provides an analytical form of the multiscaling spectrum [18].

3.4.1 The Multifractal Model of Asset Returns

In the Multifractal Model of Asset Returns the price of a financial asset is viewed as multiscaling process with heavy tails and long memory, and the fluctuations of the volatility are introduced describing the trading time as generated by the cumulative density function of a random multifractal measure [22].

Given the price of a financial asset $\{P(t); 0 \leq t \leq T\}$, the process $X(t)$ is defined as:

$$X(t) = \ln P(t) - \ln P(0) \tag{3.42}$$

Assuming the following properties:

- $X(t)$ is defined as a compound process:

  $$X(t) \equiv B_H[\theta(t)] \tag{3.43}$$

  with $B_H(t)$ fractional Brownian Motion with self-affinity index H, $\theta(t)$ stochastic trading time (increasing function of t);
- the trading time $\theta(t)$ is the c.d.f. of a multifractal measure defined on $[0, T]$, it follows that it is a multifractal process with continuous non-decreasing paths, and stationary increments;

- $\{B_H(t)\}$ and $\{\theta(t)\}$ are independent.

With this assumptions the process $X(t)$ is multifractal, with stationary increments and scaling function $\tau_X(q) \equiv \tau_\theta(Hq)$ [22].

The multifractal nature of the process imposes a multiscaling relation (3.39) and depending on the measure μ that characterizes the trading time and on the H index of the fractional Brownian motion, the process can assume very different properties. For instance, when $H \geq 1/2$ and $\mathbb{E}(\theta^{Hq})$ is finite, the price process has long memory in the value of its increments, and in general it is possible to obtain long tails in the increments distribution, long-dependence in the absolute value of price increments (i.e. volatility clustering) and absence of correlation in simple price increments [22].

3.4.2 Modelling Financial Time Series Using Multifractal Random Walks

A multifractal random walk (MRW) process $X(t)$ is the limit process ($\Delta t \to 0$) of a standard random walk $X_{\Delta t}(t)$ with a stochastic variance:

$$X(t) = \lim_{\Delta t \to 0} X_{\Delta t}(t) = \lim_{\Delta t \to 0} \sum_{k=1}^{t/\Delta t} \epsilon_{\Delta t}[k] e^{\omega_{\Delta t}[k]} \tag{3.44}$$

in which $e^{\omega_{\Delta t}[k]}$ is the stochastic volatility, and $\epsilon_{\Delta t}$ is a Gaussian white noise of variance $\sigma^2 \Delta t$ independent of $\omega_{\Delta t}$. In order to obtain an exact multiscaling spectrum for time scales smaller than the integral scale L, the process $\omega_{\Delta t}$ must be a stationary Gaussian process such that:

- $\mathbb{E}(\omega_{\Delta t}[k]) = -Var(\omega_{\Delta t}[k])$;
- $Cov(\omega_{\Delta t}[k], \omega_{\Delta t}[l]) = \lambda^2 \ln \rho_{\Delta t}[|k-l|]$ where

$$\rho_{\Delta t}[k] = \begin{cases} \frac{L}{(|k|+1)\Delta t} & \text{for } |k| \leq L/\Delta t - 1 \\ 1 & \text{otherwise} \end{cases}$$

  From the last relation it follows that the volatility is correlated up to a time lag L.

From this properties the following multiscaling spectrum follows:

$$\mathbb{E}(|X(t+l) - X(t)|^q) = K_q l^{\zeta_q} \tag{3.45}$$

$$\zeta_q = (q - q(q-2)\lambda^2)/2 \tag{3.46}$$

The parameter λ^2 is called the intermittency factor and controls the non linearity of ζ_q, when $\lambda^2 = 0$ the process $X(t)$ is a Brownian motion with linear scaling function [12, 50].

This process can thus be used to describe the return fluctuations of financial time series: the price of an asset $P(t)$ is modelled by $e^{X(t)}$ where $X(t)$ is a MRW [12]. From data one can extract the parameters of the model in the following way:

- the variance σ^2 can be extracted using the relation $Var(X(t)) = \sigma^2 t$;
- L and λ can be obtained from the approximated form of the magnitude correlation of the process:

$$C_\omega(\tau, l_1, l_2) = Cov(\omega(t, l_1), \omega(t+\tau, l_2)) \simeq -\lambda^2 \ln(\frac{\tau}{L}), \qquad L > \tau \gg max(l_1, l_2)$$

What makes this model very interesting is that it reproduces the main observed empirical stylized facts : absence of correlation between price variations, long-range volatility correlations, linear and non-linear correlation between assets. Furthermore, the multiscaling spectrum is exactly known and just three parameters are needed to control the properties of time series generated with this model: σ^2 controls the variance of the fluctuations, λ^2 controls the scale invariance properties and the volatility correlation, L controls the volatility decorrelation scale [12].

3.5 Methodologies to Measure Scaling Properties in Time Series

As showed in many works [49, 34, 19, 18, 35, 51] multiscaling is a widely accepted stylized fact in financial time series and many different tools to measure this properties have been tested and devised. In this work it is placed focus on the Generalized Hurst Exponent method [34, 35] and on a variation of this technique [19] that is more well-founded from the point of view of statistical significance.

3.5.1 Generalized Hurst Exponent Method

This method is a generalization of the previously shown R/S method to measure the standard Hurst exponent, and employs the q-th order moments of the distribution of the increments of a time series (i.e. returns for financial time series) to extract its scaling behaviour [34]. Moreover, with this tool one can study the functional behavior of the scaling exponents and detect if the observed time series presents multiscaling properties.

Given $X(t)$ a process with stationary increments, the Generalized Hurst Exponent method (GHE) considers the following function:

$$\Xi(\tau, q) = \mathbb{E}[|X(t+\tau) - X(t)|^q] \sim K_q \tau^{qHq} \tag{3.47}$$

where $q = \{q_1, ..., q_{max}\}$ is the set of moments considered and $\tau = \{\tau_1, ..., \tau_{max}\}$ is the set of time scales over which returns are computed. K_q is the q-moment for $\tau = 1$, and H_q is the so-called generalized Hurst exponent which depends on q [18]. It is clear that the method assumes that the process under study satisfies the multiscaling relation (3.39) and therefore it is expected that qH_q is a concave function due to Hölder's inequality [22]. For a standard Brownian motion, which is an uniscaling process, $H_q = H = 0.5$ regardless of q, while processes with $H > 0.5$ ($H < 0.5$) are said to be persistent (anti-persistent) [34, 18]. In order to define a multiscaling proxy the non-linearity of the function qH_q is studied as it is has been shown that only uniscaling processes have a scaling exponents that is linear in q

[22, 19]. This property can therefore be analyzed performing a linear regression of Equation (3.47) in log-log scale:

$$\ln(\Xi(\tau, q)) = qH_q ln(\tau) + \ln(K_q) \tag{3.48}$$

Certainly the linearity of the left hand side with respect to $\ln(\tau)$ must be checked, and if it holds the method computes different slopes of the straight lines at different q [35]. The so-called multiscaling proxy can be devised by fitting the measured scaling exponent with a second degree polynomial [18, 35]:

$$qH_q = Aq + Bq^2 \tag{3.49}$$

It follows that if the empirical $\hat{B} = 0$ the process is uniscaling, while if $\hat{B} \neq 0$ the process is multiscaling [35, 19, 18].

Usually, when studying financial time series

$$|X(t+\tau) - X(t)| = |\ln P(t+\tau) - \ln P(t+\tau)| = |r_\tau(t)|$$

is the log-return process and in this dissertation τ is always expressed in days when not specified. It is obvious that the choice of τ_{min}, τ_{max} and q_{min}, q_{max} are fundamental to obtain reliable results which give insights about the real scaling behavior of the underlying process.

The choice of q

In order to choose the correct range of moments to perform the previously described estimates, the prescriptions present in [18] can be followed. To have a robust measure of multiscaling, it is necessary to have $q < \alpha$ where α is the tail exponent of the return distribution, otherwise the behaviour found by not considering this fact is severely biased [8, 11, 18]. Since it has been empirically shown that financial time series have fat tails with tail exponents ranging from ~ 1.5 to ~ 4 [11, 8], in this project a conservative approach by selecting $0 < q \leq 1$ is chosen.

The choice of τ

Regarding the time aggregation instead one can initially choose to employ a heuristic approach selecting for daily data $\tau \in [1,19]$ days, which has been proved to be good enough to highlight the multiscaling behaviour in financial time series [34]. Nevertheless, this range has exhibited some biases caused by autocorrelations and power laws [35] and therefore in the next section a well founded way to extract τ_{max} will be shown.

3.5.2 Dynamical Weighted Generalized Hurst Exponent Method

Starting from the scaling assumption in Equation (3.47) one can perform the same regression in Equation (3.49) in rolling time windows of length Δt in order to measure how H_q evolves in time. In a few recent articles [17, 52], a modification of this dynamical GHE method, the weighted GHE (wGHE), has been defined with the objective to give more importance to recent events in the computation of the time average performed in Equation (3.47). In particular, when summing within a time interval $[t-\Delta t, t]$ of length Δt, each term of the time series is weighted in such a way that more recent terms give higher contributions to the sum used to compute the moments. The average present in Equation (3.47) is replaced by the following expression:

$$\mathbb{E}\left[f\left(r_\tau(t)\right)\right]_\theta = \sum_{s=0}^{\Delta t-1} w_s\left(\theta\right) f\left(r_\tau(t-s)\right), \tag{3.50}$$

where f is generic function of the returns $r_\tau(t)$. The weighting factor w_s can be defined as an exponentially decaying function of time:

$$\begin{cases} w_s\left(\theta\right) = w_o\left(\theta\right) e^{-\frac{s}{\theta}} \\ w_o = w_o\left(\theta\right) = \frac{1-e^{-\frac{1}{\theta}}}{1-e^{-\frac{\Delta t}{\theta}}} \end{cases} \quad \text{and} \quad \sum_{s=0}^{\Delta t-1} = \omega_s(\theta) = 1 \tag{3.51}$$

and θ is the characteristic time. It follows that Equation (3.47) becomes

$$\Xi(\tau, q, \theta) = \mathbb{E}\left[|X(t+\tau) - X(t)|^q\right]_\theta \sim K_q \tau^{qH_q^{(\theta)}}, \tag{3.52}$$

where $H_q^{(\theta)}$ is the wGHE with characteristic time θ. It has been shown that $\Delta t = \theta$ is a reasonable choice as corresponds to a time window for which the last day in the past is weighted by a factor $1/e$ less than the most recent day [17]. Regarding the time aggregation $\tau \in$ [1,19] days is the most common choice [17, 52] as it allows to observe the scaling behaviour without the need to employ time windows of excessive length. Obviously considering the scaling (3.52) in a given time window that ends in t one extracts $H_q^{(\theta)}(t)$.

3.5.3 Relative Normalized and Standardized Generalized Hurst Exponent Method

Recently a novel methodology, called Relative Normalized and Standardized Generalized Hurst Exponent (RNSGHE), has been developed by Brandi and Di Matteo to quantify and test the multiscaling properties of financial time series in a statistical significant way [18]. It has already been shown that given a process $\{p(t)\}$ with stationary increments, the Generalized Hurst Exponent method considers a function of the increments defined in this way [34]:

$$\Xi(\tau, q) = \mathbb{E}[|r_\tau(t)|^q] \sim K_q \tau^{qH_q} \tag{3.53}$$

where $q = \{q_1, ..., q_{max}\}$ is the set of moments considered and $\tau = \{\tau_1, ..., \tau_{max}\}$ is the set of time scales over which returns are computed. K_q is the q-moment for $\tau = 1$, and H_q is the so-called generalized Hurst exponent which depends on q [18]. To detect multiscaling, it has been indicated that one needs to analyze the non-linearity of the scaling function qH_q present in equation (3.53). Rather than estimating it in the regression, in [18] it has been proposed to compute the value of K_q by evaluating $\Xi(1, q)$, in order to reduce the presence of possible biases introduced in the estimation. It is therefore sufficient to normalize the function $\Xi(\tau, q)$ as

$$\tilde{\Xi}(\tau, q) = \frac{\Xi(\tau, q)}{K_q}, \tag{3.54}$$

to eliminate the possible bias introduced by the estimation of K_q with the regression. If one then defines the q-order normalized moment as

$$\dddot{\Xi}(\tau, q) = \tilde{\Xi}(\tau, q)^{\frac{1}{q}} \tag{3.55}$$

the scaling relation defined in Equation (3.53) becomes:

$$\dddot{\Xi}(\tau, q) \sim \tau^{H_q}. \tag{3.56}$$

In this way, the multiscaling can be detected by fitting the measured scaling exponent with a first degree polynomial:

$$H_q = A + Bq. \tag{3.57}$$

where A is the linear scaling index an B is the multiscaling proxy (different from zero for multiscaling processes) [18, 19].

To conclude the relative structure function between two consecutive moments, namely q_i and q_j ($q_j > q_i$), can be defined as follows

$$\dddot{\Xi}(\tau, q_i, q_j) = \frac{\dddot{\Xi}(\tau, q_j)}{\dddot{\Xi}(\tau, q_i)} \sim \frac{\tau^{H_{q_j}}}{\tau^{H_{q_i}}} = \tau^{H_{q_j} - H_{q_i}} = \tau^{H(q_i, q_j)}, \tag{3.58}$$

where $H(q_i, q_j) = H_{q_j} - H_{q_i}$. Equation (3.58) can also be rewritten in the following way:

$$\begin{bmatrix} \dddot{\Xi}(\tau, 0, q_1) \\ \dddot{\Xi}(\tau, q_1, q_2) \\ \vdots \\ \dddot{\Xi}(\tau, q_{M-1}, q_M) \end{bmatrix} = \tau^{\left[H_{(0,q_1)}, H_{(q_1,q_2)}, \cdots, H_{(q_{M-1},q_M)}\right]}, \tag{3.59}$$

where M is the maximum number of moments used. This structure is useful to verify if a process is statistically multiscaling using a t-test on each estimated $H(q_i, q_j)$. Indeed, the fact that for uniscaling time series $H_q = H$ implies that the difference between different order moments is always 0, except for $H(0, q_1)$. On the other hand, for multiscaling time series this quantity should be different from 0 for all q. Using Equation (3.59) it is also possible to use an F-test to test if all the coefficients except for the first one ($H(0, q_1)$) are jointly equal to 0 against the alternative that some coefficients are different from 0. This is a weaker multiscaling test compared to the multiple t-tests. The classification defined in [18] is therefore the following: *strongly multiscaling processes* are those processes which reject both the null hypothesis for all the t-tests and the null of the F-test; *weakly multiscaling processes* are those processes for which the null hypothesis of all the t-tests is rejected but not the null of the F-test. These definitions come from the idea that if a process is multiscaling, all the $H(q_i, q_j)$ are significantly different from 0. However, if the process reconstructed with a single exponent is statistically equivalent to the one reconstructed with the full multiscaling spectrum, this means that such multiscaling behavior is weak. Finally if the null hypothesis for one or more t-tests is not rejected but the F-test rejects the null hypothesis, the process is a *non-stable multiscaling process* [18]. In this work Equation (3.59) in log-log scale will be used because it provides better results as already analyzed in [18].

The choice of q

As already explained for the GHE method the best conservative choice is to select $q \leq 1$. Usually the interval employed is $q \in (0.02,1)$ in order to have a sufficiently wide range of moment to test for the multiscaling property [19].

The choice of τ_{max}

Other than q, selecting correctly the maximum aggregation time $\tau_N = \tau_{max}$ is crucial to estimate correctly the multiscaling properties. In financial time series in fact, there exists a cutoff over which the data are uncorrelated and in order to estimate correctly the scaling exponents it is very important to avoid mixing up the correlated state with the uncorrelated one [18]. The novel method employed to extract τ_{max} is called the Autocorrelation Segmented Regression (ACSR) described in [18]. The idea of this approach is simple: first define the autocorrelation ρ of the absolute (daily in this case) return series at lag Δt as:

$$\rho(|r_1(t+\Delta t)|, |r_1(t)|) = \frac{\mathbb{E}[(|r_1(t+\Delta t)| - \mu_1)(|r_1(t)| - \mu_1)]}{\sigma_1^2} \tag{3.60}$$

where μ_1 is the average value of the absolute returns time series and σ_1^2 is its variance. Then perform a segmented regression on Equation (3.60) and define $\tau_{max} = \Delta t^*$ as the break point between the correlated state and the random state which minimizes the sum of squared residuals. The autocorrelation function of the absolute returns is assumed to have the following form:

$$\rho(|r_1(t+\Delta t)|, |r_1(t)|) = \rho(\Delta t) = \begin{cases} \alpha + \beta(\Delta t), & \text{if } \Delta t < \Delta t^* \\ \alpha + \beta(\Delta t^*), & \text{if } \Delta t \geq \Delta t^* \end{cases} \tag{3.61}$$

where α is the intercept of the regression and that can be fixed to be equal to $\rho(1)$, β is a memory exponent for the autocorrelation function, Δt is the lag at which the autocorrelation is computed, and $\widehat{\Delta t}^*$ is the estimated value of the sought breakpoint. The maximum aggregation time is therefore $\widehat{\Delta t}^* = \tau_{max}$ [18]. This method has shown very good results to detect the L parameter of time series simulated with the Bacry et al. MRW model [12, 18].

Step by step procedure

The described RNSGHE procedure, combined with the ACSR method, can therefore be summarized in this way:

1. Compute τ_{max} with the Autocorrelation Segmented Regression method, computing the autocorrelation function for time lags $\Delta t \leq \frac{1}{5}$ of the length of the returns time series in order not to bias the scaling estimation with too few values [18, 19];
2. Fix the vector of q values to be used in the estimation;
3. Perform the linear regression in log-log scale of Equation (3.59) given the fact that the linear fit has a smaller RMSE with respect to the non linear one according to [18];
4. Compute the multiscaling curvature using Equation (3.57) and test for statistical significance.

The test for statistical significance of the results requires the subsequent steps:

- Test if each scaling increment $H(q_i, q_j)$ is statistically significant through a t-test, if the null hypothesis is rejected for a coefficient it can be said that it is robustly different from zero;
- Test if all scaling increments except $H(0, q_1)$ are jointly different from zero through a F-test, if the null hypothesis is rejected $H(0, q_1)$ alone is not enough to describe the scaling behaviour;
- Perform the full regression of Equation (3.57) and test for $\widehat{A} = 0.5$ and $\widehat{B} = 0$[1] using a t-test. If this test gives a conflicting result with respect to the other two steps, a deeper analysis is required.

From this procedure it is obtained the previously described classification that can be summarized in the following table:

t-tests	**F-test**	**MS classification**
pass	pass	strongly
pass	fail	weakly
fail	pass	non-stable
fail	fail	no multiscaling

Table 3.1: Statistical tests on the scaling exponents regressions.

With "pass" it is meant that the test has rejected the null hypothesis and thus the p-value is under 0.05. This classification, as already mentioned, is however subjected to the confirmatory test on the multiscaling proxy significance that is the last step used to check if the obtained behaviour is truly multiscaling.

[1] These values correspond to the absence of multiscaling null hypothesis.

3.6 Multiscaling properties of financial data

As already stated, financial data have been shown to be multiscaling regardless of the markets and the periods analyzed [34, 19, 18, 35, 53, 49]. In this section the evidence of these scaling properties are displayed and measured on the data set analyzed in the previous chapter employing the Generalized Hurst Exponent method, its dynamical version and the last presented RNSGHE technique, which is the most advanced and the one recently used for research purposes [18, 19].

3.6.1 GHE Example on Data

It has been demonstrated that the Generalized Hurst Exponent [34, 35] is an effective tool to point out the scaling properties (3.47) of financial time series as well to measure it through the non-linear fit of the Hurst exponent (3.49). Usually when performing this kind of analysis one first checks graphically for the validity of the scaling assumption and then extracts in a statistical significant way the results of the regressions.

In this project the method presented in (3.5.3) will be mainly employed due to the recent evidence of its strength. However, in this subsection an example of the simpler GHE exponent technique is showcased on a single time series, just to emphasize that already with a more straightforward methodology it is possible to evidence the multiscaling properties.

The daily closing price time series of Apple Inc stock from 02-Jan-1990 to 30-Nov-2022 is selected, and the scaling properties of the log-returns are analyzed. As in [34], $\tau \in [1,19]$ days is chosen and 50 equally spaced values of $q \in [0.05,1]$ are picked.

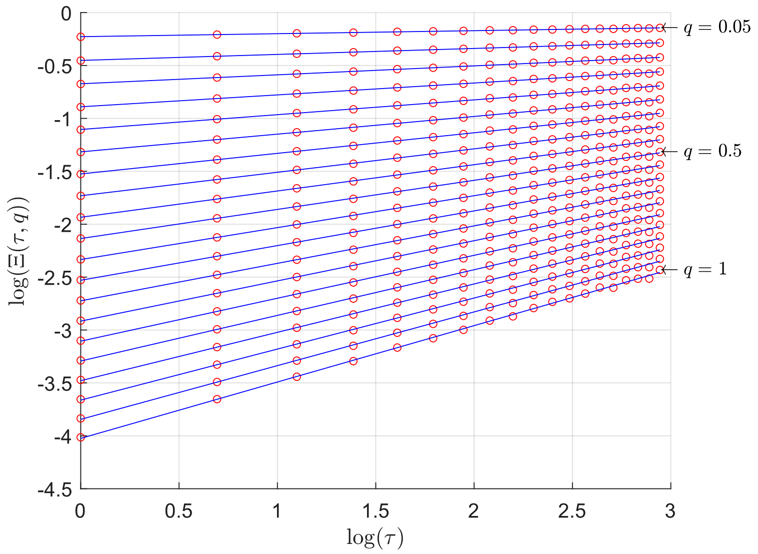

Figure 3.1: Plot of the scaling relation (3.48) for the daily closing price time series of Apple Inc stock from 02-Jan-1990 to 30-Nov-2022. Choosing $\tau \in [1,19]$ days and $q \in [0.05,1]$, the scaling is clearly observable.

It is evident that the linear relation (3.48) in log-log scale is valid and therefore one can exploit the quadratic assumption (3.49) for the Hurst exponent to extract the multiscaling proxy.

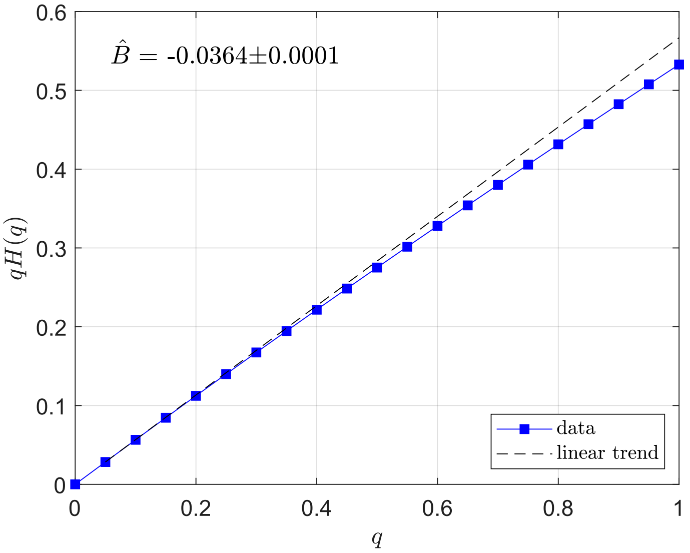

Figure 3.2: Empirical non linear relation between qH_q and q ($\tau \in$ [1,19] days and $q \in$ [0.05,1]) which highlights the multiscaling nature of the daily closing price time series of Apple Inc stock from 02-Jan-1990 to 30-Nov-2022. The multiscaling proxy $\hat{B}$ is extracted assuming (3.49) and the error reported is the Standard Error obtained from the non-linear least square technique.

From the plot it can be observed that the function qH_q bends below the linear trend and therefore the time series is said to be multiscaling. To quantify this behavior the so-called multiscaling proxy $\hat{B}$ (i.e. coefficient of the quadratic term) can be extracted through the Levenberg–Marquardt algorithm used to solve non-linear least squares problem [54]. The obtained result is significant at a 5% level and therefore the behaviour is considered valid.

3.6.2 Dynamical weighted GHE Application on Data

The same methodology can be employed in rolling time windows in order to analyze how the scaling properties of a given asset evolve in time. For instance, transitions from uniscaling to multiscaling behavior occur before critical market events, such as stock market bubbles and therefore these behaviours can be used as ’fingerprints’ of a turbulent market period as well as provide warning signals for an upcoming stock market ’bubble’ [17, 52]. Usually one can exploit a change point detection analysis [55] to devise the moments at which there are sudden changes in the Hurst exponent patterns, and this information can then be exploited both to study the changes in H_q for a given q and for the multiscaling properties of the time series.

In addition, knowing the periods in which the Hurst exponent H_q assumes values significantly different from 0.5 gives important insights about the state of the market for that particular asset: as an example, when $\hat{H}_1^\theta(t) > 0.5$ the market is in a trend-state and one can for instance buy (sell) if the price is increasing (decreasing), conversely when $\hat{H}_1^\theta(t) < 0.5$ the market is in a mean-reverting state and one can for instance buy (sell) if the price is decreasing (increasing) [18, 56]. Obviously different qs give information about the scaling at different time horizons, as smaller q weight more smaller returns that often occur at shorter time scales. In this section an application of the dynamical methodology is shown on the top 4 capitalized stocks of the data set: Apple Inc, Microsoft Corp, Alphabet Inc-Cl A, Amazon.com Inc (Alphabet Inc-Cl A has been chosen in place of Alphabet Inc-Cl C just because it has a longer time series).

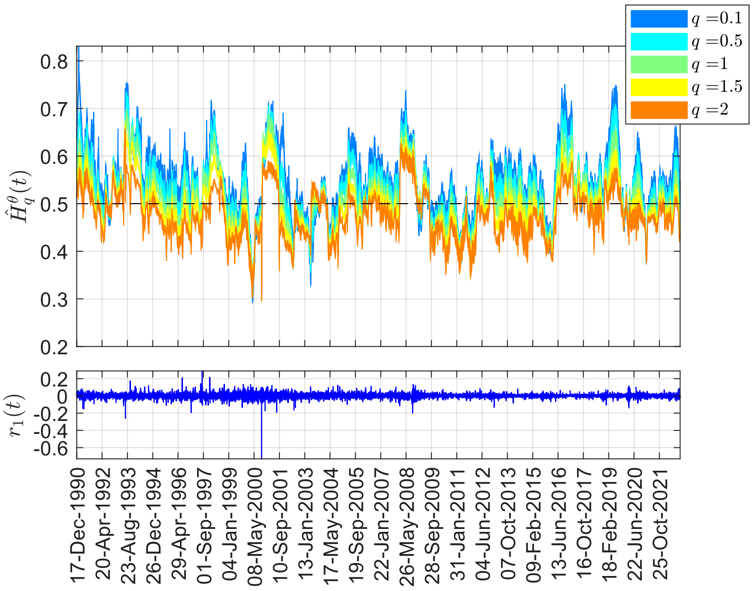

Figure 3.3: Result of $\hat{H}_q^\theta(t)$ on Apple Inc. stock daily closing price for $\Delta t = \theta = 250$ days. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit. $\tau \in [1,19]$ days, $q \in \{0.1,0.5,1,1.5,2\}$.

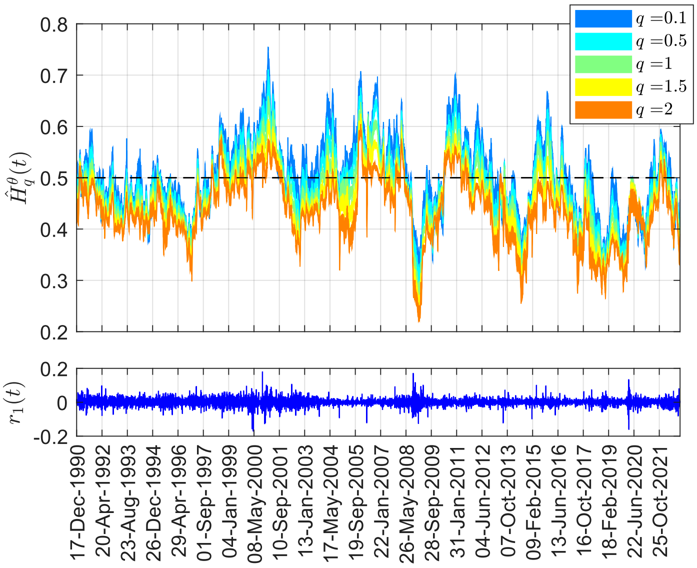

Figure 3.4: Result of $\hat{H}_q^\theta(t)$ on Microsoft Corp. stock daily closing price for $\Delta t = \theta = 250$ days. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit. $\tau \in [1{,}19]$ days, $q \in \{0.1{,}0.5{,}1{,}1.5{,}2\}$.

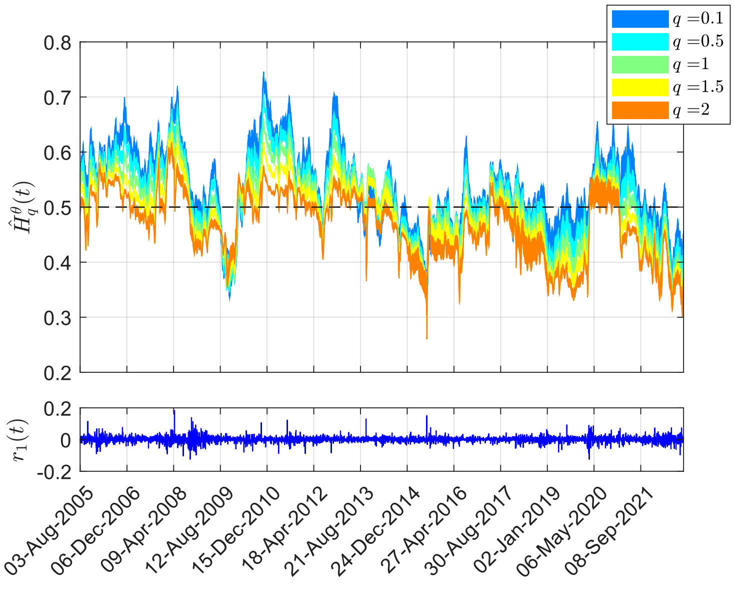

Figure 3.5: Result of $\hat{H}_q^\theta(t)$ on Alphabet Inc-Cl A stock daily closing price for $\Delta t = \theta = 250$ days. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit. $\tau \in [1{,}19]$ days, $q \in \{0.1{,}0.5{,}1{,}1.5{,}2\}$.

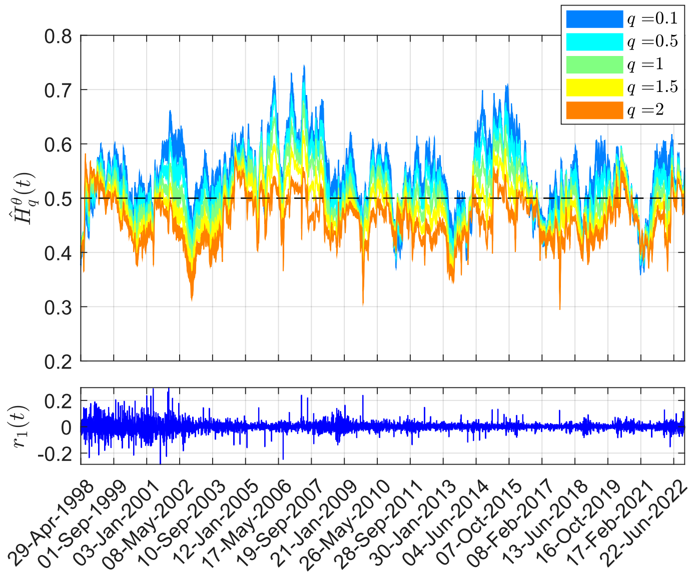

Figure 3.6: Result of $\hat{H}_q^\theta(t)$ on Amazon.com Inc stock daily closing price for $\Delta t = \theta = 250$ days. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit. $\tau \in [1{,}19]$ days, $q \in \{0.1{,}0.5{,}1{,}1.5{,}2\}$.

From these plots one can visualize the evolution of the scaling behaviour of different assets. It is confirmed that, as already observed [17, 52], sudden changes in the Generalized Hurst Exponent $\hat{H}_q^\theta(t)$ behaviour such as increasing trends from values below 0.5 to values above 0.5, occur in correspondence of periods of higher volatility which usually correspond to financial crises or market crashes. For instance, all 4 stocks show an initial decreasing trend between 2008 and 2009 for the different values of q, but the specific scaling properties of the single asset are very different. In particular, Apple Inc shows values higher than 0.5 before the decrease while Microsoft Corp values way lower than 0.5.

In general this kind of analysis can be extended as needed by including the dynamical study of the multiscaling proxy $\hat{B}$ as in [17] or by exploiting these measures as additional information for a trading strategy.

It must be underlined that in the previous plot it has been chosen to show the scaling behaviour also for values of $q \geq 1$ just to compare the different patterns. It is indeed known that financial time series are characterized by fat tailed distributions of the returns and therefore when $q \geq \alpha$ (i.e. the tail exponent) the moments can diverge causing biases in the estimates [18, 49].

Significance Test

To test the significance of the devised measures, $\hat{H}_q^\theta(t)$ is estimated also over 500 simulations of random walks of size Δt in each time window. In particular, following the idea in [17], in each time window one simulates 500 surrogate time series of length Δt generating a random walk using as the initial value the price at the beginning of the window, 0 as the average of the process and the variance of the series of the daily returns as the variance of the process:

$$\begin{cases} p(s+1) = p(s) + \mathcal{N}(0,\sigma^2) \\ \sigma^2 = Var(p(s+1) - p(s)) \end{cases} \qquad s \in [t, t+\Delta t] \tag{3.62}$$

In this way it can be tested if the values obtained on the real time series fluctuate just due to the finite size of the samples or because the underlying process is genuinely different from a purely random process [52]. For each time window the percentiles {2.5,97.5} of the distributions of $\hat{H}_q^\theta(t)$ can be computed and these quantities correspond to the bounds of the dynamical 95% confidence interval $(CI_\theta^{u/d}(t;q))$. It is clear that no useful information is obtained when the measure oscillates around the RW one (i.e. 0.5) but there are some clear trends which show quantities that truly cross the random regime and have a probability higher than the 95 % of not being originated from a simple random walk process [52].

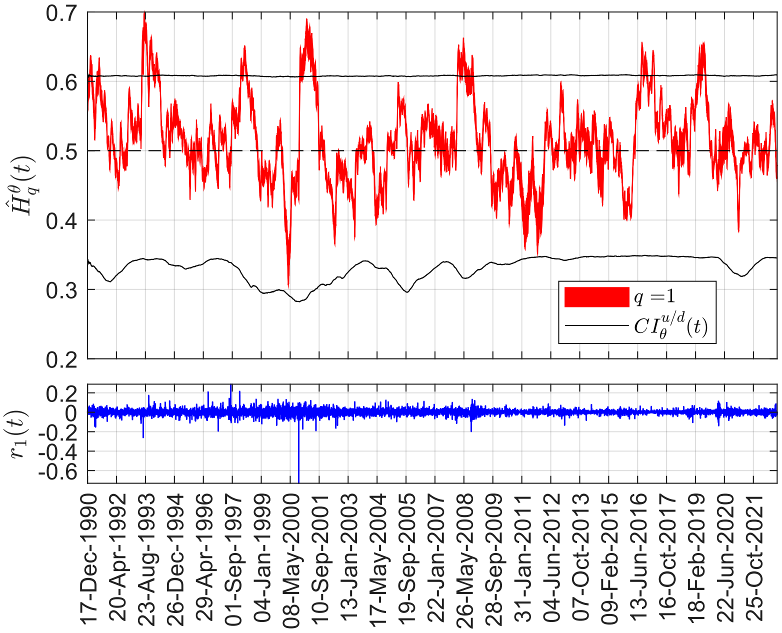

Figure 3.7: Result of $\hat{H}_1^\theta(t)$ on Apple Inc stock daily closing price for $\Delta t = 250$ days. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit. The black lines show the exponentially weighted moving average (3.50, $\Delta t = \theta = 250$ days) of the bounds of the dynamical 95 % confidence interval (for $q = 1$) computed in the previously defined way.

In this example only the $H_1^\theta(t)$ with its confidence interval is displayed to make the representation more clear.

As expected the majority of the values fall within the confidence interval due to the small size of the time windows. However, there are some clear trends in the behaviour of $H_1^\theta(t)$ which are also characterized by values of the Hurst exponent outside of the confidence interval. For instance the plot shows that between 2000 and 2001, and between 2008 and 2009, the Hurst exponent is significantly higher than 0.5 and therefore the market for this asset is in a trend-state that could be exploited to predict the price movements.

3.6.3 RNSGHE: Significant Measure of Multiscaling Properties of Data

In this last part of the chapter the Relative Normalized and Standardized Generalized Hurst Exponent Method (3.5.3) is applied to the 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990. This choice is made due to the fact that longer time series provide a scaling behaviour which is closer to the "real" one characterizing the underlying process. As explained before, $q \in [0.02,1]$ is selected in a conservative way to avoid biases in the estimate. Regarding τ_{max} instead, the autocorrelation segmented regression (3.5.3) is applied to extract the aggregation time scales over which the data are uncorrelated.

Having fixed the parameters and the methodology the following results are obtained.

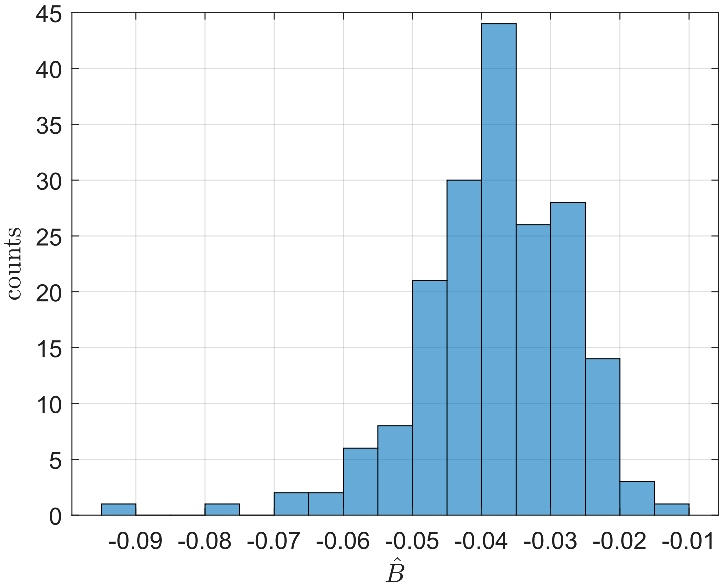

Figure 3.8: Results of the multiscaling proxy B extracted with the RNSGHE method on 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990 to 30-Nov-2022. $\tau \in [1, \tau_{max}]$ where τ_{max} is computed with the ACSR method on each stock, $q \in [0.02,1]$. The obtained values are all significant at a 5% level.

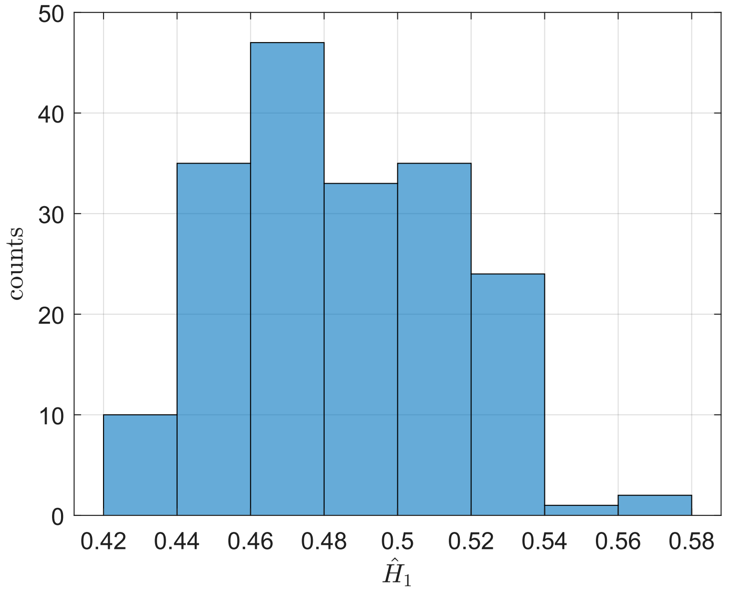

Figure 3.9: Results of H_1 extracted from the RNSGHE procedure applied on 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990 to 30-Nov-2022. $\tau \in [1, \tau_{max}]$ where τ_{max} is computed with the ACSR method on each stock, $q \in$ [0.02,1]. The obtained values are all significant at a 5% level.

As expected all the values of $\hat{B}$ are negative and the 86.1 % of the statistical test resulted in strongly multiscaling processes, while the remaining 13.9 % resulted in non-stable multiscaling processes. This outcome can be ascribed to the fact that for some tiny values of q the quantity is very small $H(q_i, q_j)$ and therefore, for a few stocks, some of the t-tests cannot reject the null hypothesis. Despite this, the multiscaling behaviour is still present, even if not in such a strong form, and therefore the multiscaling property of financial assets, as well on this data set, is confirmed.

Regarding the second plot, the generalized Hurst exponent for $q = 1$ is displayed. It is interesting to point out that almost all the values reject the null hypothesis of the Hurst exponent being 0.5 at a 1% level, except for the one which are characterized by $\hat{H}_1 \in$ (0.498,0.502). It is therefore evident that the scaling properties of these financial time series, even when considering only a single moment $q = 1$, can be different from the one of a simple Brownian motion, and taking into account this aspect, and more in general the multiscaling nature of these data, is important to correctly asses the risk for a given asset [18].

Considering also the median market capitalization of these assets, one can look for a relation between these values and the multiscaling proxy as in [53].

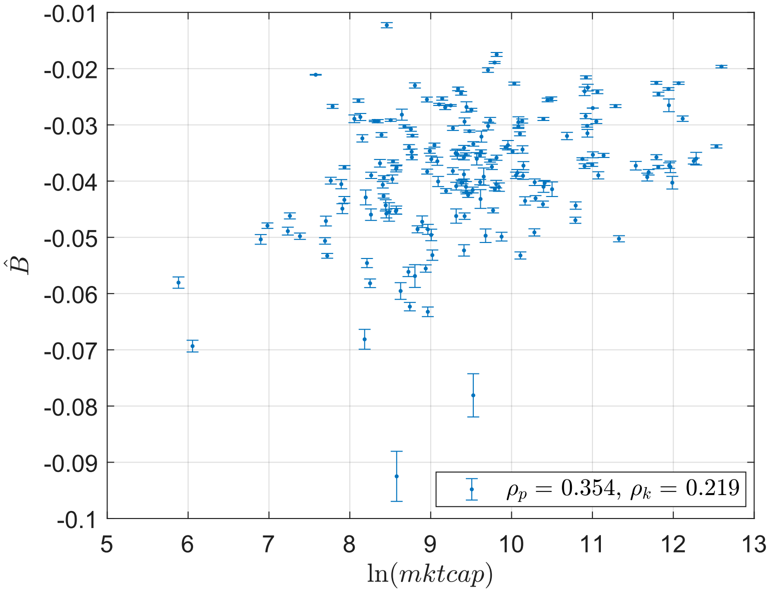

Figure 3.10: Relation between the multiscaling proxy $\hat{B}$ and the log of the median market capitalization of the 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990 to 30-Nov-2022. The Pearson and Kendall correlation measures between the two quantities are significant at a 1% level.

As already shown in the previously cited paper, there is a non trivial relation between these two quantities which highlights an interesting property: higher capitalized stocks tend to be less multiscaling. This dependence is quantified by the Pearson and Kendall correlation coefficients which are reported on the plot. This behaviour could be caused by the fact that "smaller" assets show an higher volatility correlation which is reflected on the scaling spectrum [12].

To complete, a preview of the properties that will be analyzed in the next chapter is shown. In particular considering again the 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990 to 30-Nov-2022, the average Pearson correlation coefficient $\bar{\rho}$ between them is computed over the whole period and these values are represented against the multiscaling proxy of the same time series.

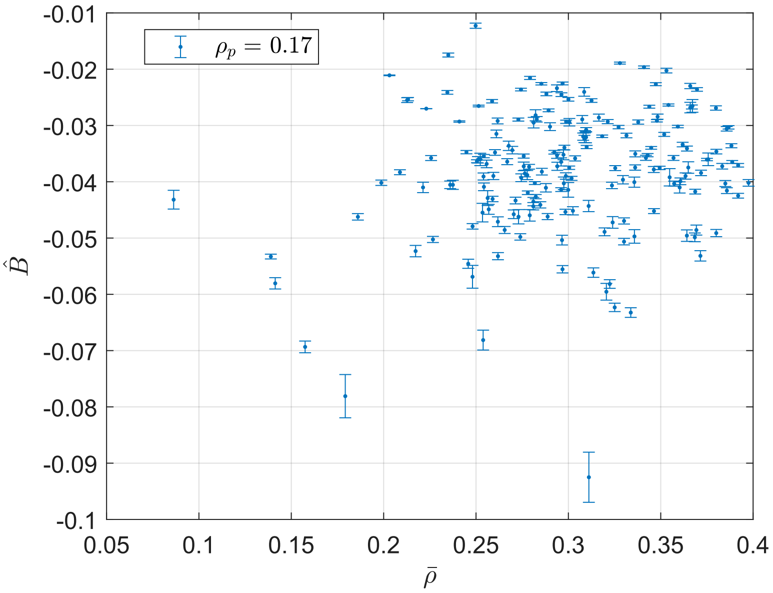

Figure 3.11: Relation between the multiscaling proxy $\hat{B}$ and the average correlation between each of the 187 stocks of the S&P 500 data set whose daily closing price is recorded from 02-Jan-1990 to 30-Nov-2022 with the others. The Pearson correlation measure between the two quantities is significant at a 5% level. The Kendall correlation coefficient is not reported because is not significant.

It is interesting to observe that a non trivial relation between $\hat{B}$ and $\bar{\rho}$ arises as already indicated in [53]. Compared to the previous results in literature, this behaviour is weaker due to the fact that the subset of stocks selected are characterized by high correlation and high market capitalization as they are all comprised in the S&P 500 index. This empirical evidence has been displayed [53] on various markets and therefore can be acknowledged as an additional stylized fact.

Chapter 4

Measures of Statistical Relations

When studying the statistical properties of financial markets one can usually focus on a single asset and devise independently how it behaves and what distinguishing traits define it. For instance, in Chapter 2 and 3 great attention has been placed on the features characterizing the asset returns time series, trying to highlight the empirical properties of their distribution and with the aim to understand more deeply the stochastic processes underlying the observed data.

While these analysis can give significant insights about the univariate properties of financial time series, it is important to remember that financial markets are complex systems made of interacting components and ignoring this structure of statistical relations may result in a lack of comprehension from a purely scientific point of view and from a risk management perspective [8, 9, 53].

In this chapter the efforts are directed towards the measure of the correlation between the price time series of the different assets comprised in the considered data set, with the objective to show how important multivariate properties can be in the study of financial data and in complex systems in general [8]. The analysis will be performed in two fashions:

- dynamically, to describe how the dependence between the various stocks evolves in time;
- statically, to showcase the topology of the correlation structure between the different assets.

4.1 Data Selection

As already presented in Chapter 2 the data set employed in this project includes time series of different length but whose end date is 30-Nov-2022 for every stock. Regarding the start date instead, there are assets whose price is reported from the first date available in the data set (02-Jan-1990), while others that are recorded from later dates. Furthermore, it is important to underline that for every asset the prices are all recorded on the same dates at the same frequency and therefore no temporal shifting between the time series must be taken into account. A simple plot to visualize the lengths of the time series employed in this project is displayed.

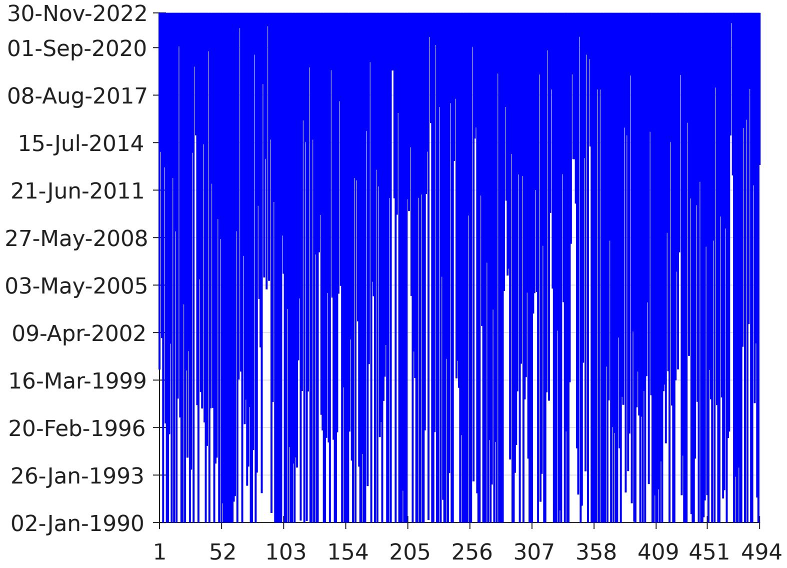

Figure 4.1: Every blue bar represents a stock and goes from the date on which its first price is recorded to the last one. It is clear that there are some stocks for which the first price comes later than the 02-Jan-1990, while all the time series end on the 30-Nov-2022.

The idea of the following sections is to compute the synchronous dependence between all the stocks whose price is recorded simultaneously, otherwise one cannot quantify how an asset is influenced by the others at the same time. The main issue is that, as already stated, not all the prices are recorded from the same date and therefore a way to select the best subset of stocks which "co-exist" is needed to perform effective analyses. Indeed, one would like to retain as many stocks as possible but without losing too much information selecting a start date which is too close to the end one.

In order to address this problem a simple method is proposed and then applied on the data set.

4.1.1 Data Selection Method

The considered data set can be described as a matrix M composed of N columns and T rows. Every column is a time series of a given stock in the set and thus each entry:

$$M_{ij} = \text{ price of stock } j \text{ at time } i \tag{4.1}$$

with $j \in [1, ..., N]$, $i \in [Date(1), ..., Date(t), ..., Date(T)]$, and $Date$ the vector of dates in which the prices of the stocks have been registered. For simplicity one can also identify the time index i just with values from $1, ..., T$ and in the following, when not specified, this choice will be made. The single time series can therefore be represented in this way:

$$M_{[:,j]} = [p_j(1), ..., p_j(t), ..., p_j(T)] \tag{4.2}$$

It is clear that when a price for stock j is not recorded at time t, the corresponding entry in the matrix M is reported as a missing value, represented as a numeric data type NaN [1] ($p_j(t) = NaN$). As already stated, the data set employed in this work is characterized by time series which share the same end index T, while are characterized by different start indices t_i^0. The purpose of the data selection mechanism is thereby to select an initial time index $t \in [1, T]$ such that a subset of the columns and rows of M, characterized by an optimal trade off between the number $n(t)$ of co-existent time series and the temporal length $T - t$ of the chosen time window, is obtained.

To perform this task one can initially perform a for-loop to count how many time series co-exist at each selected starting index t, and obtain a relationship between n and t:

$$(n(1), 1), ..., \ (n(k), t), ..., \ (N, t_{max})$$

where t_{max} is the first time index from which all the N time series in the data set co-exist. If the time series in the data set represent stocks, a set of weights w_j with $j = 1, ..., N$, can be built using the market capitalization of each stock:

$$w_j = \frac{\text{mktcap}_j}{\sum_{k=1}^{N} \text{mktcap}_k} \qquad \sum_{j=1}^{N} w_j = 1$$

In particular one can choose the average, the median or either the last available capitalization of each stock in the set depending on the purpose of the analysis.

Consequently, one can define an information function $I(t)$ such that it is proportional to the product between the information value of time length $T - t$, the number $n(t)$ of time series co-existing in the subset and the total sum of the weights associated to the time series, if these represent stocks. Obviously the three terms in the product can be tuned as needed, putting an exponent higher than 1 on the term one wants to favour. In this case it is chosen to give the same importance to all terms setting all exponents to 1.

$$\begin{cases} I(t) = \frac{(T-t)}{(T-1)} \cdot \frac{(n(t)-1)}{(N-1)} \cdot (\sum_{\text{subset(t)}} w_i) \\ t \in [1, ..., T] \\ n(t) \in [1, ..., N] \end{cases} \tag{4.3}$$

and subset(t) indicates the group of time series which co-exist from t.

It can be clearly observed that $I = 0$ if one takes a 1 length time window and/or just one time series. The information value is maximum if the time window and the number of vectors are maximized. Moreover, the sum of the weights takes into account that the information value is higher if higher capitalized stocks are included in the selection. This last point is a choice made in order to retain in the new set high capitalized stocks, which are more representative of the market and are fundamental in the study of correlations.

[1] "Not a number" in computing.

Practical Procedure To obtain the time t^*, at which the new subset must start, a simple for-loop is performed on the data matrix saving at each $t \in [1, T]$ the number $n(t)$ of time series that co-exist (that have non-NaN values before or from t). In particular, a couple (t, n) is stored in any iteration for which n is different from the previous one, obtaining thus an empirical relation $n(t)$. Furthermore, at each step of the loop the sum of the weights of all the $n(t)$ time series co-existing at that time t is computed, obtaining another vector $W(t)$. Having stored at each step this quantities it is straightforward to compute $I(t) = \frac{(T-t)}{(T-1)} \cdot \frac{(n(t)-1)}{(N-1)} \cdot W(t)$ during the iterations and at the end of the for-cycle the wanted t^* is just: $t^* = \mathrm{argmax}_t[I(t)]$, which always exists because the function lives in a finite space. A Matlab code is shown in B.

Remark About Classes. If the data-set is divided into classes (i.e. sectors for stocks), one can also apply the procedure on each subset of stocks belonging to the same class, obtaining a different t^* for each class. The greatest among these values can then be chosen in order to be sure to approximately preserve the distribution (fraction of total data) belonging to each class even in the reduced subset. However this application of the previously described method often reduces consistently the width of the time window causing a great loss of information. If, nevertheless, the average length of time series is almost the same among the different classes, even the single t^* obtained from the whole data set preserves the division among sub-groups.

4.1.2 Application on the Data Set

First of all the algorithmic procedure to extract t^* is applied on the data set. It can be observed that the function $I(t)$ defined in Equation (4.3) grows as $n(t)$ increases until a point over which the role of the factor $\frac{(T-t)}{T-1}$ becomes dominant and the information value starts to decrease.

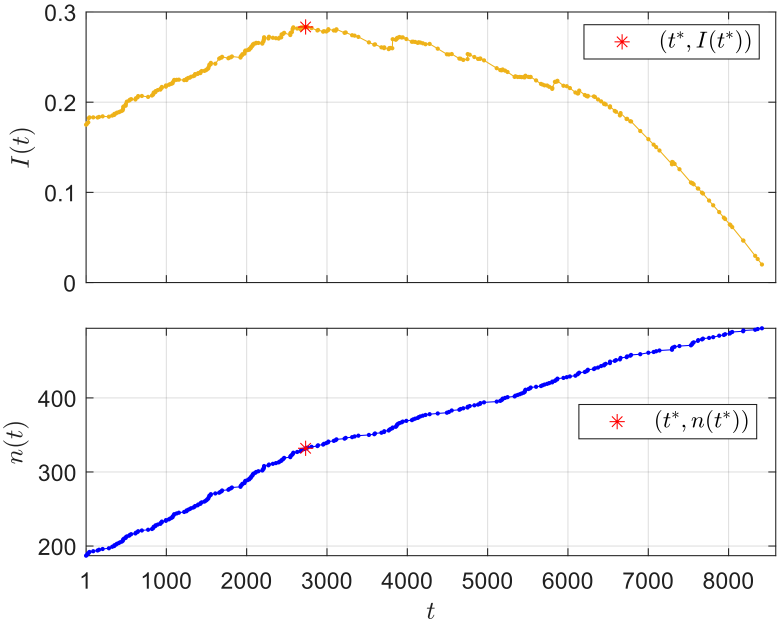

Figure 4.2: Result of the data selection procedure applied on the data set. While $n(t)$ obviously increases, $I(t)$ shows an inversion of its trend when the loss of information caused by the shortage of time points starts to become dominant. The procedure is stopped once $n(t^* < T) = N$ and this is the reason why the last computed $I(t)$ value is not equal to zero.

The results can be summarized in the following table:

$\boldsymbol{t^*}$	2734
$\boldsymbol{Date(t^*)}$	23-Jun-2000
$\boldsymbol{n(t^*)}$	332

Using these results, the new selected subset of the initial data comprises 332 stocks and goes from 23-Jun-2000 to 30-Nov-2022. To visualize this selection a plot of the time series lengths is shown.

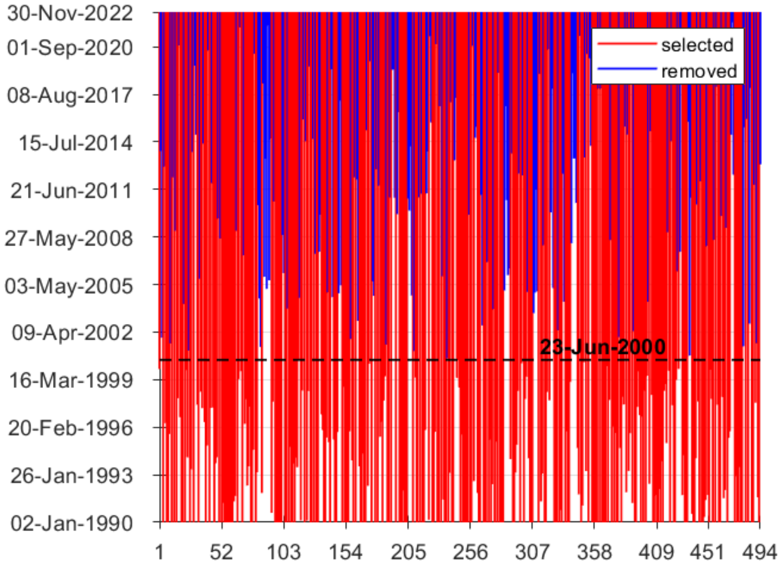

Figure 4.3: Every blue bar represents the length of the removed stocks, while red bars are the retained time series. The dashed black line shows the start date of the new selected data set.

To complete, a pie chart of the new data set is displayed. The sectors fractions are slightly modified with respect to the initial one but, for the purpose of the next analysis, it is preferred to work with this selection rather than applying the procedure described in "Remark about classes" (4.1.1). In fact, all the sectors are still represented and none of them has been modified of more than the 4%.

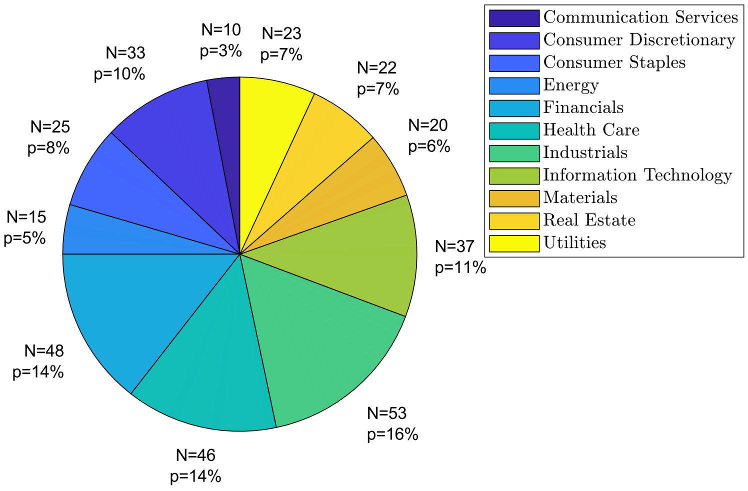

Figure 4.4: Pie chart of the new data set selected after the procedure described in this section. The number and the fraction of stocks belonging to each GICS *Sector* are also reported.

4.2 Dynamic Measure of Correlation

Since now, only the classical Pearson correlation coefficient has been presented as a measure of dependence between different variables (Subsection 2.2.3). Nevertheless, when one wants to study a system of dynamic dependency over a running time window, it is important to be aware of the excessive sensitiveness to outliers of standard measures, and it is needed to find solutions to overcome these issues. According to F. Pozzi et al. [37], a possible option is to assign a structure of weights to the observed events (daily returns in this work), and to choose a proper time window size such that the measure preserves its robustness and its statistical significance.

4.2.1 Weighted Pearson Correlation Coefficient

Even if already presented in Subsection 2.2.3, a more appropriate definition of the Pearson correlation coefficient is necessary to introduce the subsequent analyses. Given two vectors $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ of equal length L, the Pearson product-moment correlation coefficient between them is [37]:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

with

$$\sigma_{ij} = \frac{1}{L}\sum_{t=1}^{L}(x_t^i - \bar{x}^i)(x_t^j - \bar{x}^j) \tag{4.4}$$

$$\sigma_i = \sqrt{\frac{1}{L}\sum_{t=1}^{L}(x_t^i - \bar{x}^i)^2} \quad \text{the same for } j \tag{4.5}$$

$$\bar{x}^i = \frac{1}{L}\sum_{t=1}^{L} x_t^i \quad \text{the same for } j \tag{4.6}$$

If one thinks of these vectors as the sections of two given asset return time series $r_i(t)$ and $r_j(t)$ with $t \in [t' - \Delta t + 1, t']$ (i.e $L = \Delta t$), it is easy to understand that $\rho_{ij} = \rho_{ij}(t'; \Delta t)$ gives the linear correlation between the returns of the two assets in the chosen time window $[t' - \Delta t + 1, t']$.

If therefore one considers a set of n stocks' price time series, the correlation matrix between their returns time series at aggregation time τ is defined by this notation:

$$\rho_{ij}(t; \Delta t) \text{ s.t. } \begin{cases} i, j = 1, ..., n \\ t = \Delta t, ..., T - \tau \end{cases} \tag{4.7}$$

In this definition the correlation at time t is the value obtained having used all the data points in the set $[t - \Delta t + 1, t]$ [28, 29, 37]. If one then averages over all the elements in the data set:

$$\bar{\rho}(t; \Delta t) = \frac{1}{n(n-1)}\sum_{j=1}^{n}\sum_{i \neq j}^{n}\rho_{ij}(t; \Delta t) \qquad t = \Delta t, ..., T - \tau \tag{4.8}$$

the obtained values describe a single time series which gives an estimate of the mean of the dynamic correlation between the returns of all the stocks.

The main issue with this simple measure of correlation is that it counts every point in a time window in the same manner without giving more importance to the events which are closer to the given t. Moreover, this quantity can get unreliable in presence of fat-tailed distributions, it is not robust to outliers and the correlation matrix is not even invertible when $\Delta t < n + 1$ [37].

To try and overcome these issues the so-called weighted Pearson correlation coefficient between two vectors $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ of equal length L, can be defined as:

$$\sigma_{ij}^w = \sum_{t=1}^{L} w_t (x_t^i - \bar{x}_i^w)(x_t^j - \bar{x}_j^w) \tag{4.9}$$

$$\sigma_k = \sqrt{\sum_{t=1}^{L} w_t (x_t^k - \bar{x}_k^w)^2} \tag{4.10}$$

$$\bar{x}_k^w = \sum_{t=1}^{L} w_t x_t^k \tag{4.11}$$

$$\tag{4.12}$$

where $\sum_{t=1}^{L} w_t = 1$.

As before, if one thinks of these vectors as the sections of two given asset return time series $r_i(t')$ and $r_j(t')$ with $t' \in [t - \Delta t + 1, t]$ (i.e $L = \Delta t$), the weighted correlation between them can be written as:

$$\rho_{ij}^w(t; \Delta t) = \frac{\sigma_{ij}^w}{\sigma_i^w \sigma_j^w} \tag{4.13}$$

where for a set of n time series $i, j = 1, ...n$.

Averaging over all the elements in the data set one obtains:

$$\bar{\rho}^w(t; \Delta t) = \frac{1}{n(n-1)} \sum_{j=1}^{n} \sum_{i \neq j}^{n} \rho_{ij}^w(t; \Delta t) \qquad t = \Delta t, ..., T - \tau \tag{4.14}$$

In order to give a meaning to this measure the weights must assume proper values and since the main idea is to give more importance to the recent past the so-called "exponential smoothing" is adopted [37]:

$$w_t = w_0 \exp\left(\frac{t - \Delta t}{\theta}\right) \tag{4.15}$$

$$w_0(\theta) = \frac{1 - e^{-1/\theta}}{1 - e^{-\Delta t/\theta}} \tag{4.16}$$

$$\tag{4.17}$$

θ is a characteristic time which can be tuned to change the properties of the weighted averages: when $\theta \to \infty$ the weights are uniform, while when $\theta \to 0$ events in the past are less and less relevant and recent data points become the most important. Finally, the definition of w_0 comes from the constraint that the weighted correlation matrix must have the same positive semi-definiteness as the standard Pearson correlation matrix [37].

4.2.2 Measure of Dynamic Correlation on Data

The aim is thus to apply the described procedure on the group of 332 stocks extracted from the original data set and to compare the standard measure of correlation with the weighted one. In order to improve the correlation matrix numerical stability by avoiding the excessive distortions in the distribution of coefficients, the collapse of eigenvalues, the decrease of the estimated rank or the increase in the condition number of the largest full-rank sub-matrices, $\Delta t \in [50, 250]$ and $\Delta t/3$ are chosen [37].

The results for the averages presented in the Equations (4.8),(4.14) performed over all the stocks in the available period are shown for different values of Δt:

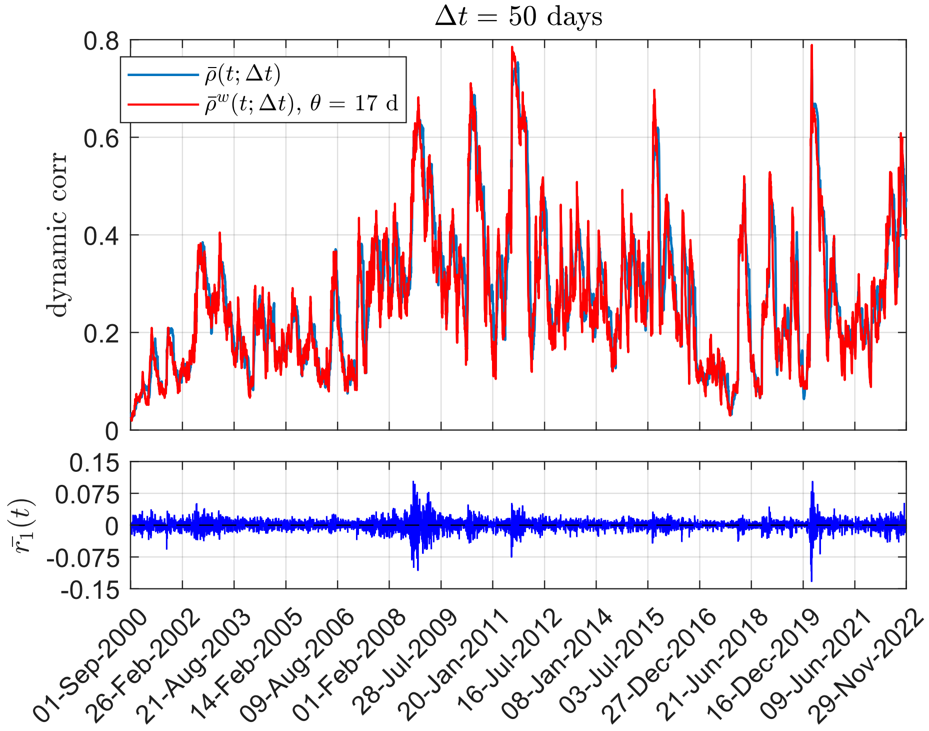

Figure 4.5: Average standard Pearson correlation coefficient vs its weighted version between daily returns time series (95 % significance level). The time window used is $\Delta t = 50$ days and $\theta = 17$ days. Using a small time window does not produce results which are considerably different between the two measures. On the bottom plot the daily returns averaged over all the stocks considered. Period of higher volatility are characterized by higher correlation. Overall the correlation is positive along the whole period.

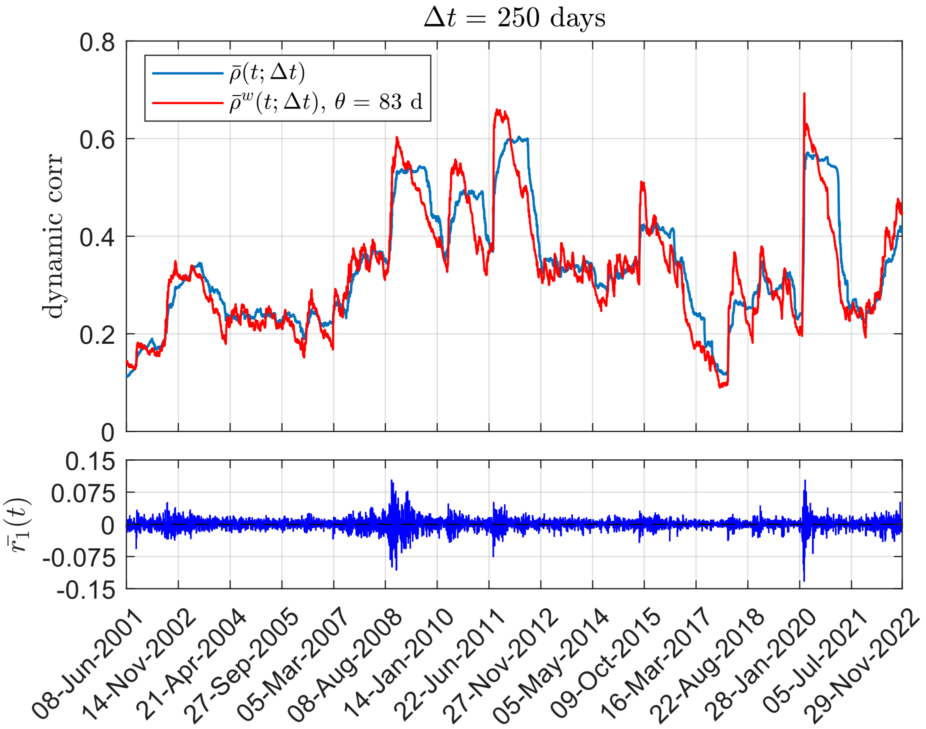

Figure 4.6: Average standard Pearson correlation coefficient vs its weighted version between daily returns time series (95 % significance level). The time window used is $\Delta t = 250$ days and $\theta = 83$ days. $\bar{\rho}^w(t;\Delta t)$ shows sharper peaks when the volatility is high, but the persistence is way shorter due to the weighting structure. On the bottom plot the daily returns averaged over all the stocks are considered. Period of higher volatility are characterized by higher correlation. Overall the correlation is positive along the whole period.

First of all it is important to remark that the displayed measures are statistically significant. In fact, for the standard Pearson correlation coefficient the p-value is computed for each correlation coefficient $\rho_{ij}(t;\Delta t)$ using a simple t-test. The obtained results whose p-val > 0.05 are set to zero because the null hypothesis of no correlation cannot be rejected and after this validation procedure the averages are performed [28, 29, 53, 37].

Regarding the weighted measure instead, it is preferred to perform a bootstrap resampling [9, 37]. For each time window the weighted correlation coefficients are computed, then every time series within the same window is randomly shuffled and the ρ^w_{ij} are computed again. This operation is performed 500 times for each time window in order to obtain a distribution for each coefficient from which a confidence interval can be extracted. Thereby, if for a single ρ^w_{ij} the value 0 is included within the 2.5 and 97.5 percentiles of its distribution, the null hypothesis cannot be rejected and the correlation coefficient is set to zero. In this way it is ensured that all the computed quantities are significant at a 95 % level.

Having assessed the statistical significance of the employed measured, the different behaviour of the two quantities can be observed. It is clear that the correlation rises during the period of financial crises (i.e. 2008, 2011, 2020) and of market shocks. This behaviour is usually described with the term *herd effect*, and is the tendency of investors to collectively overreact during financial crisis, when panic spreads through the market [37, 57]. It is important to underline that when market uncertainty increases, as in high volatility periods, risk managers must take into account this strong interdependence between different assets in order, for instance, to achieve an effective diversification when building investment portfolios [37]. The weighted correlation however tends to decay much faster after these events, thanks to the fact that the weighting structure reduces the persistence of shocks happened in the past. In particular when the 2 correlation measures are strongly different we can say that the market is in an extremely volatile period, in which spurious correlations tend to make the difference between $\bar{\rho}_t$ and $\bar{\rho}_t^w$ very unstable [37].

Just to visually give some further information on the selected data set a plot of the same dynamical weighted correlation but averaged over each sector is realized to highlight if some sub group of stocks tend to be less influenced by the movements of the other belonging to the same market.

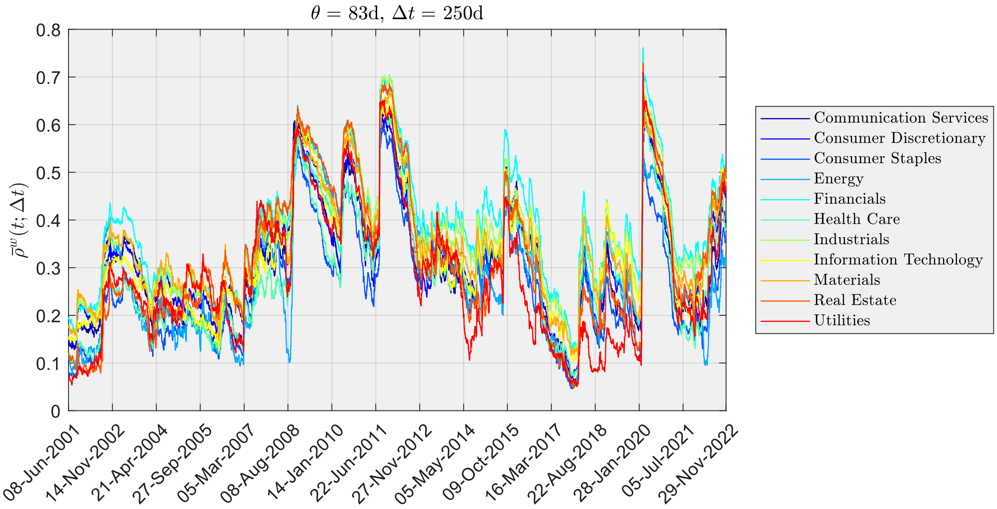

Figure 4.7: Average weighted Pearson correlation coefficient between daily returns time series (95 % significance level) of different sectors. The time window used is $\Delta t = 250$ days and $\theta = 83$ days. It is clearly observable that sectors show different behaviours, particularly after market shocks.

It is clear that not all sectors show the same dynamic behavior and the differences between them can be employed for practical purposes.

As an example, the Utilities sector showed a lower correlation with respect to the other mainly in the period that went from Jan-2014 to May-2020. The aforementioned period, as seen in figure 4.5,4.6, is basically characterized by financial stability, which means that this sector is on average less correlated with the other stocks when the market has an average low volatility. This property could therefore come in useful for portfolio building in stable periods, while in presence of market shocks the benefit would be strongly reduced.

4.3 Static Analysis of the Correlation Structure

From the dynamical study of correlations one can observe how the interdependence between different elements of a complex systems evolves. However, if the aim is to analyze the static structure of dependence, computed selecting a sufficiently long time period, another approach must be followed. It is in fact necessary to find a proper way to visualize the many reciprocal measures one computes and, in most cases, to single out the key information [58, 59, 36, 60, 9].

To show these ideas, the analysis begins with the computation of the Pearson correlation coefficient matrix, computed by considering the time series of the daily returns of all the 332 stock selected. It is chosen to employ all the period from 23-Jun-2000 to 30-Nov-2022 in order to exploit maximally the available data set, and with the objective to encompass in the analysis periods of both high and low volatility. In this way the herd effect can be mitigated and the measures obtained are more meaningful.

In the following picture an heat map showing the values of the Pearson correlation coefficients between all the couples of stocks in the data set is shown:

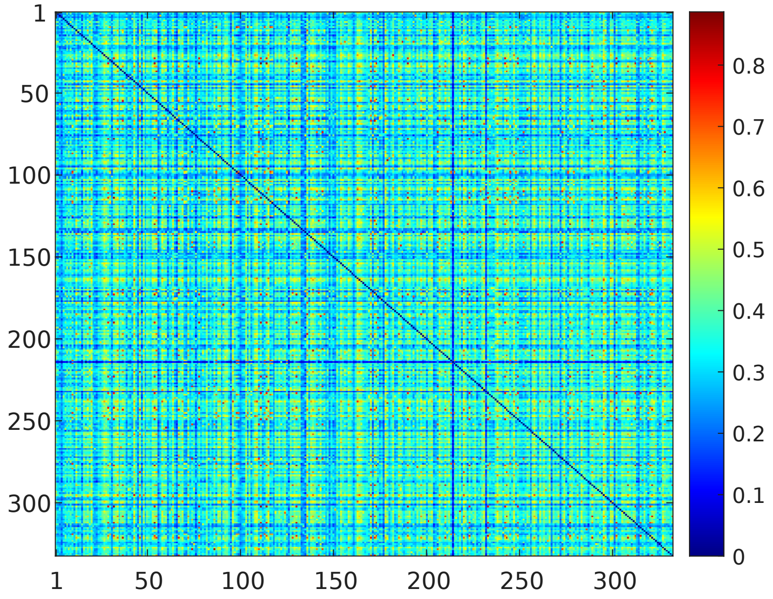

Figure 4.8: Heat map showing the values of the Pearson correlation coefficients computed using the daily returns time series of the 332 stocks in the data set from 23-Jun-2000 to 30-Nov-2022. The picture is obviously symmetric and the diagonal is totally black because it has been removed. The values have a 95 % level of statistical significance and are always non negative, showing an overall positive correlation between the different assets price variations.

It is very interesting to observe that there are some stocks for which the correlation with every other element in the set is very low. For instance, the stock identified with the number 214 corresponds to Newmont Corporation which is a gold mining company. Indeed, when investors are worried about economic instability, they often seek the store of value that gold offers. As a result, gold prices can rise or remain stable even when other asset classes, such as stocks or bonds, are experiencing significant declines [61] and this can be one of the reasons why this stock showcases a very low set of correlation coefficients with the other assets.

4.3.1 Minimum Spanning Tree

In general, in a complex system consisting of many interacting elements, one can use appropriate tools to represent the intricate structure of relations between its elements. One of the main ideas that has gained great success in the last twenty years is the use of networks, mathematically defined with the term *graph* [62, 59, 36, 58]. One can associate a node with each element and an edge with each interaction/relation. In the case of financial markets, each stock price is related to the price of all other stocks, and in the present case the displayed correlation matrix (figure 4.8) can be visualized in a different way employing a network structure.

The first straightforward idea is indeed to define directly a graph considering the correlation matrix, from which the diagonal has been removed, as the adjacency matrix of a weighted undirected graph [58]. In this way a complete graph with N nodes (stocks) and $N(N-1)/2$ edges (correlations values) is obtained. The problem is that most of the information contained in a correlation matrix is redundant [58] and from a visualization point of view representing a network with $N(N-1)/2$ edges is quite complicated and usually impractical for large N (i.e. $N \gtrsim 10^2$). Moreover, a complete graph representation does not allow to identify the presence of clusters within the correlation matrix, while their detection can be useful to define an intrinsic taxonomy of the selected data set and to devise groups of highly correlated nodes within the market.

To pursue this objective one first needs to define a proper metric that quantifies the "distance" between two variables, daily returns time series in the present case, based on their correlation. Indeed one can directly define a metric using as a distance a function of the simple Pearson correlation coefficient [60, 58]:

$$d(i,j) = \sqrt{2(1-\rho_{ij})} \tag{4.18}$$

It is interesting to observe that this definition satisfies the three axioms of a metric distance [60]:

1. $d(i,j) = 0$ if and only if $i = j$ as $\rho_{i=j} = 1$;
2. $d(i,j) = d(j,i)$ as the correlation coefficient is symmetric;
3. $d(i,j) \leq d(i,k) + d(k,j)$ as the defined distance is equivalent to the euclidean distance between the two vectors on which ρ_{ij} is computed.

It directly follows that from the Pearson correlation matrix one can define a distance matrix $D_{ij} = d(i,j)$ that again can be seen as the adjacency matrix of a weighted graph [60, 58, 36].

What comes next is to find a connected graph whose topological structure represents the correlation among the different elements but that is greatly reduced in the number of edges with respect to the complete graph. The simplest solution is the so-called spanning tree, a graph with no cycles that connects all the nodes. It follows that the structure with these properties such that retains the maximum possible number of correlations is the well-known Minimum Spanning Tree (MST), which can be extracted from a complete graph by a simple algorithmic procedure [9, 60, 58]. In this work the famous Kruskal's algorithm [63] is employed, whose procedure can be briefly summarized for the present study[58, 64]:

1. create a forest (set of disjoint trees) F, at the beginning each nodes is a tree itself;

2. create an ordered list S of edges (i, j), ranking them by increasing $d(i, j)$ defined as in Equation (4.18);

3. Take the first element in the list and add the edge to the graph;

4. Take the next element and add the edge if the resulting graph is still a forest or a tree;

5. Iterate point 4 until S is empty or F becomes the spanning tree of the complete graph.

If the starting graph has N nodes, the resulting MST tree has $N - 1$ edges such that the sum of their weights $d(i, j)$ is minimized compatibly with the fact that the obtained structure is a tree which includes all the initial N nodes. For a graph with E edges and N nodes, Kruskal's algorithm can be shown to run in $O(E \log E)$ time, all with simple data structures

The results of this procedure can be shown on the correlation structure between the 332 stocks selected from the SP&500 index. The correlation, has already stated, is computed using the Pearson correlation coefficient between the daily log returns time series of the selected stocks from 23-Jun-2000 to 30-Nov-2022 (4.8). From these quantities, the distance matrix D is simply devised and the resulting MST is obtained:

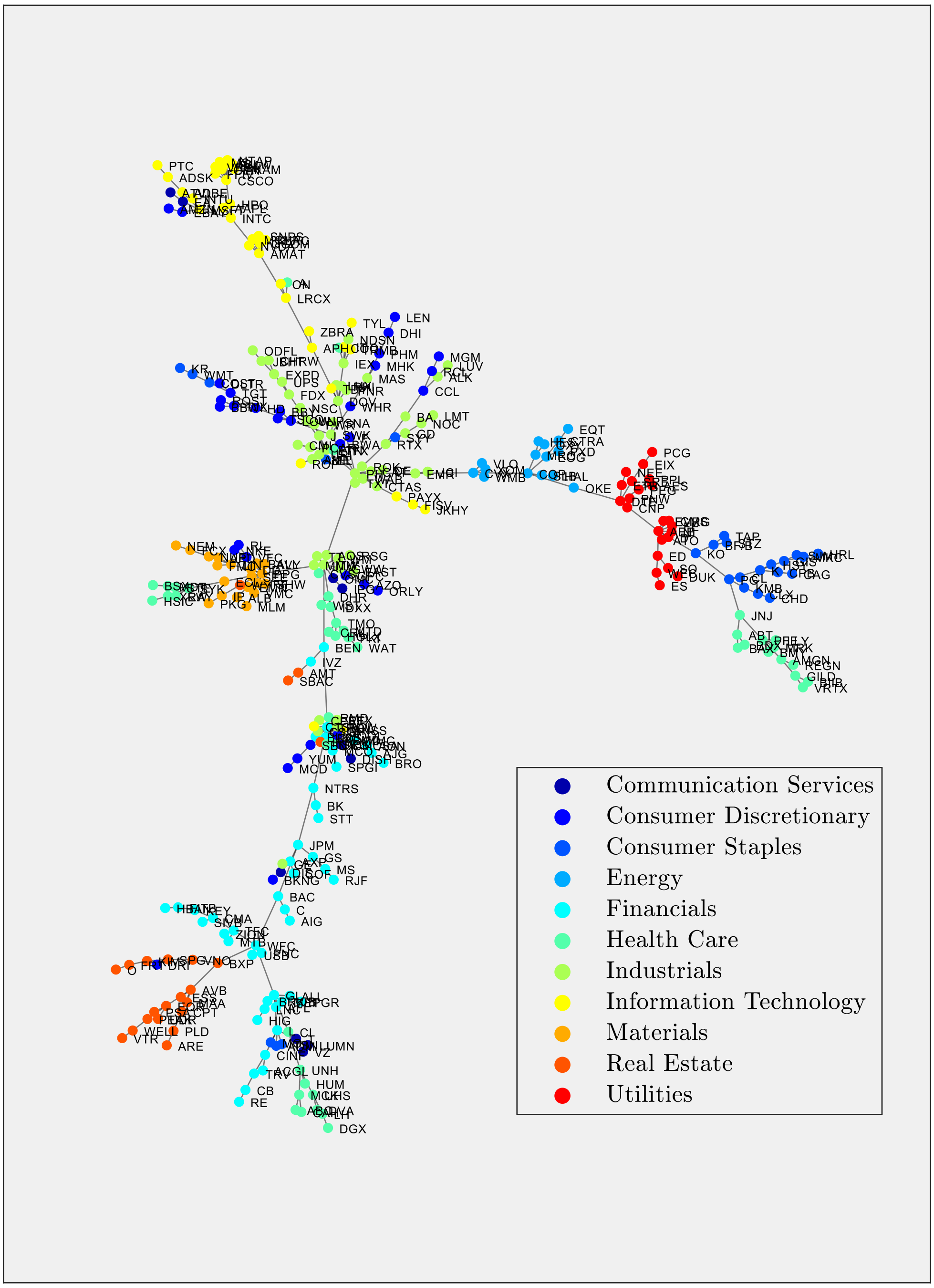

Figure 4.9: MST built from the cross-correlation matrix of the daily returns of 332 stocks comprised in the S&P 500 index over the period from 23-Jun-2000 to 30-Nov-2022. The colors represent the different GICS classification of the various stocks. It can be observed how naturally, using the algorithmic procedure, the sector clusters are almost completely retrieved.

It is evident that the devised structure naturally shows clusters which are well compatible with the GICS sector classification. Moreover, edges between stocks belonging to different sectors are also significant.

As an example, the subtree that has as a root the PPG node is analyzed.

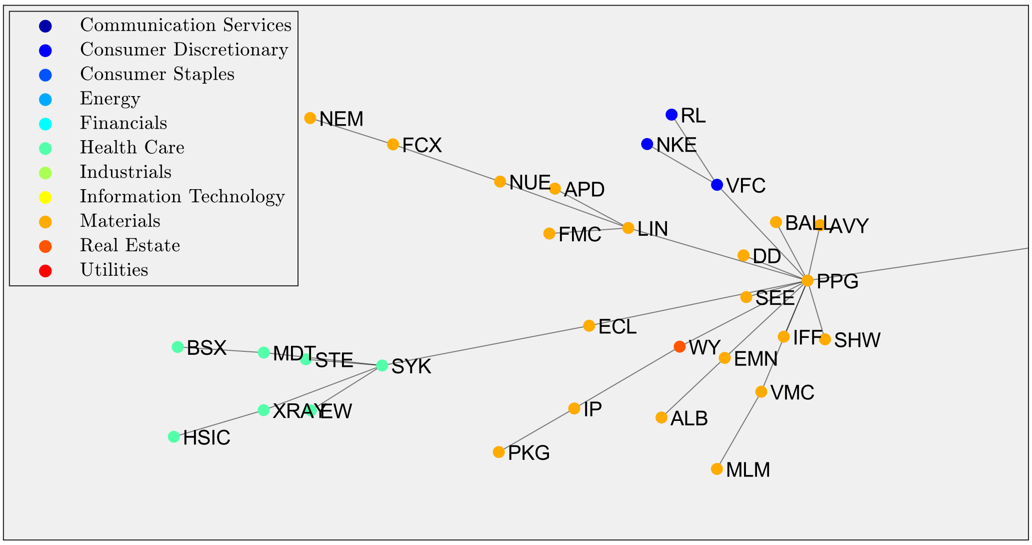

Figure 4.10: PPG subtree of the MST built from the cross-correlation matrix of the daily returns of 332 stocks comprised in the S&P 500 index over the period from 23-Jun-2000 to 30-Nov-2022. The colors represent the different GICS classification of the various stocks. The PPG node is the root of the subtree and is connected to many other subtrees due to its importance in the market.

Ppg Industries Inc. is an American company global supplier of paints, coatings, optical products, specialty materials, chemicals, glass and fiber glass with a large amount of activity in refinishing products [58]. It is indeed connected to 13 other companies 11 of which belong to the same sector. The two exceptions are VFC and WY nodes: the first corresponds to VF Corporation which controls JanSport, Eastpak, Timberland, and The North Face brands, while the second indicates the Weyerhaeuser Company which operates in three major business: timberlands, wood products and real estate. The first edge between PPG and VFC can be naturally understood because to produce complex refined products different chemicals and specialty materials are needed. On the top of that, it is also natural that WY and PPG are correlated as both are related to building materials for homes and other structure. Moreover, if one looks at the nodes connected to the VFC one, both the two companies Nike Inc (NKE) and Ralph Lauren Corp (RL) operate in the Clothing business.

What makes this information filtering technique very powerful is that the MST structure has been extracted solely from the cross-correlation matrix without any other a priori information about the system, and the flexibility of the procedure makes it very applicable also on data sets belonging to other fields of study [58, 59].

4.3.2 Other Methods and Ideas

Even if very powerful, the Minimum Spanning Tree does not allow complex connections between group of stocks as triangular cycles (3-cliques). In order to retain the filtering effectiveness of the MST, but also allowing the presence of more complex structures which involve more links, other information filtering tool have been proposed [59, 36, 58, 9]. For instance, one can build graphs embedded on surfaces with a given genus g (number of holes in the surface) and, with a simple modification of the previously described algorithm, a spanning graph with $3n - 6 + 6g$ edges can be obtained [58]. When $g = 0$ the structure devised is a triangulation of a topological sphere and is called Planar Maximally Filtered Graph (PMFG) [58, 9, 59]. It can also be proven that the MST is always a subgraph of the PMFG, and the latter is just an extension of the former with more links to enhance the complexity and the thoroughness of the description [58].

To conclude, it is cited also the possibility to build a hierarchical tree from the previously defined metric (Equation 4.18). One can indeed define from $d(i, j)$ the so-called subdominant ultrametric distance $d^{<}(i, j)$ as the maximum value of any Euclidean distance $d(k, l)$ detected by moving in single steps from i to j through the shortest path connecting i and j in the MST. By exploiting the detected subdominant ultrametric space it is possible to obtain a taxonomy of the analyzed data set which naturally gives a kind of hierarchical organization that is able to isolate economically meaningful groups of stocks [60].

In order to avoid making this chapter too long in terms of methodologies, it has been chosen to provide results just for the dynamical correlations and for the MST. Nevertheless, it is clear that extending the analysis with other tools, as the last two mentioned, can be very useful to gain other insights about the interdependence structure of the selected data set.

Chapter 5

Physics Inspired Modeling and the PUCK Model

In this chapter it is presented a relatively novel stochastic model for market price called the PUCK model [65], which has been employed as a novel type of time series data analysis tool, as well as to describe the behaviour of a particular type of trader in an agent-based model of financial markets [66, 67].

The scaling behaviour of the price time series obtained with the PUCK model is analyzed to understand if it is possible to obtain multiscaling properties just by changing its parameters. Moreover, employing both the dynamical weighted GHE and the RNSGHE methods described in chapter 3, a relation between these parameters and the scaling exponents is retrieved.

To complete, the analysis of data is performed with both the PUCK model and the dynamical weighted GHE method and the results are compared.

5.1 Physics Inspired Modeling

The stylized facts presented in chapter 2 have been used in the last 40 years as a benchmark to test the validity of every novel financial model [8, 10, 48, 12, 13]. In particular, the recent developments in the description of financial markets have been characterized by the employment of the so-called agent-based models (ABM), in which, the macroscopic dynamics, for instance of the price of an asset, is modeled describing the microscopic behaviour of the single traders (agents) [10].

This framework naturally originates from statistical physics considerations as the idea is to devise the macroscopic properties of a system from its microscopic interactions. For instance, in [68] Bak et al. the order book is modeled as a physical reaction diffusion process: two types of particles are inserted on each side of a pipe and move randomly, every time two particles collide, they are annihilated and two new particles are inserted. In this case particles are the orders (sell/buy, 2 types), the finite pipe represents the order book and the collision is a transaction whose price $p(t)$ is recorded [68, 10].

Another peculiar example of an analogy between a model of a physical system and a typical financial problem is presented in [69]. In this paper it is shown that the classical portfolio optimisation problem: "given a set of financial assets, characterized by their average return and their risk, what is the optimal weight of each asset, such that the overall portfolio provides the best return for a fixed level of risk, or conversely, the smallest risk for a given overall return?" [69], leads to equations which are analogous of those defining the locally stable configurations in a spin-glass.

The model by Takayasu et al. [13] which will be presented later on in the chapter, does not describe how some macroscopic properties arise from the interaction of microscopic components of the system. It is rather a random walk model in which the "walker" is subjected to a force described by a time-dependent potential function whose center is given by a moving average of market price. This force directly modifies the diffusion properties of the process and can also be related to the strategy of the dealers in the market. In addition, the continuous limit of the model gives as a result the Langevin equation with fluctuating viscosity and mass.

There are many other examples of models which could be cited that try to apply a physical framework or approach to financial markets [10]. It is however preferred to avoid presenting other instances in order to keep the focus on the aspects which are needed to introduce the following sections.

5.1.1 Minimal Agent Based Model for Financial Markets

Lux-Marchesi Model

At the end of the Nineties Lux and Marchesi introduced an agent based model to show that the empirical characteristics of financial prices can emerge from the interactions of a large ensemble of market participants [70, 71]. In this model the pool of traders is divided into two groups: "fundamentalists" who follow the efficient market's hypothesis and expect the price p_t to follow a fundamental value p_f, which is the discounted sum of expected future earnings; and the "noisy traders" who do not believe in an immediate tendency of the price to revert to its underlying fundamental value and attempt to identify price trends and patterns [70] (these are further divided into optimists who always buy and pessimists who always sell). Moreover, the single agents can move from one class to the other and the account of the behaviour of other traders as a source of information, which results in a tendency towards herding behaviour, is also inserted in this framework. It is important to underline that the novelty of this model has been to adopt a mass-statistical formalization inspired by statistical physics: individuals react to certain economic forces by modifying their behaviour with a certain (endogenous) probability. Assuming a Gaussian external driving force which affects the market through the operations of fundamentalist traders, the model has naturally led to fat tailed distributions of the returns, absence of autocorrelations for the price variations and strong persistence in the volatility. It follows that the scaling properties, as they are absent in the external driving force, are generated by the interaction of economic agents with heterogeneous beliefs and strategies in the simulated market [70, 71].

Alfi-Cristelli-Pietronero-Zaccaria Model

This model presented in [66, 67] is inspired by the Lux-Marchesi one but is much simpler with respect both to the number of parameters and the rules for the dynamics. The main elements are the following:

- "Fundamentalists": these agents have as reference a fundamental price p_f derived from standard economic analysis of the value of the stock. Their strategy is to trade when the price departs from the reference value and bet on the fact that the price will return to the reference value. These traders are usually represented by institutions, their time scale is relatively long, and they tend to stabilize the price around the reference value.
- "Chartists": these agents consider only the price time series and tend to follow positive or negative trends. These traders have usually a time horizon shorter than the fundamentalists and they are responsible for the large price fluctuations which correspond to bubbles or crashes. Moreover, they induce a destabilizing tendency in the market and in the Lux-Marchesi model they are called "noisy traders".
- Herding effect: the tendency to follow the strategy of the other traders; it is also complemented considering the possibility that traders can change their strategy from fundamentalist to chartist and vice-versa depending on various elements.
- Price behaviour: each agent looks at the price from her perspective and derives a signal from its value which will be crucial in deciding her strategy.

In this model the concept of self-organization arises very naturally. A price which is very stable demotivates agents to trade this stock and will naturally lead to a decrease of the number of agents. On the other hand, a small number of agents leads to large fluctuations in the price which presents opportunities of arbitrage that will appeal more traders. It follows that the system will self-organizes around the number of traders which corresponds to a situation of intermittency, leading to a state which corresponds to the empirical stylized facts [66].

As in the Lux-Marchesi model the agents are divided into two classes, but in this case there's no need to divide chartists into two further subcategories (optimists and pessimists). This last class is indeed described by a potential method [72, 65, 13] such that its agents try to follow the trend and bet that the price will further move away from the actual price, in such a way that they create a local bubble which destabilizes the market. The stochastic equation for the price which describes this behaviour is nothing but the main equation of the PUCK model by Takayasu et al [13], which is introduced in the next section.

5.2 PUCK model

In [13] it is proposed a model of a random walker in a randomly changing potential function called the PUCK model (Potentials of Unbalanced Complex Kinetics model). In this model the center of the potential function moves with the moving average of the random walker's position, and the potential function is given by a quadratic function with its curvature slowly changing around zero. It can be written the following mathematical form:

$$x(t+1) - x(t) = -\frac{d}{dx} U_M(x;t)|_{x=x(t)-x_M(t)} + f(t) \tag{5.1}$$

$$U_M(x;t) \equiv \frac{b(t)}{M-1} \frac{x^2}{2} \tag{5.2}$$

$$x_M(t) \equiv \frac{1}{M} \sum_{k=0}^{M-1} x(t-k) \tag{5.3}$$

where $f(t)$ is a random external noise (usually Gaussian with zero mean and unitary variance), $b(t)$ is the coefficient of the quadratic potential, M is the size of moving average to define the center of potential function $x_M(t)$. The model has been originally built to describe high frequency data in financial markets but in this work it is applied for daily closing prices as the statistical procedures that will be employed have been extensively tested on these kind on data [18, 17, 52].

The main idea is to measure the multiscaling properties of the model with the aim to devise a relation between the generalized Hurst exponent H_q and the coefficient $b(t)$. In this manner it could be possible both to obtain an interpretation of H_q in term of the coefficient of the quadratic potential of the model, and to possibly exploit scaling exponents to extract $b(t)$ from data.

5.2.1 Continuum Limit of the PUCK Model

It can be shown that the Langevin equation with fluctuating viscosity and mass is derived as a continuous limit of the PUCK model [73]. First of all Equation (5.1) can be rewritten in this form:

$$P(t+\Delta t) - P(t) = -\frac{\partial}{\partial x} \Phi_M(x,t) \bigg|_{x=\frac{P(t)-P_M(t)}{M-1}} + F(t) \tag{5.4}$$

in which $F(t)$ is and independent random noise, and the potential function $\Phi_M(x,t)$ can be expanded as:

$$\Phi_M(x,t) = \sum_{n=1}^{\infty} a_n(t;M) \frac{x^n}{n} \tag{5.5}$$

In real data usually the observed potential functions are such that $\langle P(t+\Delta t) - P(t) \rangle \simeq 0$ and therefore $a_1(t;M) \simeq 0$. On the other hand, $a_2(t;M)$ is non negligible and its dependence from M is very weak [73]. It follows that this expression is nothing but the generalization of Equation (5.2) and thus $a_2(t;M) \propto b(t)$.

Considering the limit for $\Delta t \to 0$ keeping $\tau = M\Delta t$ constant, Equation (5.4) becomes:

$$\frac{d}{dt} P(t) = -\sum_{n=1}^{\infty} a_n(t;\tau) \left[\frac{P(t) - P_\tau(t)}{\tau} \right]^{n-1} + G(t) \tag{5.6}$$

where $P_\tau(t) \equiv \frac{1}{\tau} \int_{t-\tau}^{t} P(s)ds$, $a_n(t;\tau) \equiv \lim_{\Delta t \to 0} a_n(t; \tau/(\Delta t))(\Delta t)^{n-2}$ and $G(t) \equiv \lim_{\Delta t \to 0} F(t)/\Delta t$.

For small positive τ, comparing the resulting equation with the standard Langevin equation in 1 dimensional space of location $r(t)$ one obtains:

$$m\frac{d^2}{dt^2}r(t) + \mu\frac{d}{dt}r(t) = f(t) \tag{5.7}$$

From this expression the market viscosity corresponds to $\mu = 1 + \frac{a_2(t;\tau)}{2}$ and the mass of the market price is given by $m = -\frac{a_2(t;\tau)}{6}\tau$. When $-2 < a_2(t;\tau) < 0$, that corresponds to negative $b(t)$, the motion of the market is described by an ordinary Langevin equation with positive mass and positive viscosity. When $a_2(t;\tau) > 0$ the mass can be negative and for $a_2(t;\tau) < -2$ also the viscosity becomes negative. It follows that a direct correspondence with a physical situation is possible only when the market is in, as it will be defined in the next section, a "trend following" state [73].

5.2.2 Constant b

The coefficient $b(t)$ of the quadratic potential defines the behaviour of the walker and the force to which it is subjected. In this section the relevant properties at constant b are presented. Three main cases can be acknowledged:

- $b = 0$ corresponds to a simple random walk, no potential is present and therefore no force is acting on the walker;
- for $b > 0$ the random walker is attracted to the moving average of its own path, the diffusion becomes slower than the random case; this behaviour can be called as "mean reverting";
- for $b < 0$ the random walker is pushed away from the moving average of its traces and the walker diffuses faster than the random case; this behaviour can be called as "trend follower".

There is a sharp transition in the diffusion properties of the model at $b = -2$: the repulsive force from the center of the potential function is larger than the effect of the random noise $f(t)$ and $x(t)$ follows an exponential growth which can be interpreted as market crashes or bubbles [13]. For positive large b instead the potential force is so strong that the motion becomes a diverging oscillation [13]. In particular the stochastic process defined by equation (5.1) can be proven to be non-stationary for $b \leq -2$. As an example the paths generated by the model at constant b are shown.

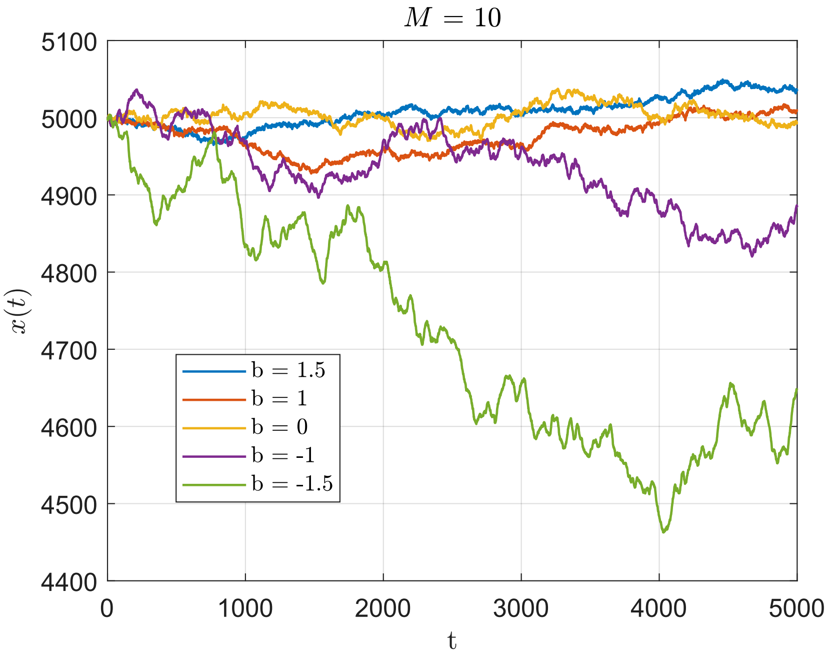

Figure 5.1: Path simulated with Equation (5.1) fixing $M = 10$ and $b(t) = const = b$. The initial position is $x(1) = 5000$ and the path is $T = 5000$ steps long.

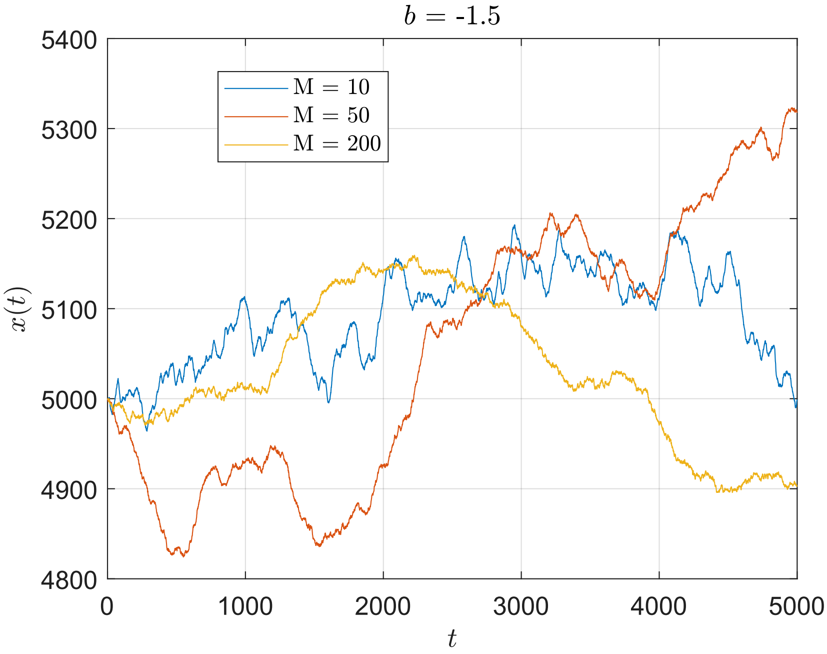

Figure 5.2: Path simulated with equation (5.1) fixing $b(t) = -1.5$ for every t and varying M. The initial position is $x(1) = 5000$ and the path is $T = 5000$ steps long.

Already from the sample paths it is clear that varying the parameters of the model the time series have very different properties: negative b produce smoother paths compared to positive b when M is fixed and not too large (fig. 5.1); when increasing M at fixed b the produced paths seem to preserve their diffusion properties but tend to lose their long-range dependence.

From the point of view of the statistical properties of the time series generated Takayasu et al. showed that, when $b(t) = const$ and the noise $f(t)$ in Equation (5.1) is a white Gaussian noise[13]:

- The cumulative distribution of price differences $\Delta x(\tau; t) = x(t + \tau) - x(t)$ is well approximated by a gaussian with variance depending on b and M, therefore no heavy tails are observable;
- The autocorrelation of $v(t) = x(t) - x(t - 1)$ is always positive and decays exponentially for any negative b-value, while for a positive b-value it is characterized by an oscillatory behavior;
- The autocorrelation of the volatility, defined as $v(t)^2$, always decays exponentially and therefore no longer correlation and resulting volatility clustering can be observed at constant b;
- Analyzing the behavior of $\sigma^2(t) = \langle x(t) - x(0) \rangle$, the model shows slower abnormal diffusion for $b > 0$ and faster abnormal diffusion for $b < 0$ regardless of M (to be proved).

5.2.3 Random b(t)

In order to try to reproduce some of the empirical stylized facts which are well known to characterize financial data [11, 8], Takayasu et al. considered the case that the potential coefficient $b(t)$ changes randomly with time. In particular, assuming that $b(t)$ follows a random walk in a fixed potential function, its process can be described by this equation:

$$b(t + 1) = (1 - c_0)b(t) + g(t) \tag{5.8}$$

where c_0 is a constant $\in$ [0,1] and $g(t)$ is a normal Gaussian noise with zero mean and variance G [13].

The relevant case for this discussion is when c_0 is not so small compared with G (i.e. $c_0 = 0.0015$ and $G = 0.000784$) and the probability of $b(t) \leq -2$ is negligible [13]. In fact, in this case some of the basic statistical properties of the model become similar to that of real market price fluctuations[13, 11, 8]:

- The cumulative distribution of the price variation $v(t)$ has heavy tails, well approximated by power laws;
- The autocorrelation function for $v(t)$ decays rapidly to zero;
- The autocorrelation function for the volatility $v(t)^2$ slowly decays to zero and therefore the volatility clustering can be observed;
- Abnormal diffusion can be found for small time scales, while for large time scales the normal diffusion property is preserved.

It must be highlighted that these properties emerge only if the simulations one performs are long enough (i.e. $T \sim 10^6$ time steps) and not for every simulation conducted. This model is thus not ideal for daily data modeling. Moreover, when c_0 is relatively large compared with the G (i.e. $c_0 = 0.02$ and $G = 0.000784$) the statistical properties of prices are confirmed to be nearly equivalent to the case of normal random walk. Finally, when c_0 is very close to zero (i.e. $c_0 = 0.0001$ and $G = 0.000784$) $b(t)$ strongly fluctuates, and there is a non negligible probability that $b(t) \leq -2$ making the process unstable and non-stationary [13].

As an example a path generated by the model when $b(t)$ follows Equation (5.8) in the only case of interest ($c_0 = 0.0015$ and $G = 0.000784$) is shown:

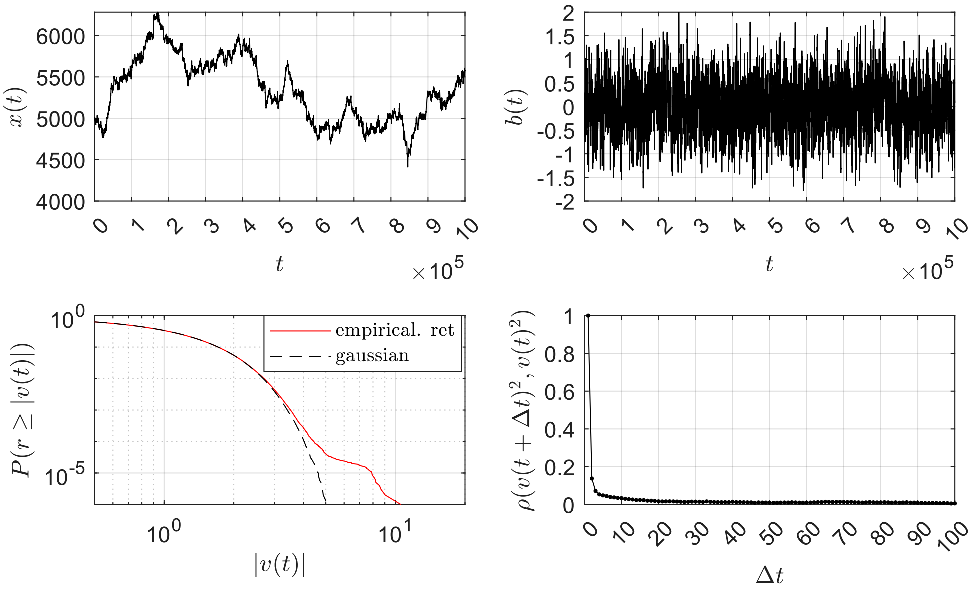

Figure 5.3: Path simulated with Equation (5.1), $M = 2$ and $b(t)$ following Equation (5.8) with $c_0 = 0.0015$ and $G = 0.000784$. The initial position is $x(1) = 5000$ and the path is $T = 10^6$ steps long. The cumulative distribution function for the price variation $v(t)$ and the autocorrelation of the volatility $v(t)^2$ are also shown.

From the plot it can be deducted that the random process for $b(t)$ affects both the diffusion properties, as already mentioned in [13], and the roughness of the path.

5.3 Measuring Scaling Properties of Simulated Time Series

5.3.1 Case of Constant b

At $b(t) = b = \text{const}$ it is employed the RNSGHE method (3.5.3) to study the scaling properties of the absolute value of the price variation $|x(t+\tau) - x(t)|$ of time series simulated using the PUCK model (5.1), choosing $f(t)$ to be a gaussian noise with $\mu = 0$ and $\sigma = 1$. The following scaling relation is therefore assumed:

$$\Xi(\tau, q) = \mathbb{E}[|x(t+\tau) - x(t)|^q] \sim K_q \tau^{qH_q} \tag{5.9}$$

The model at constant b produces price variations which are normally distributed [13] and therefore one could choose also $q \geq 1$ without affecting the estimate [18]. For simplicity, and coherence with the subsequent analysis, 50 equally spaced values of $q \in [0.02,1]$ are chosen. Regarding τ_{max}, for the estimates at constant b, $\tau_{max} = 200$ has been chosen. For $b \geq 0$ the volatility $v(t)^2$ shows an oscillating behaviour around zero [13] and therefore the choice of τ_{max} should not bias the computation of the multiscaling proxy $\hat{B}$ [35]. On the other hand, when $b < 0$ the volatility shows a positive autocorrelation that decays exponentially [13]. The main issue is that the ACSR method (3.5.3) is designed for real financial time series which show a power law like decay of the volatility autocorrelation [8, 18], and it is therefore chosen to use $\tau_{max} = 200$ also for $b \geq 0$. To finally confirm the observed behaviours the estimates are realized also setting $\tau_{max} = 19$ [35, 34] which is a value small enough to reduce the potential mixing of the high autocorrelation state with the high noise state (see appendix B).

100 simulations of $T = 5 \cdot 10^4$ time steps are performed using all the possible 36 combinations of the following set of parameters: $b = [-1.9, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 1.9]$; $M = [5,10,50,200]$.

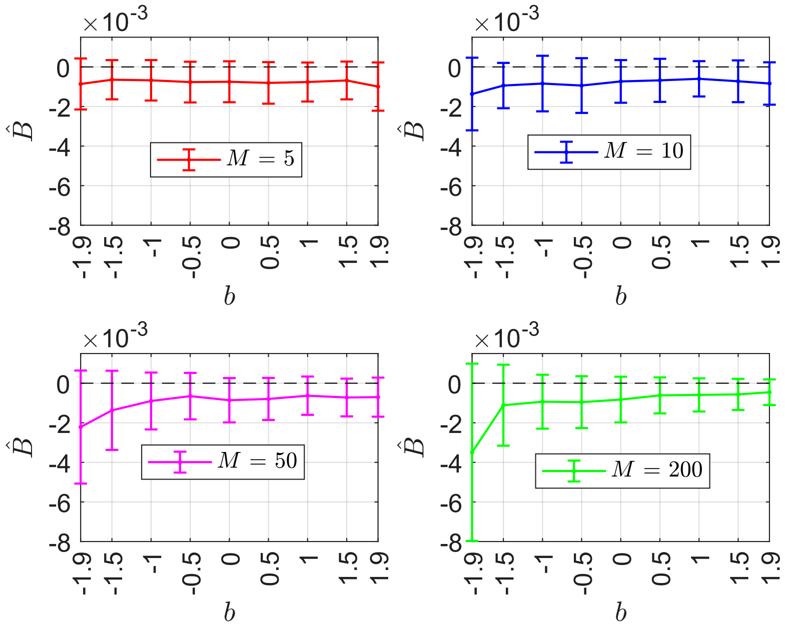

Figure 5.4: Estimate of the multiscaling proxy (Equation 3.57) varying both b and M. The RNSGHE method has been applied using $\tau \in [1,200]$, $q \in [0.02,1]$. For larger M both the average and the std of $\hat{B}$ become larger due to the higher persistence of the volatility autocorrelation. The single error bar represents the standard deviation of the estimate over the set of 100 simulations performed for each choice of the parameters.

Even if, as it could be expected [35], increasing M and decreasing b the multiscaling proxy on average increases due to the enhanced persistence of the volatility autocorrelation, all the obtained values are still not big enough to make the process multiscaling. Moreover, the standard deviation of the distribution of the estimates grows for the same reason, as the autocorrelation introduces a bias in the multiscaling proxy estimate that can also result in slightly positive values of $\hat{B}$ [35].

Regarding the multiscaling test, the $3579/3600 = 99.42\%$ of the simulations performed failed both the F-test and the t-tests (see table 3.1) while the remaining 21 rejected the null hypothesis only for the F-test. It can therefore be safely claimed that at constant b the process is not multiscaling. It follows that one can quantify the scaling properties of the process just from one exponent H_q and, given its theoretical importance, it is chosen to show the plots for $q = 1$:

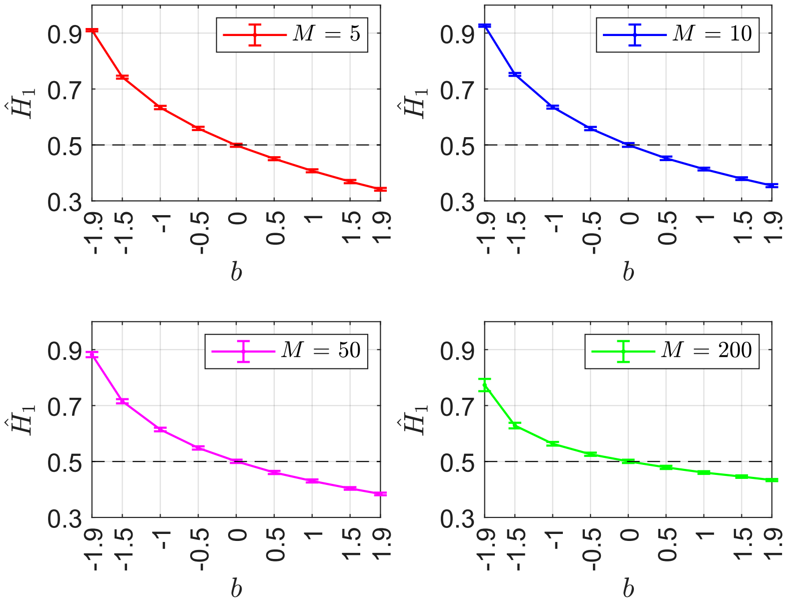

Figure 5.5: Generalized Hurst Exponent for $q = 1$ computed on simulated time series varying both b and M. The scaling exponent has been extracted using $\tau \in [1{,}200]$. It is evident that smaller values of b imply larger values of $\hat{H}_1$. The single error bar represents the standard deviation of the estimate over the set of 100 simulations performed for each choice of the parameters.

The results obtained show a clear relation between the parameter b and the estimated Hurst exponent $\hat{H}_1$: when $b \leq 0$ the time series is persistent, while $b \geq 0$ the time series is anti-persistent.

It is also interesting to observe that when M is increased up to value of order 10^4 time steps, the Hurst exponent tends to 0.5 regardless of the initial b value. As an example the case for $b = -1.9$, which is the one with the most persistence autocorrelation of the volatility, is shown.

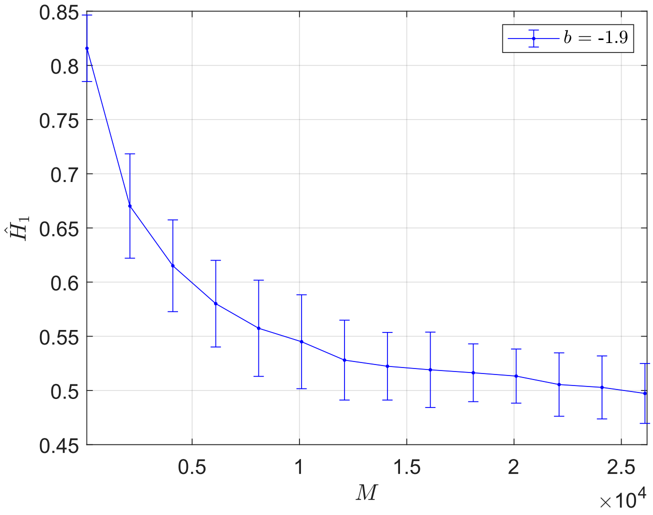

Figure 5.6: Generalized Hurst Exponent for $q = 1$ computed on simulated time series fixing $b = -1.9$ and increasing M. The scaling exponent has been extracted using $\tau \in [1{,}200]$. Increasing the value of M, the time series' scaling exponent tends to the one of a simple Brownian motion. The single error bar represents the standard deviation of the estimate over the set of 100 simulations performed for each choice of the parameters.

The behavior can be ascribed to the fact that increasing M the time series become equally smooth, regardless of b, while the diffusion properties are still determined by coefficient of the potential. As an example three simulation with $M = 2000$ and $b \in [-1.9{,}0{,}2]$ are displayed on the same plot:

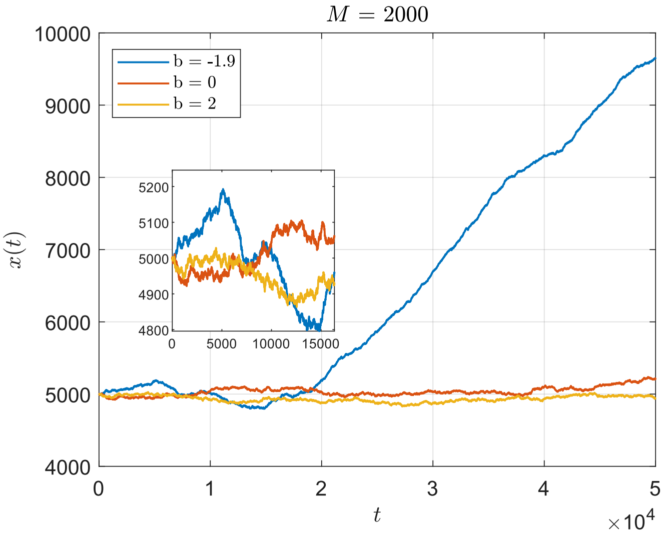

Figure 5.7: PUCK model time series simulation for $b \in [-1.9{,}0{,}2]$ and $M = 2000$. The value of M affects the properties of the long-term memory and therefore the "roughness" of the time series, while b still determines the diffusion behaviour.

5.3.2 Case of Random b(t)

In the case of $b(t)$ following Equation (5.8), it is employed again the RNSGHE method (3.5.3) to study the scaling properties of the absolute value of the price variation $|x(t+\tau) - x(t)|$ of the time series simulated. $f(t)$ is chosen again to be a gaussian noise with $\mu = 0$, $\sigma = 1$ and M is set equal to 2 just to make the observations comparable with the one of the Takayasu et al. paper [13]. The parameters of the model for $b(t)$ are chosen to be $c_0 = 0.0015$ and $G = 0.000784$ in order to let the model reproduce, at least for long simulations, some of the properties of real financial time series (5.2.3).

Again the scaling relation in Equation (5.9) is assumed, and for the estimate of H_q and B the following moments q and aggregation times τ are chosen:

- $q \in$ [0.02,1] to have a robust measure of multiscaling when having fat tails in the distribution of price variations;
- $\tau \in$ [1,200], [1,19] time steps, as the autocorrelation of the volatility $v(t)^2$ again does not follow the power law decay observable in real time series and indeed the ACSR method gives results with a great variance between similar time series.

100 simulations of length $T = 10^6$ are performed and for each of them is verified that the properties shown in Figure (5.3) are confirmed. The 95% confidence interval bounds of the distribution of the estimates of the multiscaling proxy $\hat{B}$ are therefore reported:

$\boldsymbol{\tau}$	[1,19]	[1,200]
$\hat{\boldsymbol{B}}$	(-0.1043,-0.0022) $\cdot 10^{-11}$	(-0.2476, -0.2221) $\cdot 10^{-13}$

It is clear, and confirmed by the multiscaling test, that the time series do not show multiscaling properties and thus the Hurst exponent for $q = 1$ is enough to describe the scaling of the model also in this random $b(t)$ case. The 95% confidence interval bounds of the distribution of the estimates of H_1 are reported:

$\boldsymbol{\tau}$	[1,19]	[1,200]
$\hat{\boldsymbol{H}}_{\mathbf{1}}$	(0.5036, 0.5207)	(0.5034, 0.5145)

The values obtained show that when $b(t)$ randomly fluctuates between -2 and 2 the Hurst exponent is just slightly above 0.5, which means that the scaling properties of the time series are practically equivalent to the one of a simple Brownian motion. In a sense one could approximately relate the Hurst exponent computed over the whole time series with an average of $b(t)$ over time, as long as M is not too big.

In fact, if one modifies the parameters of the model of $b(t)$ (c_0 and the average of the noise) in order to obtain a clear transition in the behaviour of this coefficient, the measure of the Hurst exponent changes accordingly:

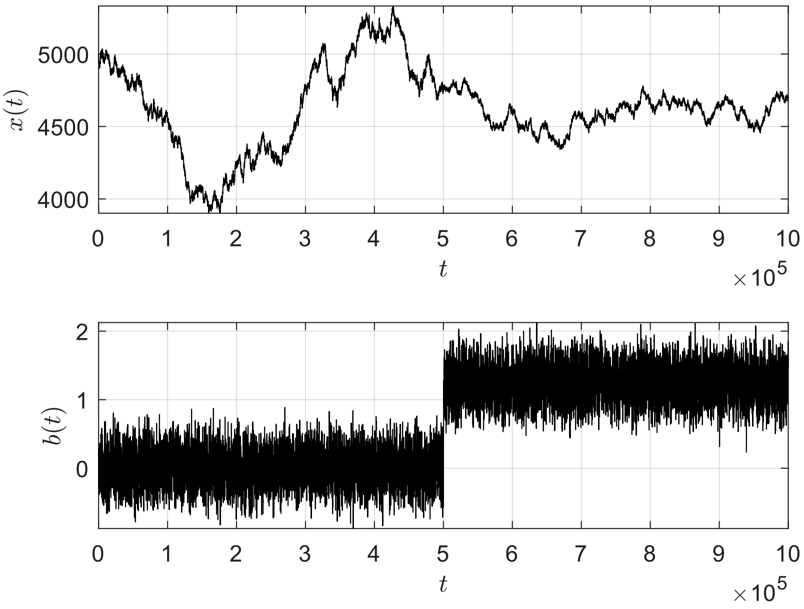

Figure 5.8: Path simulated with $M = 5$ and $b(t)$ following Equation (5.8) with $c_0 = 0.008$ and $G = 0.000784$. The initial position is $x(1) = 5000$ and the path is $T = 10^6$ steps long. From $t = 5 \cdot 10^5$ the average of the noise of the process followed by $b(t)$ has been artificially moved from 0 to 0.01.

For the depicted time series in fact one obtains ($q \in [0.02{,}1]$, $\tau \in [1{,}200]$):

$$\hat{H}_1 = 0.4514 \pm 0.0005$$

which is significantly lower than 0.5 as the time series of $b(t)$, from $T = 5 \cdot 10^5$, oscillates around a value greater than zero. The results are in agreement with the previous analysis realized at constant b (Fig. 5.5). In particular the time series, simulated setting $M = 5$, shows an average $\langle b(t) \rangle_t = 0.6250$, and if one compares this result with the top-left plot in Figure 5.5, the value of $\hat{H}_1$ obtained is perfectly consistent.

It is therefore clear that the Hurst exponent computed over the whole time series can be directly related to $b(t)$ or to be more precise, to its time average.

5.3.3 Comparison between b(t) and the Dynamical Hurst Exponent

First of all, the *w*GHE method is applied to compute the scaling of the price variation $|x(t+\tau) - x(t)|$, with $q \in [0.02{,}1]$, $\tau_{min} = 1$, $\tau_{max} = 19$ as in [34, 17, 52], to a simulated time series of length $T = 10^5$ with known random $b(t)$ following Equation (5.8) and $M = 2$. The idea is to compare the real $b(t)$ with the dynamical Hurst exponent $H_q^{\theta}(t)$ computed using a rolling time window of $\Delta t = 10^3$ time steps and setting $\Delta t = \theta$ as in [17].

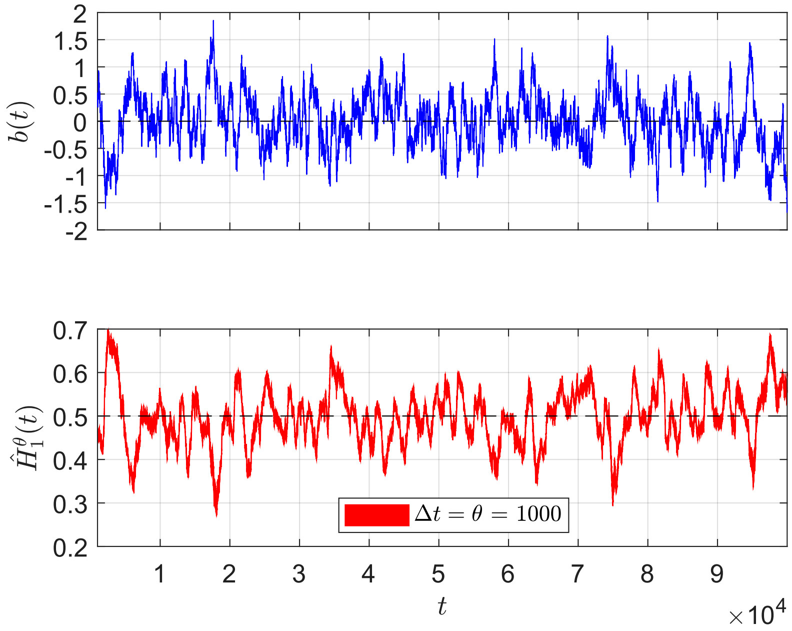

Figure 5.9: Result of $\hat{H}_1^\theta(t)$ on a time series simulated using the PUCK model with the random $b(t)$ ($c_0 = 0.0015$ and $G = 0.000784$) depicted and $M = 2$. The wGHE method is performed with $q \in [0.02,1]$, $\tau_{min} = 1$, $\tau_{max} = 19$ and $\Delta t = \theta = 1000$.

The results seem to have a close profile, but in order to quantify their similarity the Spearman and the Kendall correlation coefficient between the two time series are computed.

$\boldsymbol{\Delta t}$	**1000**
$\boldsymbol{\rho_s}$	-0.5575
$\boldsymbol{\rho_k}$	-0.3915

Table 5.1: Spearman and Kendall correlation coefficients between the time series of $\hat{H}_1^\theta(t)$ and the real $b(t)$ used in the simulation of the time series. All the values are statistically significant at 1 % level.

The correlation between the Hurst exponent measure and the real $b(t)$ time series are negative as expected. What is very interesting is that the correlation is quite strong, and therefore it is clear that one could use the Generalized Hurst exponent dynamically to extract from the time series an information equivalent to $b(t)$.

5.4 Comparison between b(t) and the Hurst Exponent on Data

To devise the value of $b(t)$ from data, it is needed to assume that this quantity remains constant within a certain time window. Indeed from equations (5.1) a simple linear model for $b(t)$ can be defined:

$$x(t+1) - x(t) = -\frac{b(t)}{M-1}(x(t) - x_M(t)) + f(t) \tag{5.10}$$

Therefore, in a selected time window of width Δt, from the angular coefficient of this linear regression the value of $b(t)$ can be extracted [13]. Obviously, assuming that for real data this coefficient remains constant over Δt is a simplification that leads to a value of $b(t)$ certainly subjected to an error. Ideally one would like to compute it from a smaller time window, in order to get a quantity closer to the real value, even if having less points to perform the regression causes inevitably an higher standard error on the angular coefficient.

5.4.1 An Idea to Estimate M

To compute $b(t)$ the value of M is also needed and one can try to extract it from the data or fixing it as in [13, 74]. If one defines: $v(t) \equiv x(t) - x(t-1)$; when b is constant the model can be viewed as an auto-regressive process [13]:

$$v(t+1) = -\frac{b}{2}\sum_{k=1}^{M-1} \omega_k v(t-k+1) + f(t) \tag{5.11}$$

where the weight function ω_k is given by:

$$\omega_k = \frac{2(M-k)}{M(M-1)} \quad , \quad \sum_{k=1}^{M-1} \omega_k = 1 \tag{5.12}$$

Assuming $b(t) = \text{const}$ in a given time window Δt, one can simply exploit the fact that the return time series is described by equations (5.11) and (5.12) to devise the coefficients of the AR model.

In order to estimate the parameters α_k of an AR(n) model (5.11) of order n (in our case $n = M - 1$) the so called Yule-Walker equations [75] can be employed:

$$\begin{cases} \gamma_m = \sum_{k=1}^{n} \alpha_k \gamma_{m-k} + \sigma^2 \delta_{m,0} & m = 0, ..., n \\ \gamma_m = \mathbb{E}[v(t)v(t-m)] \end{cases} \tag{5.13}$$

$$\begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_n \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_{-1} & \gamma_{-2} & \cdots \\ \gamma_1 & \gamma_0 & \gamma_{-1} & \cdots \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \\ \gamma_{n-1} & \gamma_{n-2} & \gamma_{n-3} & \cdots \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{bmatrix} \tag{5.14}$$

$$\gamma_0 = \sum_{k=1}^{n} \alpha_k \gamma_{-k} + \sigma^2,$$

from which the $\{\alpha_k\}$ and the variance of the noise can be extracted after having estimated empirically the autocovariance of the process. From a computational point of view the Levinson-Durbin recursion on the biased estimate of the sample autocorrelation sequence [76] is usually applied to solve these equations and therefore to finally compute the parameters. Knowing from equation (5.12) the theoretical expression of the coefficients of the model, the idea would be to use this information to devise M from the data. Indeed, for $k \geq M$ all the α_k become equal to zero and this behaviour should be observable directly from the Yule-Walker estimate of the coefficients.

The idea one could apply is exactly the same but on a small time windows of Δt time steps with the objective to find a similar behaviour even if the portion of the time series analyzed is very short. The advantage indeed would be that one could estimate M without knowing b, just by computing the breakpoint from which the coefficients of the model become zero.

Unfortunately, the procedure described showed promising results only for simulated time series with small values of M and $b = const$ far from zero. Real data are indeed characterized by $b(t)$ oscillating around zero [72, 65, 13, 74] and therefore M becomes very difficult to be extracted.

For the subsequent analysis it is therefore preferred to fix M artificially and to consequently extract $b(t)$. Indeed, if one extracts $b(t)$ using Equation (5.10):

$$b(t) = -\frac{(x(t+1) - x(t))}{(x(t) - x_M(t))} \cdot (M - 1) \tag{5.15}$$

different M values just cause a rescaling of the results.

5.4.2 Hurst Exponent vs Time Average of b(t)

Having observed in Figure (5.8) that it's reasonable to think about the existence of a relation between the time average of $\langle b(t) \rangle_t = \frac{1}{T} \sum_{t=1}^{T} b(t)$ and $\hat{H}_1$, these quantities are measured on some real financial data.

From the set of stocks comprised in the S&P 500 index, as of 30 Nov 2022, 187 stocks whose daily closing prices is recorded from 02-Jan-1990 to 30-Nov-2022 are selected. Daily prices are chosen as the methods to devise the scaling exponents are tested and built for this kind of data [18, 19]. The coefficient $b(t)$ is extracted using Equation (5.15) employing a rolling time window of $\Delta t = 1000$ days and fixing $M = 2$. As an example the time series for $b(t)$ computed on Apple Inc stock is shown:

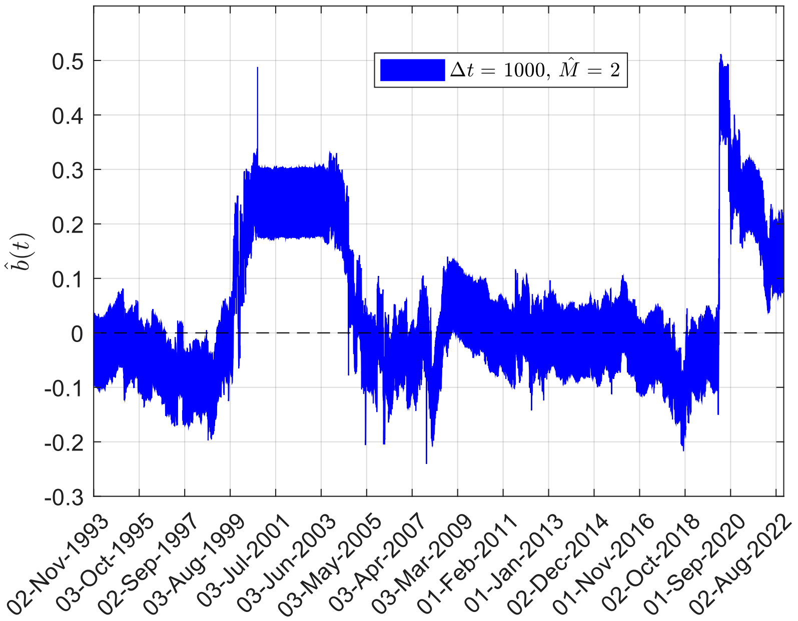

Figure 5.10: Result of the regression procedure on Apple Inc stock daily closing price for $\Delta t = 1000$ days and $M = 2$. The width of each line is equal to two standard errors of the angular coefficient as determined by the least squares linear fit.

It is clear that the measure of $b(t)$, even using large time windows, gives large standard errors which make this estimate unpractical on real data. Moreover, the value of M has been chosen a priori without knowing the actual one which describes data.

The Hurst exponent H_1 is instead measured on the absolute log-returns time series with the methodology described in section 3.5.3, setting $q \in [0.02,1]$, $\tau \in [1, \tau_{max}]$ and computing τ_{max} with the so called ACSR method.

Just to make a comparison between the two measures a scatter plot over all the data set is realized and the Spearman's correlation coefficient between the two is computed.

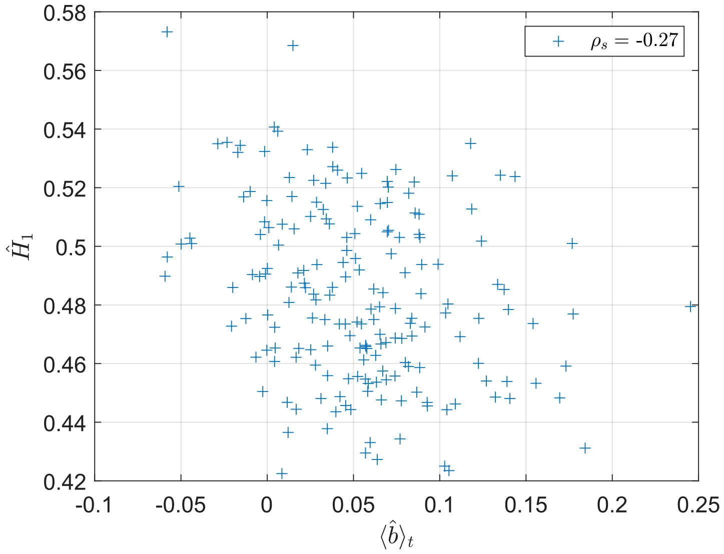

Figure 5.11: Scatter plot of the Hurst exponent for $q = 1$ computed over the whole log-returns time series of 187 stocks against the time average of $b(t)$ computed for each stock. The correlation value is statistical significant at 1 %.

The obtained value shows, as expected, a negative correlation that is also confirmed by the Kendall's correlation coefficient which is slightly lower ($\rho_k = -0.15$). One must however underline that $b(t)$ is subjected to large errors that are also amplified computing a time average, despite this a non trivial negative correlation exists between the two measures.

What should come clear is that the estimate of the coefficient of the potential of the PUCK model is not practical and brings with it many issues related both to the regression technique and to the estimate of M. Nevertheless, having found an approximate interpretation of the parameter $b(t)$ in terms of H_q it is therefore preferred to use this well established measure to determine if a particular time series is in a mean reverting state or in a trend follower state. One in fact can apply the Generalized Hurst exponent method in two different fashions: using the whole time series and therefore determining the scaling properties over the entire analyzed period; working in rolling time windows to analyze the evolution of the scaling properties.

The first approach is the one employed to extract the values shown in figure 5.11. The application of the second approach will instead be shown in the next section.

5.4.3 Comparison between b(t) and the Dynamical Hurst Exponent

To complete the investigation, the idea is thus to apply the previously described technique to extract from a real time series a dynamical information that, as showed previously, can somehow replace the unpractical $b(t)$.

The wGHE method is performed with $q \in [0.02,1]$, $\tau_{min} = 1$, $\tau_{max} = 19$ as in [34, 17, 52] on the time series of the Apple Inc stock daily closing price from 02-Jan-1990 to 30-Nov-2022. It is chosen to work on daily prices as the employed methodologies to study the scaling exponents are well-established on daily data [19, 17].

The main focus is placed on $H_q^\theta(t)$ with $q = 1$ because it is directly comparable with $b(t)$. Following the prescriptions in [17] $\Delta t = \theta$ is chosen and $\Delta t = 500$,1000 days are selected for the length of the rolling time windows.

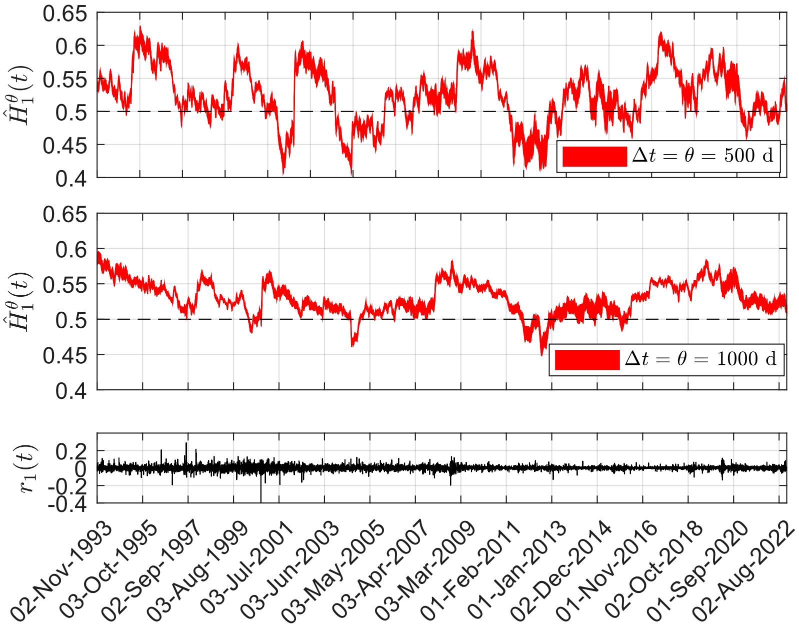

Figure 5.12: Result of $\hat{H}_1^\theta(t)$ on Apple Inc stock daily closing price (from 02-Jan-1990 to 30-Nov-2022) for $\Delta t = 500, 1000$ days. On the third plot the daily log returns time series is displayed.

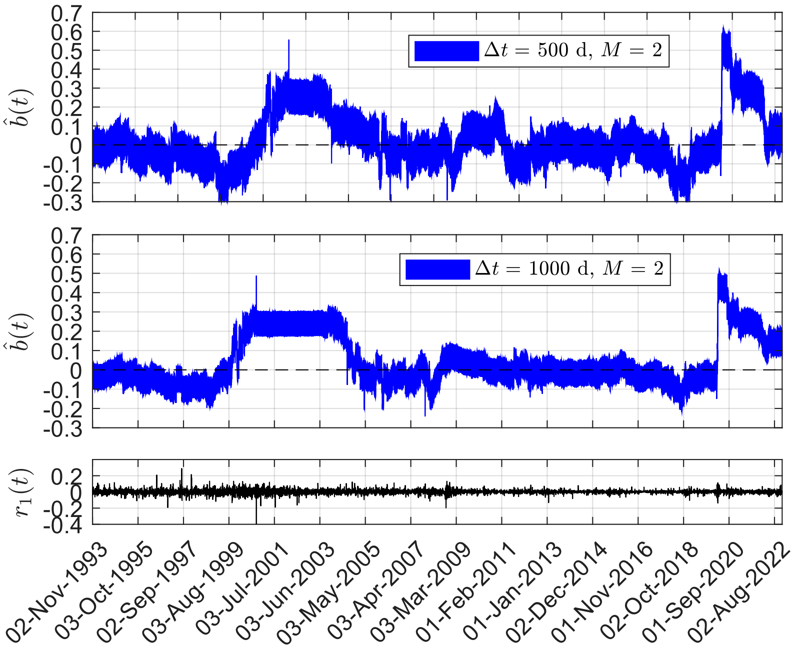

Figure 5.13: Result of $\hat{b}(t)$ on Apple Inc stock daily closing price (from 02-Jan-1990 to 30-Nov-2022) for $\Delta t = 500, 1000$ days and $M = 2$. On the third plot the daily log returns time series is displayed.

$\boldsymbol{\Delta t}$	**500 d**	**1000 d**
$\boldsymbol{\rho_s}$	-0.1527	-0.1974
$\boldsymbol{\rho_k}$	-0.1052	-0.1227

Table 5.2: Spearman and Kendall correlation coefficients between the time series of $\hat{H}_1^{\theta}(t)$ and $\hat{b}(t)$ obtained for Apple Inc stock daily closing price (from 02-Jan-1990 to 30-Nov-2022) for various time windows sizes. All the values are statistically significant at 1 % level.

The obtained results show again a negative correlation between the two measures also when the Hurst exponent is computed dynamically. The correlation is not very high due to the fact that in real data the two measures oscillate respectively around 0 and 0.5, and the real behaviour around these values is often hidden by the errors on the estimates caused by the reduced time window employed. Moreover, $b(t)$ is subjected to large errors even when using time window of $\Delta t = 1000$ days and indeed when comparing just the Hurst exponent with the real time series of $b(t)$ the correlation is stronger.

In fact, the estimate of $\hat{H}_1^{\theta}(t)$ does not carry the ambiguity of the choice of M and the great error caused by the regression for $b(t)$. Moreover, the generalized Hurst exponent extracts the full scaling spectrum of H_q and, given the multiscaling nature of financial data [34, 18], this set of information can provide more insights about the analyzed time series [17, 52].

The message that is intended to be sent is therefore that, even if the PUCK model is very intuitive and opens the possibility to describe particular states of the market, which can be directly interpreted through the behavior of the traders, when employing it as a time series analysis tool its strength weakens, as extracting its parameters from data can be challenging and often unpractical. It is therefore proposed to replace this procedure with the well established Generalized Hurst Exponent analysis (both over rolling time windows or over the whole time series), which gives statistically significant results that can be also related to the coefficient of the potential function present in the PUCK model.

Chapter 6

Conclusion

The aim of this project has been to revise some important results retrieved by studying financial markets with tools coming from physics and complexity science in general: scaling analysis, measures of correlations, physics inspired modeling. All the topics have been presented theoretically and most of the empirical results described have been confirmed on a real data set of all the stocks comprised in the S&P 500 index. An application of some of the scaling analysis methodologies has then been originally employed to study particular properties of a novel random walk model of market price by Takayasu et al. [65].

In the first chapter the main empirical *stylized facts* of financial time series have been presented: heavy tails, aggregational Gaussianity, absence of autocorrelations and volatility clustering. Together these features define the main properties shared by financial data (particularly time series), which are fundamental both to understand the validity of theoretical models and to describe some features of the analyzed assets. In the present case, the tail exponents for the distribution of the daily returns and the exponents for the decay of the volatility autocorrelation have been further examined, showing their values over different sectors and their practical meaning which could lead to useful applications in the context of risk management.

In the second chapter the concept of "Scaling properties" has been presented and it has been illustrated how it naturally arises within the context of self-similarity, scale invariance and fractals. This framework has been expanded to stochastic processes which display statistical properties of self-similarity and the Hurst exponent as a measure related to these behaviours has been presented. The discussion has therefore been enlarged to explain theoretically the concept of multifractality from which a general scaling rule has been introduced. This scaling law is the hallmark of multifractality and the starting point for every empirical data analysis in which one wants to measure multiscaling properties. Two famous stochastic models for returns which include multifractality have then been presented and the chapter has been completed with the description of various important techniques employed to measure the multiscaling properties of financial time series: the generalized Hurst exponent method, its dynamical and weighted version and its refinement which is more reliable and effective from a statistical point of view. The application of these methods has been thus shown on the analyzed data set and it has been confirmed that financial time series exhibit multiscaling properties. Moreover, through the wide literature review different practical applications of these devised measures have also been presented. It is important to underline the fact that the described framework can be naturally extended to many other complex system, indeed the methodologies are quite general and are able to extract the statistical properties of the underlying processes directly from the data.

In the third chapter the usefulness of the study of the correlations between financial assets has been demonstrated. By studying the dynamical evolution of the weighted Pearson correlation coefficient it has been highlighted how various moments of the markets can affect the interdependence structure of the stocks, while employing the network science approach it has been evidenced the natural clustering structure that arises simply from the topology of the correlation graph. It is important to point out that these techniques, particularly the information filtering tool employed (i.e. Minimum Spanning Tree), are very useful to study any complex system as it is often strongly necessary to reduce the huge amount of noisy information available in order to extract some interesting insights about the framework under study.

The last chapter has been devoted to a brief presentation of physics inspired models in the context of finance, with the aim to show another interesting approach that employs ideas from the physical field to study financial systems. The ideas behind two interesting agent based models have been presented, with the aim to show a concrete application of this interdisciplinary approach to finance as well to introduce an important framework in which the PUCK model has been employed. The remaining part of the chapter has been indeed dedicated to the presentation of the PUCK model, a stochastic model for market price by Takayasu et al. [13], that has been employed both to describe "chartists" agents in the Alfi-Cristelli-Pietronero-Zaccaria agent based model and as a time series analysis tool. The second application is the one on which the main focus has been placed in the chapter. After having described the model, the scaling properties of time series simulated with its equations have been studied and a non trivial novel relation between the Hurst exponent and the parameter $b(t)$ of the model has been devised. Moreover, the model has been found to produce uniscaling time series and a non negligible negative correlation has been retrieved both between the temporal average of $b(t)$ and the Hurst exponent computed over the whole time series, and between the time series for $b(t)$ and the time series for the weighted generalized Hurst exponent $H_q^\theta(t)$. This last dependence has been retrieved both on time series with known artificial $b(t)$ and on real time series. The conclusion is that the study of the scaling properties of time series through the generalized Hurst exponent technique is more reliable, as it is defined by statistically well founded methods not present for $b(t)$, and the results obtained can also be interpreted in terms of the correlation with the parameter $b(t)$.

In this work only a small part of the techniques, tools and models, coming from physics and complexity science and employed to study financial markets, have been presented and many more chapters would have been necessary to cover the majority of them. It has been instead preferred to focus mainly on the presented topics as they can be very useful both to researchers and practitioners. Furthermore, the unveiled framework encloses various techniques which are helpful in many different fields and understanding the theoretical intuitions behind them and how to apply these ideas can provide important instruments to analyze complex systems from different and novel perspectives.

Appendix A

Codes

A.1 Data cleaning code

```
function [indices]= data_cleaning(Prices,dt)

clusters = zeros(size(Prices,1)-1,size(Prices,2));

for i=1:size(Prices,2)
    price = Prices(~isnan(Prices(:,i)),i);
    d = (diff(price) == 0);
    cluster_size = 1;
    for j=1:length(d)-1
        if d(j) == 1 && d(j+1) == 1
            cluster_size = cluster_size + 1;
            clusters(j,i) = cluster_size;
        else
            cluster_size = 0;
        end
    end
end

indices = find(max(clusters) >= dt); %indices of problematic stocks with constant clusters longer than dt
end
```

The shown code takes in input the matrix `Prices` whose columns are time series of different length, and an integer `dt`. If the time series analyzed includes at least a cluster of constant adjacent values of length greater or equal than `dt`, the time series is considered to be problematic.

A.2 Data selection code

```
function [t_star] = selection_time(Prices,mktcap)

T = size(Prices,1);
N = size(Prices,2);
W = zeros(T,1);
pts = zeros(T,2);
n0 = 0;

w = mktcap/sum(mktcap);

j=1;
for i=1:T
    stock_idx = find(~isnan(Prices(i,:)));
    n = length(stock_idx); %number of stocks that co-exist
    t=i; %starting time of the subset
    if n ~= n0
        pts(j,1) = t;
        pts(j,2) = n;
        W(j) = sum(w(stock_idx));
        n0 =n;
        j = j+1;
    end
    if n == N
        break
    end
end

pts(j:end,:) = [];
W(j:end) = [];

I = ((T-pts(:,1))/(T-1)).*((pts(:,2)-1)/(N-1)).*W;
[I_max,k] = max(I);
t_star = pts(k,1);
n_star = pts(k,2);
end
```

The shown code takes in input the matrix `Prices` whose columns are time series of different length, and a vector `mktcap` of the market capitalization of each stock. It performs a `for` loop to compute iteratively the information function defined in Equation (4.3). The loop stops when the number of selected columns of the `Prices` matrix is equal to `N`. The procedure returns the t index for which $I(t)$ is maximum.

Appendix B

Scaling PUCK model

B.1 Scaling at constant b with $\tau_{max} = 19$

The previously obtained results (see subsection 5.3.1) are confirmed using a different aggregation time interval: $\tau \in [1{,}19]$ is chosen to check if the previous results where an artifact of the selected time scales [35, 34]. As before 100 simulations of $T = 5{\cdot}10^4$ time steps are performed using all the possible 36 combinations of the following set of parameters: $b = [-1.9, -1.5, -1, -0.5{,}0{,}0.5{,}1{,}1.5{,}1.9]$; $M = [5{,}10{,}50{,}200]$.

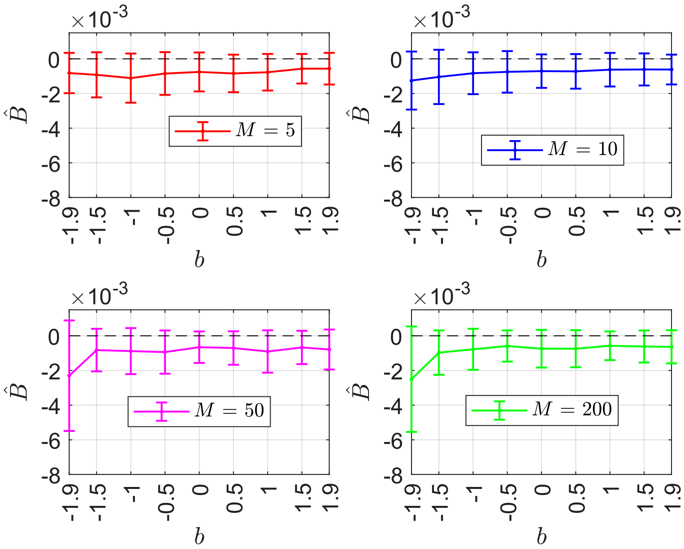

Figure B.1: Estimate of the multiscaling proxy (Equation 3.57) varying both b and M. The RNSGHE method has been applied using $\tau \in [1{,}19]$, $q \in [0.02{,}1]$. For larger M both the average and the std of $\hat{B}$ become larger due to the higher persistence of the volatility autocorrelation. The single error bar represents the standard deviation of the estimate over the set 100 simulation performed for each choice of the parameters.

The multiscaling proxy properties are practically the same as the one displayed in figure 5.4, and no major differences can be observed. It is confirmed the role of the autocorrelation which reduces the average $\hat{B}$ and increases the standard deviation of the results.

Regarding the multiscaling test, the $3597/3600 = 99.92\%$ of the simulations performed failed both the F-test and the t-tests (see table 3.1) while the remaining 21 rejected the null hypothesis only for the F-test. It is therefore confirmed that at constant b the process is not multiscaling. The scaling properties of the process for $q = 1$ are shown:

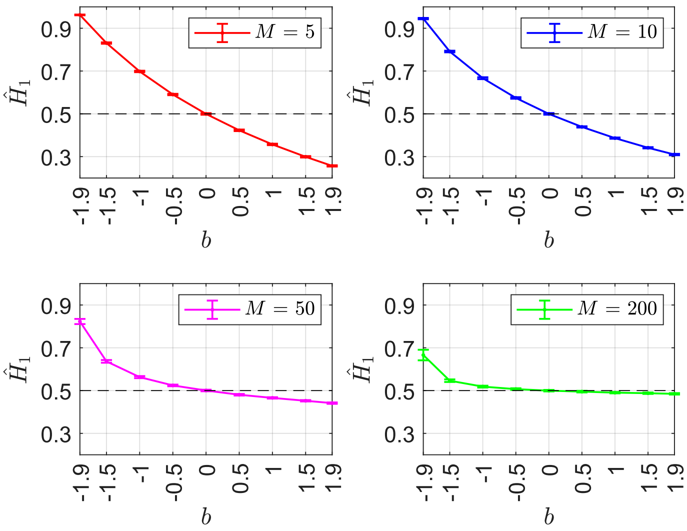

Figure B.2: Generalized Hurst Exponent for $q = 1$ computed on simulated time series varying both b and M. The scaling exponent has been extracted using $\tau \in [1,19]$. It is evident that smaller values of b imply larger values of $\hat{H}_1$. The single error bar represents the standard deviation of the estimate over the set 100 simulation performed for each choice of the parameters.

The relation between $\hat{H}_1$ and b is qualitatively the same as the one depicted in figure 5.5. For larger values of M, the behaviour is less evident when b is larger. This is of course caused by the choice of $\tau_{max} = 19$, that for values of $M \gtrsim 50$ becomes too small to detect the long-term memory of the time series at larger time scales.

Bibliography

[1] Tomaso Aste and Tiziana Di Matteo. «Introduction to complex and econophysics systems: A navigation map». In: *Complex physical, biophysical and econophysical systems*. World Scientific, 2010, pp. 1–35 (cit. on pp. 1, 10, 11, 17).

[2] Ryszard Kutner, Marcel Ausloos, Dariusz Grech, Tiziana Di Matteo, Christophe Schinckus, and H Eugene Stanley. *Econophysics and sociophysics: Their milestones & challenges*. 2019 (cit. on pp. 1–3, 5).

[3] RJ Buonocore, N Musmeci, T Aste, and T Di Matteo. «Two different flavours of complexity in financial data». In: *The European Physical Journal Special Topics* 225 (2016), pp. 3105–3113 (cit. on pp. 1, 3).

[4] Rosario Nunzio Mantegna. «Lévy walks and enhanced diffusion in Milan stock exchange». In: *Physica A: Statistical Mechanics and its Applications* 179.2 (1991), pp. 232–242 (cit. on pp. 1, 2, 10).

[5] H Eugene Stanley and Rosario N Mantegna. *An introduction to econophysics*. Cambridge University Press, Cambridge, 2000 (cit. on pp. 1, 2).

[6] Ettore Majorana. «Il valore delle leggi statistiche nella fisica e nelle scienze sociali». In: *Scientia* 36.71 (1942) (cit. on p. 1).

[7] H Eugene Stanley et al. «Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics». In: *Physica A: Statistical Mechanics and its Applications* 224.1-2 (1996), pp. 302–321 (cit. on p. 2).

[8] Anirban Chakraborti, Ioane Muni Toke, Marco Patriarca, and Frédéric Abergel. «Econophysics review: I. Empirical facts». In: *Quantitative Finance* 11.7 (2011), pp. 991–1012 (cit. on pp. 2, 5, 6, 8–10, 13, 17, 19–21, 23, 27, 40, 55, 73, 79, 81).

[9] Tiziana Di Matteo. *Lecture notes for the course of Econophysics*. King's College London, 2023 (cit. on pp. 2, 3, 9, 10, 13, 55, 64, 66, 68, 71).

[10] Anirban Chakraborti, Ioane Muni Toke, Marco Patriarca, and Frédéric Abergel. «Econophysics review: II. Agent-based models». In: *Quantitative Finance* 11.7 (2011), pp. 1013–1041 (cit. on pp. 2, 73, 74).

[11] Rama Cont. «Empirical properties of asset returns: stylized facts and statistical issues». In: *Quantitative finance* 1.2 (2001), p. 223 (cit. on pp. 2, 5, 6, 8–11, 13, 17, 19–21, 23, 24, 27, 40, 79).

[12] Emmanuel Bacry, Jean Delour, and Jean-François Muzy. «Modelling financial time series using multifractal random walks». In: *Physica A: statistical mechanics and its applications* 299.1-2 (2001), pp. 84–92 (cit. on pp. 5, 10, 37–39, 43, 53, 73).

[13] Misako Takayasu, Kota Watanabe, Takayuki Mizuno, and Hideki Takayasu. «Theoretical base of the PUCK-model with application to foreign exchange markets». In: *Econophysics Approaches to Large-Scale Business Data and Financial Crisis*. Springer. 2010, pp. 79–98 (cit. on pp. 5, 73–75, 77, 79–81, 84, 87, 88, 96).

[14] Jiwon Ma Will Kenton Michael J Boyle. *S&P 500 Index: What It's for and Why It's Important in Investing*. 2023. URL: `https://www.investopedia.com/terms/s/sp500.asp` (visited on 05/24/2023) (cit. on pp. 6, 8).

[15] John Y Campbell, Andrew W Lo, A Craig MacKinlay, and Robert F Whitelaw. «The econometrics of financial markets». In: *Macroeconomic Dynamics* 2.4 (1998), pp. 559–562 (cit. on p. 6).

[16] MSCI Inc. *The Global Industry Classification Standard (GICS ®)*. 2023. URL: `https://www.msci.com/our-solutions/indexes/gics` (visited on 05/26/2023) (cit. on pp. 7, 13).

[17] Ioannis P Antoniades, Giuseppe Brandi, L Magafas, and Tiziana Di Matteo. «The use of scaling properties to detect relevant changes in financial time series: A new visual warning tool». In: *Physica A: Statistical Mechanics and its Applications* 565 (2021), p. 125561 (cit. on pp. 9, 40, 41, 46, 49, 76, 85, 91, 92).

[18] Giuseppe Brandi and Tiziana Di Matteo. «On the statistics of scaling exponents and the multiscaling value at risk». In: *The European Journal of Finance* 28.13-15 (2022), pp. 1361–1382 (cit. on pp. 9, 37, 39–45, 47, 49, 52, 76, 81, 88, 92).

[19] Giuseppe Brandi and T Di Matteo. «Multiscaling and rough volatility: An empirical investigation». In: *International Review of Financial Analysis* 84 (2022), p. 102324 (cit. on pp. 9, 35, 39, 40, 42–45, 88, 91).

[20] Benoit B Mandelbrot and Benoit B Mandelbrot. *The variation of certain speculative prices*. Springer, 1997 (cit. on pp. 9, 20).

[21] Anand Banerjee, Victor M Yakovenko, and Tiziana Di Matteo. «A study of the personal income distribution in Australia». In: *Physica A: statistical mechanics and its applications* 370.1 (2006), pp. 54–59 (cit. on p. 9).

[22] Benoit B Mandelbrot, Adlai J Fisher, and Laurent E Calvet. «A multifractal model of asset returns». In: (1997) (cit. on pp. 10, 33–40).

[23] Parameswaran Gopikrishnan, Martin Meyer, LA Nunes Amaral, and H Eugene Stanley. «Inverse cubic law for the distribution of stock price variations». In: *The European Physical Journal B-Condensed Matter and Complex Systems* 3.2 (1998), pp. 139–140 (cit. on p. 11).

[24] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. «Power-law distributions in empirical data». In: *SIAM review* 51.4 (2009), pp. 661–703 (cit. on p. 11).

[25] William R Leo. *Techniques for nuclear and particle physics experiments: a how-to approach.* Springer Science & Business Media, 2012 (cit. on p. 13).

[26] Antonios Antypas, Phoebe Koundouri, and Nikolaos Kourogenis. «Aggregational Gaussianity and barely infinite variance in financial returns». In: *Journal of Empirical Finance* 20 (2013), pp. 102–108 (cit. on p. 17).

[27] L Kullmann, J Töyli, J Kertesz, A Kanto, and K Kaski. «Characteristic times in stock market indices». In: *Physica A: Statistical Mechanics and its Applications* 269.1 (1999), pp. 98–110 (cit. on p. 17).

[28] RA Fisher. *Statistical Methods for Research Workers, ed 13, p 285.* 1958 (cit. on pp. 19, 27, 61, 64).

[29] Maurice George Kendall et al. «The advanced theory of statistics.» In: *The advanced theory of statistics.* 2nd Ed (1946) (cit. on pp. 19, 61, 64).

[30] Adrian Pagan. «The econometrics of financial markets». In: *Journal of empirical finance* 3.1 (1996), pp. 15–102 (cit. on p. 19).

[31] Eugene F Fama. «Efficient capital markets: II». In: *The journal of finance* 46.5 (1991), pp. 1575–1617 (cit. on p. 19).

[32] Burton G Malkiel. «Efficient market hypothesis». In: *Finance* (1989), pp. 127–134 (cit. on p. 19).

[33] JP Bouchaud and M Potters. «Theorie des Risques Financiers (Alea-Saclay-Eyrolles, Paris, 1997)». In: *Google Scholar Theory of Financial Risks and Derivative Pricing (Cambridge University Press, Cambridge, 2003)* () (cit. on p. 21).

[34] Tiziana Di Matteo. «Multi-scaling in finance». In: *Quantitative finance* 7.1 (2007), pp. 21–36 (cit. on pp. 27, 33–35, 37, 39–41, 45, 81, 85, 91, 92, 101).

[35] Riccardo Junior Buonocore, Tomaso Aste, and Tiziana Di Matteo. «Measuring multiscaling in financial time-series». In: *Chaos, Solitons & Fractals* 88 (2016), pp. 38–47 (cit. on pp. 27, 39, 40, 45, 81, 82, 101).

[36] Michele Tumminello, Tiziana Di Matteo, Tomaso Aste, and Rosario N Mantegna. «Correlation based networks of equity returns sampled at different time horizons». In: *The European Physical Journal B* 55 (2007), pp. 209–217 (cit. on pp. 27, 66, 67, 71).

[37] Francesco Pozzi, Tiziana Di Matteo, and Tomaso Aste. «Exponential smoothing weighted correlations». In: *The European Physical Journal B* 85 (2012), pp. 1–21 (cit. on pp. 27, 61–65).

[38] Leo P Kadanoff. «Scaling and universality in statistical physics». In: *Physica A: Statistical Mechanics and its Applications* 163.1 (1990), pp. 1–14 (cit. on p. 29).

[39] Grigory Isaakovich Barenblatt. *Scaling.* Vol. 34. Cambridge University Press, 2003 (cit. on pp. 29, 31).

[40] Luca Dall'Asta. *Lecture Notes on Statistical Field Theory.* Politecnico di Torino, 2023 (cit. on p. 29).

[41] Benoit B Mandelbrot and Benoit B Mandelbrot. *The fractal geometry of nature*. Vol. 1. WH freeman New York, 1982 (cit. on p. 30).

[42] Bo Qian and Khaled Rasheed. «Hurst exponent and financial market predictability». In: *IASTED conference on Financial Engineering and Applications*. Proceedings of the IASTED International Conference Cambridge, MA. 2004, pp. 203–209 (cit. on p. 33).

[43] Harold Edwin Hurst. «Long-term storage capacity of reservoirs». In: *Transactions of the American society of civil engineers* 116.1 (1951), pp. 770–799 (cit. on p. 34).

[44] Harold Edwin Hurst. «Long term storage». In: *An experimental study* (1965) (cit. on p. 34).

[45] Uriel Frisch and Giorgio Parisi. «Fully developed turbulence and intermittency». In: *New York Academy of Sciences, Annals* 357 (1980), pp. 359–367 (cit. on p. 35).

[46] Benoit B Mandelbrot and Benoit B Mandelbrot. «Intermittent turbulence in self-similar cascades: divergence of high moments and dimension of the carrier». In: *Multifractals and 1/f Noise: Wild Self-Affinity in Physics (1963–1976)* (1999), pp. 317–357 (cit. on p. 35).

[47] L Pietronero and AP Siebesma. «Self-similarity of fluctuations in random multiplicative processes». In: *Physical review letters* 57.9 (1986), p. 1098 (cit. on p. 35).

[48] Benoit B Mandelbrot. «A multifractal walk down Wall Street». In: *Scientific American* 280.2 (1999), pp. 70–73 (cit. on pp. 37, 73).

[49] Zhi-Qiang Jiang, Wen-Jie Xie, Wei-Xing Zhou, and Didier Sornette. «Multifractal analysis of financial markets: a review». In: *Reports on Progress in Physics* 82.12 (2019), p. 125901 (cit. on pp. 37, 39, 45, 49).

[50] Emmanuel Bacry, Jean Delour, and Jean-François Muzy. «Multifractal random walk». In: *Physical Review E* 64.2 (2001), p. 026103 (cit. on pp. 37, 38).

[51] Riccardo J Buonocore, Tomaso Aste, and Tiziana Di Matteo. «Asymptotic scaling properties and estimation of the generalized Hurst exponents in financial data». In: *Physical Review E* 95.4 (2017), p. 042311 (cit. on p. 39).

[52] Raffaello Morales, Tiziana Di Matteo, Ruggero Gramatica, and Tomaso Aste. «Dynamical generalized Hurst exponent as a tool to monitor unstable periods in financial time series». In: *Physica A: statistical mechanics and its applications* 391.11 (2012), pp. 3180–3189 (cit. on pp. 40, 41, 46, 49, 50, 76, 85, 91, 92).

[53] RJ Buonocore, G Brandi, RN Mantegna, and T Di Matteo. «On the interplay between multiscaling and stock dependence». In: *Quantitative Finance* 20.1 (2020), pp. 133–145 (cit. on pp. 45, 52, 54, 55, 64).

[54] Ananth Ranganathan. «The levenberg-marquardt algorithm». In: *Tutoral on LM algorithm* 11.1 (2004), pp. 101–110 (cit. on p. 46).

[55] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. «Optimal detection of changepoints with a linear computational cost». In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598 (cit. on p. 46).

[56] Petr Kroha and Miroslav Skoula. «Hurst Exponent and Trading Signals Derived from Market Time Series.» In: *ICEIS (1)*. 2018, pp. 371–378 (cit. on p. 47).

[57] Sushil Bikhchandani and Sunil Sharma. «Herd behavior in financial markets». In: *IMF Staff papers* 47.3 (2000), pp. 279–310 (cit. on p. 65).

[58] Tomaso Aste, William Shaw, and Tiziana Di Matteo. «Correlation structure and dynamics in volatile markets». In: *New Journal of Physics* 12.8 (2010), p. 085009 (cit. on pp. 66–68, 70, 71).

[59] Michele Tumminello, Tomaso Aste, Tiziana Di Matteo, and Rosario N Mantegna. «A tool for filtering information in complex systems». In: *Proceedings of the National Academy of Sciences* 102.30 (2005), pp. 10421–10426 (cit. on pp. 66, 67, 70, 71).

[60] Rosario N Mantegna. «Hierarchical structure in financial markets». In: *The European Physical Journal B-Condensed Matter and Complex Systems* 11 (1999), pp. 193–197 (cit. on pp. 66–68, 71).

[61] Dirk G Baur and Thomas KJ McDermott. «Why is gold a safe haven?» In: *Journal of Behavioral and Experimental Finance* 10 (2016), pp. 63–71 (cit. on p. 67).

[62] Albert-László Barabási and Réka Albert. «Emergence of scaling in random networks». In: *science* 286.5439 (1999), pp. 509–512 (cit. on p. 67).

[63] Joseph B Kruskal. «On the shortest spanning subtree of a graph and the traveling salesman problem». In: *Proceedings of the American Mathematical society* 7.1 (1956), pp. 48–50 (cit. on p. 68).

[64] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms. 2001.* 2009 (cit. on p. 68).

[65] Misako Takayasu, Takayuki Mizuno, and Hideki Takayasu. «Theoretical analysis of potential forces in markets». In: *Physica A: Statistical Mechanics and its Applications* 383.1 (2007), pp. 115–119 (cit. on pp. 73, 75, 88, 95).

[66] V Alfi, Matthieu Cristelli, L Pietronero, and A Zaccaria. «Minimal agent based model for financial markets I: origin and self-organization of stylized facts». In: *The European Physical Journal B* 67 (2009), pp. 385–397 (cit. on pp. 73, 75).

[67] Valentina Alfi, Matthieu Cristelli, Luciano Pietronero, and A Zaccaria. «Minimal agent based model for financial markets II: statistical properties of the linear and multiplicative dynamics». In: *The European Physical Journal B* 67 (2009), pp. 399–417 (cit. on pp. 73, 75).

[68] Per Bak, Maya Paczuski, and Martin Shubik. «Price variations in a stock market with many agents». In: *Physica A: Statistical Mechanics and its Applications* 246.3-4 (1997), pp. 430–453 (cit. on p. 73).

[69] Stefano Galluccio, Jean-Philippe Bouchaud, and Marc Potters. «Rational decisions, random matrices and spin glasses». In: *Physica A: Statistical Mechanics and its Applications* 259.3-4 (1998), pp. 449–456 (cit. on pp. 73, 74).

[70] Thomas Lux and Michele Marchesi. «Scaling and criticality in a stochastic multi-agent model of a financial market». In: *Nature* 397.6719 (1999), pp. 498–500 (cit. on p. 74).

[71] Thomas Lux and Michele Marchesi. «Volatility clustering in financial markets: a microsimulation of interacting agents». In: *International journal of theoretical and applied finance* 3.04 (2000), pp. 675–702 (cit. on p. 74).

[72] Misako Takayasu, Takayuki Mizuno, Takaaki Ohnishi, and Hideki Takayasu. «Temporal characteristics of moving average of foreign exchange markets». In: *Practical fruits of econophysics: proceedings of the third Nikkei econophysics symposium*. Springer. 2006, pp. 29–32 (cit. on pp. 75, 88).

[73] Misako Takayasu and Hideki Takayasu. «Continuum limit and renormalization of market price dynamics based on PUCK model». In: *Progress of Theoretical Physics Supplement* 179 (2009), pp. 1–7 (cit. on pp. 76, 77).

[74] Arthur Matsuo Yamashita Rios de Sousa, Hideki Takayasu, and Misako Takayasu. «Random coefficient autoregressive processes and the PUCK model with fluctuating potential». In: *Journal of Statistical Mechanics: Theory and Experiment* 2019.1 (2019), p. 013403 (cit. on pp. 87, 88).

[75] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015 (cit. on p. 87).

[76] Monson H Hayes. *Statistical digital signal processing and modeling*. John Wiley & Sons, 1996 (cit. on p. 88).