Department of Economics and Finance

MSc in Economics and Finance - Finance major

Chair of Empirical Finance

Early detection of financial bubbles

Candidate: Vincenzo Guarracino

Student ID: 773751

Supervisor: Prof. Paolo Santucci De Magistris

Co-supervisor: Prof. Guido Traficante

Academic Year 2024/2025

Contents

Introduction

Bubbles have played a significant role throughout the history of finance. Since the "tulip bubble" of 1636–1637, they have had the power to shatter entire economies and, in exceptional cases, turn the tide of history. Bubbles can happen in any market, be it the stock market, credit, derivatives, crypto, and many others. In particular, stock market bubbles can be categorized into three types:

1. Endogenous, which arise from internal positive feedback loops within the market. Examples of this type are the Dotcom and the 2021–2022 bubbles.
2. Exogenous, such as the Covid crash of 2020, caused by the news of lockdown.
3. Mixed, such as the 2008 bubble, which originates from the subprime market and is transmitted later on in the major stock indexes.

Furthermore, a bubble can typically be decomposed into three different stages:

1. a random walk with positive drift, that represents the secular positive trend of the major stock markets;
2. an AR(1) process with $\phi_1 > 1$, which is the upward explosive phase;
3. a random walk with negative drift, which is the decline phase.

The topic of financial bubbles has always interested both economists and market investors alike. Economists have tried to explain the existence—or the impossibility—of bubbles through different theoretical frameworks, while analysts have spent great effort developing methods useful to detect stock market speculative regimes before their collapse.

In particular, the Efficient Market Hypothesis (EMH) claims that all available information is instantaneously reflected in prices, which therefore follow a martingale process. Within this paradigm, bubbles cannot be detected ex ante, since any deviation from fundamentals would be arbitraged away. On the opposite side, Behavioral Finance stresses the role of cognitive biases, herding, and reflexivity in generating self-reinforcing feedback loops that detach prices from intrinsic values. These distortions are not immediately corrected by rational arbitrage, and can therefore lead to the endogenous formation of bubbles. A third view, originating from Minsky, focuses on credit cycles and the progressive accumulation of financial fragility as a structural cause for booms and crashes.

This thesis pursues a twofold objective. First, to demonstrate that financial bubbles can be detected *ex ante*, i.e., before they reach their tipping point and crash. Second, to compare econometric and non-econometric approaches to early bubble detection, evaluating their relative accuracy, robustness, and interpretability. This comparative analysis allows to assess not only whether these methods work, but also how and why their performance varies across different market environments.
In particular, four distinct approaches will be considered:

1. Random Coefficient Autoregression (RCA), an econometric technique designed for real-time monitoring of regime shifts in time series.

2. The Log-Periodic Power Law Singularity (LPPLS) model, based on the idea of accelerating log-periodic structures before the crash.

3. Topological Data Analysis (TDA), a geometric approach based on the persistent homology of sliding windows over market prices.

4. Permutation Entropy (PE), a nonparametric, information-theoretic measure of randomness versus order in a time series.

These four approaches are tested on two empirical cases: the S&P 500 and Natural Gas (NYMEX) bubble of 2021–2022. Through their application, it will be demonstrated that financial bubbles can be identified before they peak, and that speculative episodes—even in highly liquid and informationally efficient markets—display a consistent endogenous component that can be measured and monitored.

This thesis contributes to the literature in three main ways:

1. It offers a systematic and comparative analysis of both econometric and non-econometric bubble detection techniques, an area in which existing studies are typically fragmented or methodologically isolated.

2. It expands the application of recent models—such as RCA and TDA—to asset classes beyond equities, such as energy futures, where their effectiveness is less documented.

3. It provides evidence supporting the hypothesis that financial bubbles, far from being unpredictable anomalies, can be anticipated through changes in the statistical and structural properties of price time series.

The empirical findings show that RCA is the most effective method in terms of early detection and signal stability. When coupled with exogenous variables—like the VIX or WTI—it provides both an early warning signal and an interpretable explanation of the regime change. LPPLS, while capable of predicting the timing of the crash with reasonable precision, is sensitive to the choice of in-sample windows and prone to instability. TDA offers a powerful visual and structural insight into market phases, particularly when applied to multi-dimensional time series, but its computational intensity and sensitivity to thresholds make it less practical. PE, although elegant and computationally efficient, underperforms in complex or noisy environments such as commodity futures.
This thesis shows that bubbles are not purely exogenous or random phenomena, but instead are shaped by internal market dynamics (particularly feedback loops, misperceptions, and herding behavior) that manifest themselves well before prices reach unsustainable levels, thus giving credit to Behavioral Finance. The combined use of econometric and non-econometric approaches allows for a more nuanced and comprehensive view of these phenomena, bridging theoretical models and empirical applications in financial market analysis.

1 Theoretical Frameworks

1.1 Efficient Market Hypothesis and Rational Bubbles

The efficient market hypothesis (Fama, 1970; Fama,1976) states that, regarding public markets, prices incorporate all available information, formally

$$f_m(p_{1t}, \ldots, p_{nt} | \phi^m_{t-1}) = f(p_{1t}, \ldots, p_{nt} | \phi_{t-1}). \tag{1.1}$$

where ϕ_{t-1} is the complete available information set at time $t-1$ and

$$f(p_{1t}, \ldots, p_{nt} | \phi_{t-1}) \tag{1.2}$$

is the joint density of prices of n assets conditioned on the available information set. The letter m stands for market, thus Equation 1.1 expresses the idea that the information used by it at time $t-1$ coincides with all available information and that the market uses this information correctly to determine asset prices. Furthermore, the EMH distinguishes three levels of information processing by investors (Fama, 1970):

1. The weak form: current prices incorporate only data about past prices and volumes. This implies the ineffectiveness of the usage of technical analysis to achieve above-average returns.

2. The semi-strong form: current prices incorporate all publicly available information, including earnings data. This demonstrates that also fundamental analysis cannot be used to outperform stock indexes.

3. The strong form: current prices incorporate also privately held information, implying that neither insider traders or investors who have privileged access to information (e.g. hedge funds which have real time access to brokers order flow) can outperform consistently in the long run.

While the forms 1. and 2. have been empirically tested (Fama et al., 1969 ; Fama, 1970; Jensen, 1978) particularly through "event studies", the third one has been criticized, being it also much more difficult to test. Generally speaking, the EMH implies that financial bubbles cannot be tested, and that prices follow a random walk

$$X_t = X_{t-1} + \epsilon_t \tag{1.3}$$

and that they are martingales, that is

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t \tag{1.4}$$

as stated by Fama, 1970. Brock (1982) and Tirole (1982) constructed a theoretical argument against the existence of bubbles in stock prices: assuming a constant number of asset holder with infinite planning horizon, if bubbles existed, investors would expect utility gains from selling the stock and never repurchasing it
However, Lo & MacKinlay (1988)demonstrated that prices do not always follow a random walk and that they can shift toward an explosive behavior, while others (Blanchard & Watson, 1982; Diba & Grossman, 1988) tried to explain the existence of bubbles in a rational market framework. Particularly, Diba & Grossman (1988) demonstrated the existence of strictly positive rational bubbles, i.e. the possibility of overestimation of assets values from the majority of investors, starting from the maximization of the following expected utility of a typical household over an infinite horizon:

$$\mathbb{E}_t \left[\sum_{\tau=t}^{\infty} \beta^{\tau-t} u(c_\tau) \right], \quad 0 < \beta < 1 \tag{1.5}$$

where c_τ is a stochastic process representing consumption of a single perishable good, and β is the discount factor for future consumption. The fact that present consumption is preferred over future consumption implies that the discount factor is lower than unity. The utility function $u(\cdot)$ is strictly concave, increasing and continuously differentiable. Each period, the household has an endowment y_τ of the aforementioned consumption good. The household can smooth consumption by acquiring shares, s_τ, at the price p_τ units of consumption good per share. Each share pays a dividend d_τ units of consumption good per period. The budget constraint faced by the household at a given point in time is

$$c_\tau + p_\tau(s_{\tau+1} - s_\tau) \leq y_\tau + d_\tau s_\tau \tag{1.6}$$

and the first order condition for the utility maximization is

$$p_\tau u'(c_\tau) = \beta \mathbb{E}_{\tau+1}\left[(\rho_{\tau+1} + d_{\tau+1})u'(c_{\tau+1})\right]. \tag{1.7}$$

By normalizing the number of shares per capita to 1, the market clearing condition is

$$c_\tau = y_\tau + d_\tau \quad \text{for all} \quad \tau \geq t \tag{1.8}$$

As done by Lucas (1978), it is possible to put Equation 1.8 into Equation 1.7 to get the expression

$$\mathbb{E}_t q_{t+1} - \beta^{-1} q_t = -\mathbb{E}_t\left[u'(y_{t+1} + d_{t+1})d_{t+1}\right] \tag{1.9}$$

where

$$q_t \equiv u'(y_t + d_t)p_t. \tag{1.10}$$

Equation 1.9 has a forward-looking solution that can be expressed as F_t, i.e. the market-fundamentals component of q_t, which is

$$F_t = \sum_{j=1}^{\infty} \beta^j \mathbb{E}_t\left[u'(y_{t+j} + d_{t+j})d_{t+j}\right] \tag{1.11}$$

Particularly, if the household is risk neutral, Equation 1.11 reduces the stock price to the sum of present values of expected future dividends. However, Equation 1.9 has also a general solution that includes not only a fundamental component, but also a positive rational bubble factor, that is

$$q_t = B_t + F_t \tag{1.12}$$

where B_t is the solution to the expectational difference equation

$$\mathbb{E}_t B_{t+1} - \beta^{-1} B_t = 0 \tag{1.13}$$

In this framework, market bubbles are endogenous, since they are the product of expectations of investors, which can anticipate the fact that other investor will feed the bubble in the future and therefore buy the asset in order to profit from this subsequent upward movement. However, this model does not explain how bubbles peak and crash, i.e. how B_t changes over time, often because of exogenous informational shocks, and why in many cases financial markets can remain persistently undervalued after a crash, i.e. how there can be a negative bubble, not only a positive one.

1.2 Behavioral Finance and Soros' Reflexivity

After the EMH started to prevail in the academic field, a growing number of psychologists and economists started to gather empirical evidence that investors are not perfectly rational, they do not always process information correctly, being afflicted by cognitive biases, and thus the degree of distortion of prices with respect to fundamentals may vary significantly, from negligible to consistent. The prospect theory (Kahneman & Tversky, 1979) demonstrates through a series of surveys that outcomes obtained with certainty are overweighted with respect to uncertain outcomes. If a choice between two positive events must be done, the so called ”certainty effect” makes people prefer a smaller sure gain over a larger gain that is merely probable, whereas in a choice between negative outcomes this same effect drives towards a larger merely probable loss over a certain smaller one. Also, if two similar events are compared, people tend to overestimate differences and underestimate analogies between them; this explains why it is a common bias among investors to insist on the fact that ”this time is different”, justifying the absence of a bubble with a different macroeconomic and technological environment. In addition to that, if markets were perfectly efficient, every specific investor which tries to outsmart it by gathering additional information would never be compensated for bearing the cost, therefore there must be a variable degree of non-incorporation of information into markets, assuming that investors are not equally informed (Grossman & Stiglitz, 1980).

The overreaction to positive and negative news is another reason for the erroneous incorporation of information into public markets (De Bondt & Thaler, 1985).

Shiller (2003) explores the feedback loop mechanism that underlies the development and crash of every speculative bubble: if prices go up for a long time, this attracts the attention of the media and academia towards that particular asset class and new theoretical models to justify this rise are created, thus enhancing a new round of price increases. If this feedback loop is not interrupted, after many round a speculative bubble is formed, in which high current prices are only supported by future expectations of equal or larger price increases. When these expectations of price increase are not met anymore, the bubble eventually bursts even if there is no news to support radical changes in the market's fundamentals, bringing prices to an unsustainable low level.

In his Reflexivity theory, Soros (2013) deepens the perspective on feedback loops given by Shiller (1990): they can be positive or negative; the positive ones are self-reinforcing, bringing about increasing detachment of prices from fundamentals, while negative ones are self-correcting, bringing prices from undervaluation to their fair value. Also, investors are not only perfect or imperfect, passive analyzers of information but have a performative role towards the market and, as a consequence, on future information about them. As an example, during 2021, Bulge Bracket investment banks forecasted a positive annual return for the S&P 500 also for 2022, thus allowing a prolonged positive self-reinforcing pattern that would be abruptly corrected the following year; in a counterfactual world, it is possible to imagine a negative reaction to a forecast of opposite sign made by these banks bringing about a milder correction and preventing the insurgence of speculative phenomena. According to the famous Hungarian speculator, the most important cognitive flaw for investors is confirmation bias: market operators tend to trust the vision of neighboring practitioners if these views are similar to their own and are prone to gather news that confirms their previous beliefs. Interestingly, a K-nearest neighbors setting for traders' behavior is also one of the key assumptions of the LPPLS model (Sornette & Johansen, 1997). Daniel, Hirshleifer and Subramanyam (1999) demonstrated that feedback loops are created mainly by a "self-attribution bias," identified by psychologist Daryl Bem (1965) as a pattern of behavior in which people attribute events similar to their view as the product of their skill, whereas events contrary to their vision are attributed to bad luck. Shiller (1990) synthesized feedback loops through an autoregressive distributed lag model, in which present stock returns are the weighted sum of past prices changes, to which exponentially declining weights are attributed. Thus, current price changes are explained mainly, but not exclusively, by recent changes. This framework has been proved by Jegadeesh and Titman (1993), which found that stock that showed exceptionally high six-months returns beat stock showing exceptionally low returns by 12 percent over the following year. This demonstrated that investors are afflicted by a "temporal proximity bias", i.e. they give more importance to recent events with respect to past ones not because of an objective greater impact on market fundamentals, but only because of their recency. Also, Shiller (1990) does not make any difference between institutional investors and uninformed traders, since it appears clearly that in most cases more expert investors amplify, rather than diminish, speculative feedback loops. De Long et al. (1990b) explains this behavior with the rational concern of experienced investors for the risk generated by the unexperienced ones and with the cost that offsetting this risk would be necessary to bear. Also in the case of behavioral finance bubbles are a substantially endogenous phenomenon, arising from erroneous information processing by practitioners.

1.3 Credit-Debt Cycles as a cause for bubbles

Minsky (1992) describes bubbles as the result of the change in attitude of businesses, households and bankers towards credit. He distinguishes three phases:

1. ”Hedge Finance”, in which people households and companies borrow money well knowing that they have the capacity to pay both principal and interest without rolling over the debt.
2. ”Speculative Finance”, in which people borrow now money in order to pay the principal, even if future revenues still have the capacity to pay the interest.
3. ”Ponzi Scheme Finance”, and it happens when borrowers have to borrow again to pay both the principal and the interest. This is the time in which crashes and recessions happen.

Although this framework is useful to describe crises that arise from high-leverage industries, such as the Subprime crisis of 2008, it cannot explain neither stock market bubbles arising from completely new technological trends (such as the Dotcom bubble in 2000) nor bubbles in new asset classes such as cryptocurrencies, since the impact of leverage on inflows in this new asset class is negligible.

1.4 The 2021-2022 bubble

In order to face the Covid crisis and avoid a complete collapse of internal demand due to lockdown measures, in 2020 central banks all around the world initiated Quantitative Easing measures and lowered policy rates to historically low levels. Fiscal Policy was oriented toward direct stimulation for households and businesses. After the pandemic ended, the delayed return to a normal monetary policy (the "tapering" phase) stimulated above-average stock market returns (the S&P 500 gained 16.16% in 2020 and 26.89% in 2021). The US stock market was flooded by fintech IPOs, particularly trading platforms such as Robinhood, which went public on Nasdaq on July 29, 2021. On November 2021, the Cyclically Adjusted PE ratio, formulated by Shiller in his book *Irrational Exuberance* (2000), reached 38.58, close to the historical maximum touched at the peak of the Dotcom bubble (44.19, December 1999). The most liquid cryptocurrencies reached new historical highs (Bitcoin was quoted $ 64402.50 on November 13 2021, while on the same day Ethereum reached the value of $ 4646.21). The Non-Fungible Token market was born, attracting $ 10 Billion mainly from HNWI investors. Many world-renowned multinationals, such as Maersk Line, announced the implementation of Blockchain solutions.

However, some hedge fund managers started to sense the irrational behavior of the markets: on 22 October 2021, David Tepper, founder and CEO of Appaloosa Capital Management, with a track record of 25% average return since 1993, gave an interview to CNBC in which he stated that "there are not really any great asset classes now"; as a matter of fact, the S&P 500 peaked on 31 December 2021 at 4776.18. At the beginning of 2022, rampant inflation in the EU and the US, caused by the delayed effect of demand subsidies, provoked a general increase in interest rates, while the Ukrainian war, which began on 20 February, caused the expulsion of Russia from the energy international market because of new American sanctions. This caused a spike in commodity prices, further intensifying inflationary pressures on advanced economies. In addition to this, earnings of the main technology companies, the "Magnificent Seven", largely disappointed market expectations. The S&P 500 declined until 14 October, losing approximately 25% of its value. This represents the most significant US stock market drawdown since the 2008 crisis. Only at the end of 2023 did the US stock market recover from this crash. In the following chapters, it will be demonstrated that this bubble cannot be considered as exogenous and that the assumption of existence of an endogenous component can be useful to detect it in advance.

2 Random Coefficient Autoregression

2.1 Introduction and literature review

Many econometric techniques have been developed for real-time detection of financial bubbles. Some of them are based on the calibration of p-values of the Augmented Dickey-Fuller test, to achieve the superuniformity property, that is to have a uniform distribution conditioned to the fact that the bubble has not yet developed (Genoni et al., 2023). Other models try to monitor in real time the change in the coefficient of an AR (1), to detect the transition time from a random walk to an explosive regime (Whitehouse et al., 2023). In addition to that, many *ex post* techniques of detection of bubbles have been formalized in recent years, such as the supremum-ADF test (Philips et al., 2011; Philips & Yu, 2011), and the generalized sup-ADF test (Philips et al., 2015a;Philips et al., 2015b).

The first model that will be applied is the Random Coefficient Autoregression (RCA) model, in which the AR coefficient is modified by a stochastic error term. A specific test statistic, belonging to the weighted-CUSUM family, will be used to detect market regime shifts in real time. The approach followed is that of Horváth & Trapani, 2022 and Horváth & Trapani, 2023. In this context, no prior knowledge about heteroskedasticity is needed.

2.2 Univariate Model

2.2.1 Stationarity conditions and model assumptions

The RCA model is

$$y_i = (\beta_i + \epsilon_{i,1})y_{i-1} + \epsilon_{i,2}. \tag{2.1}$$

As per Aue et al., 2006, in Equation (2.1) the stationarity or lack thereof is determined by $E \log |\beta_0 + \epsilon_{0,1}|$:

1. if $-\infty \leq E \log |\beta_0 + \epsilon_{0,1}| < 0$, then y_i converges exponentially fast to a strictly stationary solution $\forall y_0$
2. if $E \log |\beta_0 + \epsilon_{0,1}| > 0$, then y_i is nonstationary with $|y_i| \xrightarrow{a.s.} \infty$ exponentially
3. if $E \log |\beta_0 + \epsilon_{0,1}| = 0$, then $|y_i| \overset{P}{\to} \infty$ but a rate slower than exponential.

Two assumptions must be held in all three cases:

1. $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are IID random variables random variables with
   - $E[\epsilon_{i,1}] = E[\epsilon_{i,2}] = 0$
   - $0 < E[\epsilon_{i,1}^2] = \sigma_1^2 < \infty$ and $0 < E[\epsilon_{i,2}^2] = \sigma_2^2 < \infty$
   - $E[\epsilon_{i,1}\epsilon_{i,2}] = 0$
   - $E[|\epsilon_{i,1}|^\nu] < \infty$ and $E[|\epsilon_{i,2}|^\nu] < \infty$ for some $\nu > 2$
2. β_i is constant over the training set $\{y_i, 1 \leq i \leq m\}$, i.e. the *non-contamination assumption.*

In case of nonstationarity, three additional assumptions are required:

1. $\epsilon_{0,2}$ has a bounded density
2. $\epsilon_{i,1}$ and $\epsilon_{i,2}$ are independent $\forall$ i
3. $P\left\{(\beta_0 + \epsilon_{0,1})y_0 + \epsilon_{0,2} = x\right\} = 0 \ \forall \ -\infty < x < \infty$

The null hypothesis is that, in the test set after m, β_i remains constant, that is

$$H_0 : \beta_0 = \beta_{m+1} = \beta_{m+2} = \dots \tag{2.2}$$

2.2.2 Weighted CUSUM detector

The Weighted Least Squares estimator using the training set is

$$\hat{\beta}_m = \left(\sum_{i=1}^{m} \frac{y_{i-1}^2}{1 + y_{i-1}^2} \right)^{-1} \left(\sum_{i=1}^{m} \frac{y_i y_{i-1}}{1 + y_{i-1}^2} \right). \tag{2.3}$$

The cumulative sum (CUSUM) process of the WLS residual is

$$Z_m(k) = \left| \sum_{i=m+1}^{m+k} \frac{(y_i - \hat{\beta}_m y_{i-1})y_{i-1}}{1 + y_{i-1}^2} \right|, \quad k \geq 1 \tag{2.4}$$

Under the null hypothesis of no change of the estimator, the residuals have zero mean; therefore, the partial sum process $Z_m(k)$ should fluctuate around zero as well. If at k^* there is a break, then $\hat{\beta}_m$ is a biased estimator for the autoregressive coefficient β_{m+k^*+1}. A break is therefore signaled if $Z_m(k)$ exceeds a threshold defined by the following boundary function

$$g_{m,\psi}(k) = c_{\alpha,\psi} \delta m^{1/2} \left(1 + \frac{k}{m}\right) \left(\frac{k}{m+k}\right)^{\psi}. \tag{2.5}$$

with $0 \leq \psi \leq 1/2$ and δ defined as

$$\delta^2 = \begin{cases} a_1 \sigma_1^2 + a_2 \sigma_2^2, & \text{if } -\infty \leq E \log |\beta_0 + \epsilon_{0,1}| < 0, \\ \sigma_1^2, & \text{if } E \log |\beta_0 + \epsilon_{0,1}| \geq 0. \end{cases} \tag{2.6}$$

where σ_1^2 and σ_2^2 are variance of the errors, while a_1 and a_2 are defined as

$$a_1 = E \left(\frac{\bar{y}_0^2}{1 + \bar{y}_0^2}\right)^2, \quad \text{and} \quad a_2 = E \left(\frac{\bar{y}_0}{1 + \bar{y}_0^2}\right)^2 \tag{2.7}$$

where $\bar{y}_i$ is the stationary solution of Equation 2.1. The point of change of market regime is found at a stopping time $\tau_{m,\psi}$ defined as

$$\tau_{m,\psi} = \begin{cases} \inf\{k \geq 1 : Z_m(k) \geq g_{m,\psi}(k)\}, \\ \infty, \quad \text{if} \quad Z_m(k) < g_{m,\psi}(k) \quad \text{for all} \quad 1 \leq k < \infty. \end{cases} \tag{2.8}$$

The constant $c_{\alpha,\psi}$ is chosen to ensure that the I type error is lower than α under the null and that, under the alternative, $\lim_{m\to\infty} P\{\tau_{m,\psi} < \infty \mid H_1\} = 1$.

In the previous case, we operate in an "open-ended" monitoring, i.e. the sequential monitoring lasts for the entire out of sample. However, there is also a "closed-ended" framework, which can be applied by stopping the

monitoring after a given observation m^* because no break has been found or the training period may be extended; in the present thesis, the open-ended approach will be followed.
In three recent contributions,Fremdt (2015) , Kirch and Stoehr (2022a) and Kirch and Stoehr (2022b) created a different class of detector, which searches for the maximum value of $Z_m(k)$ in a given subset of observations from 1 to k, the Page-CUSUM process. However, since the accuracy of detection of regime change does not improve significantly, while the assumptions became much stricter, the analysis will be limited to the standard weighted CUSUM process.

2.2.3 Asymptotics

The formulation of the limits of the test statistics depends on the choice of ψ, the use of an "open-ended" or "closed-ended" approach and the type of detector (weighted CUSUM or Page-CUSUM).

Under the null hypothesis, $\forall\ \psi < 1/2$, it holds that

$$\lim_{m\to\infty} P\left\{\tau_{m,\psi} = \infty\right\} = P\left\{\sup_{0\le u\le 1} \frac{|W(u)|}{u^{\psi}} < c_{\alpha,\psi}\right\}. \tag{2.9}$$

where $W(\cdot)$ is a Wiener process.

Under H_1 we assume a change in the deterministic part of the autoregressive coefficient

$$y_i = \begin{cases} (\beta_0 + \epsilon_{i,1})y_{i-1} + \epsilon_{i,2}, & 1 \le i \le m + k^*, \\ (\beta_A + \epsilon_{i,1})y_{i-1} + \epsilon_{i,2}, & i > m + k^*. \end{cases} \tag{2.10}$$

with $\beta_0 \neq \beta_A$ and k^* is the time of regime change. The change can happen between two stationary regime, two nonstationary regimes, from a stationary to a nonstationary one and vice versa.

We assume

1. that market regime transition depends on the size of the training set m

$$\Delta_m = \beta_A - \beta_0 \tag{2.11}$$

2. that $k^* = O(m)$, i.e. that the breaking point gets further as the training set becomes larger

Following these two last assumptions, if it is true that

$$\lim_{m\to\infty} m^{1/2}|\Delta_m| = \infty \tag{2.12}$$

then, $\forall\ \psi < 1/2$,

$$\lim_{m\to\infty} P\left\{\tau_{m,\psi} < \infty \mid H_A\right\} = 1 \tag{2.13}$$

is also true.

2.3 Multivariate Model

Astill et al. (2023) proved that, by integrating exogenous variables into the RCA model, it is possible to have quicker detection of bubbles. The multivariate expansion of Equation 2.1 is

$$y_i = (\beta_i + \epsilon_{i,1})y_{i-1} + \lambda_{\mathbf{0}}^\top \mathbf{x}_i + \epsilon_{i,2}, \tag{2.14}$$

where y_0 is an initial value and $\mathbf{x}_i \in \mathbb{R}^p$. Equation 2.14 is a dynamic model with exogenous covariates, with λ_0 constant over time.
A weighted CUSUM detector is based on the following WLS loss function

$$G_m(\beta, \lambda) = \sum_{i=2}^{m} \frac{(y_i - \beta y_{i-1} - \lambda^\top \mathbf{x}_i)^2}{1 + y_{i-1}^2}. \tag{2.15}$$

The estimators of β_0 and λ_0 are defined as

$$(\hat{\beta}_m, \hat{\lambda}_m) = \arg\min_{\beta,\lambda} G_m(\beta, \lambda),$$

and satisfy the following condition

$$\frac{\partial}{\partial\beta} G_m(\hat{\beta}_m, \hat{\lambda}_m) = -2\sum_{i=2}^{m} \frac{(y_i - \hat{\beta}_m y_{i-1} - \hat{\lambda}_m^\top \mathbf{x}_i)y_{i-1}}{1 + y_{i-1}^2} = 0, \tag{2.16}$$

The weighted CUSUM detector is a generalization of Equation 2.4

$$Z_m^X(k) = \left| \sum_{i=m+1}^{m+k} \frac{(y_i - \hat{\beta}_m y_{i-1} - \hat{\lambda}_m^\top \mathbf{x}_i)y_{i-1}}{1 + y_{i-1}^2} \right|. \tag{2.17}$$

In addition to the assumptions contained in Chapter 2.2.1, it is also necessary to assume that

1. the exogenous covariates $\mathbf{x}_i$ form a weakly dependent and stationary process
2. $\mathbf{x}_i$ are independent of $\epsilon_{i,1}$ and $\epsilon_{i,2}$

The boundary function is defined as

$$g_m^{(x)}(k) = c_{\alpha,\psi}^{(x)} \delta_x^2 m^{1/2} \left(1 + \frac{k}{\delta_{x,d}^2 m}\right) \left(\frac{k}{\delta_{x,d}^2 m + k}\right)^{\psi}, \tag{2.18}$$

where $c_{\alpha,\psi}^{(x)}$ is a critical value, and

$$\delta_x^2 = \begin{cases} \delta_{x,2}^2/\delta_{x,1}, & \text{if } -\infty \leq E\log|\beta_0 + \epsilon_{0,1}| < 0, \\ \sigma_1, & \text{if } E\log|\beta_0 + \epsilon_{0,1}| > 0, \end{cases} \tag{2.19}$$

$$\delta_{x,d}^2 = \begin{cases} \delta_{x,2}^2/\delta_{x,1}^2, & \text{if } -\infty \leq E\log|\beta_0 + \epsilon_{0,1}| < 0, \\ 1, & \text{if } E\log|\beta_0 + \epsilon_{0,1}| > 0, \end{cases} \tag{2.20}$$

with

$$\delta_{x,1}^2 = \mathbf{a}^\top \mathbf{QCQa}, \quad \text{and} \quad \delta_{x,2}^2 = \sigma_1^2 E\left(\frac{\overline{y_0}}{1+\overline{y_0}^2}\right)^2 + \sigma_2^2 E\left(\frac{\overline{y_0}}{1+\overline{y_0}^2}\right)^2. \tag{2.21}$$

where $\mathbf{a}$ is a vector of weights, $\mathbf{Q}$ is a matrix including information on past normalized data, while $\mathbf{C}$ is the covariance matrix of the errors of the model. $\overline{y_0}$ is the stationary solution of the training set, while σ_1^2 and σ_2^2 are the variances of the innovations $\epsilon_{i,1}$ and $\epsilon_{i,2}$.
The stopping time rule definition is the same as Equation (2.8), albeit with different CUSUM detector and boundary function. The asymptotic theory is also the same with respect to the univariate model (Equation 2.9-2.13).

2.4 Simulations

First, I provide the results of simulations to find empirical rejection frequencies of the null hypothesis under the null of no changepoint under an "open-ended" framework. Only weighted CUSUM is considered. The coefficient β_0 is chosen in order to represent nonstationary regimes, random walks and explosive regimes, while $\sigma_1 = 0.1$ and $\sigma_2 = 0.2$.

ψ	p	$m = 50$	$m = 100$	$m = 200$	$m = 400$
0.00	0	0.053	0.017	0.015	0.003
0.00	1	0.034	0.021	0.003	0.006
0.00	2	0.027	0.014	0.003	0.004
0.25	0	0.070	0.032	0.017	0.006
0.25	1	0.031	0.017	0.016	0.005
0.25	2	0.034	0.024	0.007	0.007
0.45	0	0.067	0.035	0.034	0.027
0.45	1	0.043	0.020	0.014	0.009
0.45	2	0.031	0.020	0.012	0.015

Table 2.1: Weighted CUSUM - Open-Ended - Rejection Frequencies ($\beta_0 = 0.5$, $\alpha = 0.01$, p is the number of covariates)

ψ	p	$m = 50$	$m = 100$	$m = 200$	$m = 400$
0.00	0	0.062	0.043	0.028	0.018
0.00	1	0.043	0.033	0.023	0.007
0.00	2	0.049	0.027	0.016	0.004
0.25	0	0.058	0.046	0.029	0.027
0.25	1	0.049	0.047	0.025	0.015
0.25	2	0.043	0.037	0.018	0.024
0.45	0	0.067	0.046	0.030	0.041
0.45	1	0.056	0.034	0.047	0.043
0.45	2	0.074	0.047	0.035	0.037

Table 2.2: Weighted CUSUM - Open-Ended - Rejection Frequencies ($\beta_0 = 1$, $\alpha = 0.01$,p is the number of covariates)

ψ	p	$m = 50$	$m = 100$	$m = 200$	$m = 400$
0.00	0	0.021	0.010	0.001	0.002
0.00	1	0.020	0.008	0.005	0.001
0.00	2	0.014	0.010	0.005	0.006
0.25	0	0.013	0.007	0.010	0.006
0.25	1	0.021	0.015	0.011	0.004
0.25	2	0.021	0.014	0.007	0.003
0.45	0	0.022	0.011	0.008	0.010
0.45	1	0.012	0.007	0.008	0.010
0.45	2	0.022	0.007	0.015	0.006

Table 2.3: Weighted CUSUM - Open-Ended - Rejection Frequencies ($\beta_0 = 1.05$, $\alpha = 0.01$, p is the number of covariate)

3 The LPPLS model

3.1 Introduction and literature review

The second *ex ante* technique that will be presented is the Log-Periodic Power Law Singularity (LPPLS) model, also known as the Johansen-Ledoit-Sornette (JLS) model. This model affirms that speculative phenomena are not the result of an exponential increase in price since the exponential increase in prices is only the result of a compound interest with a constant compound rate. Rather, bubbles happen when prices assume a super-exponential behavior, i.e. when they increase at a rate that is increasing itself. This super-exponential growth happens because of positive self-reinforcing feedback loops in the valuations of assets among both traders and investors; this growth stops at a finite-time singularity, which happens at time t_c, before the crash starts to unfold. As said in Chapter 1.2, positive feedback loops are the result of an instinctive impulse of humans not only to imitate others but also to consider mainly data that confirm their previous beliefs, without listening to dissenting voices. The positive feedback phenomenon is not only characteristic of bubbles in asset pricing but can also happen in option hedging, insurance portfolio strategies, procyclical financing of firms by banks, and even network effects on social media. The LPPLS, appeared for the first time in Sornette & Johansen (1997), has been improved over time by Jonasen, Ledoit & Sornette (2000) and by Filimonov & Sornette (2013). It has been used to successfully detect bubbles and crashes in emerging markets (Jiang et al., 2010), in advanced economies (Zhou et al., 2008; Zhou & Sornette, 2003; Zhou & Sornette, 2006) and in commodities markets (Sornette et al., 2009). Also, confidence intervals have been constructed to detect probabilities of positive bubbles in real time (Shu & Song, 2024).

3.2 Assumptions

1. Prices accelerate hyperbolically, i.e. via a super-exponential trajectory with compounding growth rate. This price trajectory can be synthesized by the following formula (Sornette & Johansen, 1997):

   $$\log p(t) = A + B(t_c - t)^m \tag{3.1}$$

   with $B < 0$ and $0 < m < 1$ and t_c being the end of the bubble (and the beginning of the crash).

2. Price growth is directly proportionate to the probability of a crash, which grows as price approaches to the singularity.

3. Competition between two sets of market operators (short-term traders and long-term investors) bring log-periodic oscillations to the super-exponential trend. As the price approaches t_c, their wavelength diminishes and their frequence grows.

4. Traders operate in a herd-like environment, and the choice they make (to be a short-term or long-term operator) is highly influenced by how other near traders behave. In mathematical terms, this would be a k-means clustering setting.

3.3 Model Derivation

The LPPLS model aims to describe the dynamics of financial bubbles by characterizing their intrinsic instability and the buildup toward a finite-time singularity. Following Sornette & Johansen (1997), we begin with the standard price dynamic expressed as a stochastic differential equation:

$$\frac{dp}{p} = \mu(t)dt + \sigma(t)dW - \kappa dj. \tag{3.2}$$

Here, $p(t)$ is the asset price, $\mu(t)$ is the drift term, $\sigma(t)dW$ is the stochastic component following a standard Wiener process, and κdj models a discrete price drop associated with a crash. The variable dj is a jump process that takes the value 0 before the crash and 1 after it. The occurrence of this jump is governed by a time-varying crash hazard rate $h(t)$, which expresses the probability that a crash occurs in the infinitesimal interval $[t, t + dt]$, conditional on survival until t.

Given this framework, the expected value of the jump component becomes:

$$E_t[dj] = h(t)dt. \tag{3.3}$$

To incorporate the idea that the likelihood of a crash increases with the unsustainable growth of the bubble, the hazard rate is modeled with a log-periodic power law form:

$$h(t) = B'(t_c - t)^{m-1} + C'(t_c - t)^{m-1} \cos\left(\omega \ln(t_c - t) - \phi'\right). \tag{3.4}$$

This function captures both the accelerating risk (through the power law term) and the cyclical fluctuations often seen in price trajectories near speculative peaks (through the cosine term). These oscillations are interpreted as the result of alternating phases of bullish and bearish sentiment among heterogeneous traders, consistent with behavioral finance theories of herding, imitation, and confirmation bias.

Assuming no arbitrage in a risk-neutral world, the expected return of the asset must be zero. Taking expectations in Equation (1), and using the properties $E_t[dW] = 0$ and $E_t[dj] = h(t)dt$, we get:

$$E_t[dp] = \mu(t)p(t)dt - \kappa p(t)h(t)dt = 0 \tag{3.5}$$

which implies:

$$\mu(t) = \kappa h(t). \tag{3.6}$$

Thus, the drift term (i.e. the expected return) grows proportionally with the perceived crash probability. Conditioning on the fact that the crash has not yet occurred, the stochastic differential equation becomes:

$$\frac{dp}{p} = \kappa h(t)dt + \sigma(t)dW. \tag{3.7}$$

Taking expectations and integrating the deterministic part, we obtain the expected log-price:

$$E_t[\ln p(t)] = A + B(t_c - t)^m + C(t_c - t)^m \cos\left(\omega \ln(t_c - t) - \phi\right). \tag{3.8}$$

Here, A is the expected log-price at the critical time t_c, the moment in which the bubble peaks. $B = -\kappa B'/m$ controls the magnitude of the accelerating growth, and $C = -\kappa C'/\sqrt{m^2 + \omega^2}$ governs the amplitude of log-periodic fluctuations. The exponent $m \in (0, 1)$ ensures that prices remain finite up to the singularity, while ω and ϕ control the frequency and phase of oscillations, respectively.

Filimonov & Sornette (2013) proposed an alternative formulation to simplify estimation. They expand the cosine term into sine and cosine components using the identities

$$C_1 = Ccos\phi \tag{3.9}$$

and

$$C_2 = Csin\phi \tag{3.10}$$

which represent the amplitude of log-periodic oscillations. Thus, we obtain the following equation as a result:

$$\begin{aligned} E_t[\ln p(t)] = A &+ B(t_c - t)^m \\ &+ C_1(t_c - t)^m \cos[\omega \ln(t_c - t)] \\ &+ C_2(t_c - t)^m \sin[\omega \ln(t_c - t)]. \end{aligned} \tag{3.11}$$

This reparameterization eliminates the need to estimate the phase ϕ directly and reduces the number of nonlinear parameters from four (t_c, m, ω, ϕ) to three (t_c, m, ω), enhancing numerical stability.

3.4 Estimation

The model is Equation (3.11) is characterized by three nonlinear parameters (t_c, m, ω) and four linear ones (A,B,C_1,C_2). Following the steps of Shu & Song (2024) the estimation can be done by minimizing the sum of squared residuals (SSR)

$$\begin{aligned} F(t_c, m, \omega, A, B, C_1, C_2) = \sum_{i=1}^{N} \Bigg[& \ln p(\tau_i) - A - B(t_c - \tau_i)^m \\ & - C_1(t_c - \tau_i)^m \cos(\omega \ln(t_c - \tau_i)) \\ & - C_2(t_c - \tau_i)^m \sin(\omega \ln(t_c - \tau_i)) \Bigg]^2 . \end{aligned} \tag{3.12}$$

In order to have a "pure" nonlinear optimization, we consider the four linear parameters as dependent on the nonlinear ones, thus having the following optimization problem:

$$\{t_c, \hat{m}, \hat{\omega}\} = \arg \min_{t_c, m, \omega} F_1(t_c, m, \omega) \tag{3.13}$$

where the cost function F_1 is

$$F_1(t_c, m, \omega) = \min_{(A,B,C_1,C_2)} F(t_c, m, \omega, A, B, C_1, C_2),$$

and this optimization problem can be rewritten as:

$$(\hat{A}, \hat{B}, \hat{C}_1, \hat{C}_2) = \arg \min_{(A,B,C_1,C_2)} F(t_c, m, \omega, A, B, C_1, C_2) \tag{3.14}$$

$$= \arg \min_{(A,B,C_1,C_2)} \sum_{i=1}^{N} [\ln p(\tau_i) - A - Bf_i - C_1 g_i - C_2 h_i]^2 \tag{3.15}$$

where

$$f_i = (t_c - t_i)^m, \tag{3.16}$$

$$g_i = (t_c - t_i)^m \cos(\omega \ln(t_c - t_i)), \tag{3.17}$$

and

$$h_i = (t_c - t_i)^m \sin(\omega \ln(t_c - t_i)). \tag{3.18}$$

Now the LPPLS estimation problem has been divided into a two steps optimization: first, we estimate the non-linear parameters using metaheuristic algorithms and derivative based methods, then, keeping the non-linear parameters constant, it is possible to proceed to estimate the linear ones with Ordinary Least Squares.

3.4.1 Estimation of non-linear parameters

Since the cost function is extremely complex, it is useful to use a metaheuristic algorithm to have good preliminary conditions for the nonlinear parameters. One of the most used derivatives-free optimization algorithms used for this purpose is the taboo search algorithm (Cvijovic & Klinowski, 1995). This algorithm starts by defining a search space S, thus the problem becomes

$$\min_{s \in S} f(s).$$

where s is the starting solution. We define a neighborhood $N(s)$ as

$$N(s) = \{\, s' \in S \mid d(s, s') \leq \delta\}.$$

where $d(s, s')$ is the distance (e.g. Euclidean) between the first and any other solution in that given neighborhood. Each of the already verified solutions is included in a taboo (i.e. non-touchable) list, and cannot be considered as feasible unless it satisfies the following condition:

$$f(s_n) < min\ f(s_{1:n-1}).$$

The algorithm stops either when the value of the objective function is under a predetermined value or the improvement of the objective function becomes negligible. The boundaries of the search space has been suggested by Shu & Zhu, 2020b:

1. $m \in [0, 1]$
2. $\omega \in [1, 50]$
3. $t_c \in [t_2, t_2 + \frac{t_2 - t_1}{3}]$, where t_1 and t_2 are respectively the beginning and the end of in-sample data
4. $\frac{m|B|}{\omega\sqrt{C_1^2 + C_2^2}} \geq 1$ to ensure the non-negativity of crash hazard rate $h(t)$.

Once the boundaries are defined, the Taboo Search algorithm is executed to generate robust initial values for the nonlinear parameters t_c, m, ω. These initial values are then refined using the Levenberg–Marquardt (LM) algorithm, a damped least-squares optimization method designed to minimize the nonlinear sum of squared residuals (SSR). The LM algorithm iteratively solves the following system:

$$(J^\top J + \lambda I)\Delta p = -J^\top r, \tag{3.19}$$

where J is the Nx3 (where N is the number of in-sample observations) Jacobian matrix of partial derivatives of the residuals with respect to the nonlinear parameters, λ is the damping parameter, I is the identity matrix, Δp is the update vector for the nonlinear parameters, and r is the residual vector. The Jacobian matrix J is defined as:

$$J_{ij} = \frac{\partial r_i}{\partial \theta_j}$$

where $\theta_j \in \{t_c, m, \omega\}$, and $r_i = \ln p(t_i) - f(t_i; \theta)$. These partial derivatives are computed numerically using finite difference approximations.

3.4.2 Estimation of linear parameters

Once the optimal nonlinear parameters $\hat{t}_c, \hat{m}, \hat{\omega}$ are estimated, the model becomes linear in the parameters A, B, C_1, C_2, and can be estimated using Ordinary Least Squares (OLS).

We define the transformed variables:

$$f_i = (t_c - t_i)^m, \quad g_i = f_i \cos(\omega \ln(t_c - t_i)), \quad h_i = f_i \sin(\omega \ln(t_c - t_i)),$$

so that the regression model can be written as:

$$\ln p(t_i) = A + B f_i + C_1 g_i + C_2 h_i + \varepsilon_i.$$

The model in matrix form becomes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where:

$$\mathbf{X} = \begin{pmatrix} 1 & f_1 & g_1 & h_1 \\ 1 & f_2 & g_2 & h_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & f_N & g_N & h_N \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} A \\ B \\ C_1 \\ C_2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \ln p(t_1) \\ \ln p(t_2) \\ \vdots \\ \ln p(t_N) \end{pmatrix}.$$

The OLS estimator is then:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This completes the two-step estimation process: first the nonlinear parameters via Taboo Search and Levenberg-Marquardt optimization, and then the linear coefficients via closed-form OLS regression.

3.5 Simulations

First, the model is applied to three different types of time series, each of them with log-periodic oscillations:

1. An AR process $y_t = (\phi_1 y_{t-1} + \epsilon_t)(1 + 0.1\ cos(2\pi t/20))$ with $\phi_1 = 0.8$.
2. A log-periodic random walk
3. An explosive process with log-periodic oscillations $y_t = (\phi_1 y_{t-1} + \epsilon_t)(1 + 0.1\ cos(2\pi t/20))$ with $\phi_1 = 1.05$

Each of the 1000 simulations has 500 steps. Without prior constraints on parameter bounds, the LPPLS is applied on each time series to verify whether the four aforementioned conditions are respected. If the fitted LPPLS respects all of the four conditions, then the simulation is classified as a bubble. The percentages of simulations detected as bubbles have the following percentages:

AR(1) with ϕ=0.8	**RW**	**AR(1) with $\phi = 1.05$**
0.00%	0.81%	100%

Table 3.1: Simulation Results

The distribution of m, t_c and ω are provided for all three models

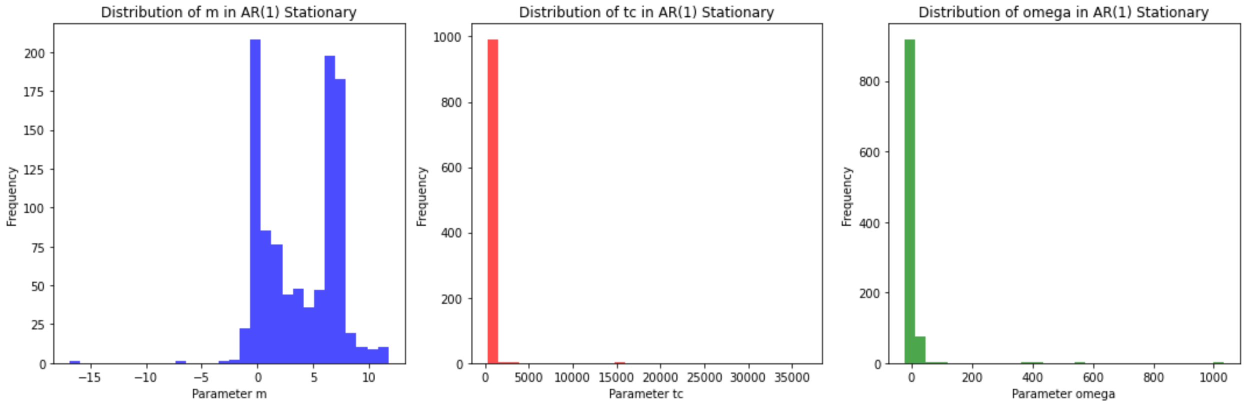


Figure 3.1: Stationary AR(1) distributions of m (blue), t_c (red) and ω (green)

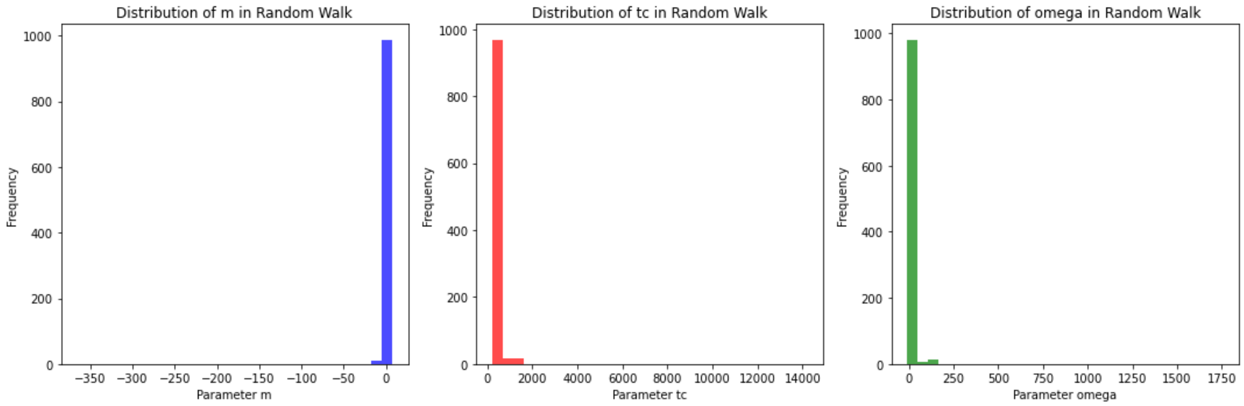


Figure 3.2: RW distributions of m (blue), t_c (red) and ω (green)

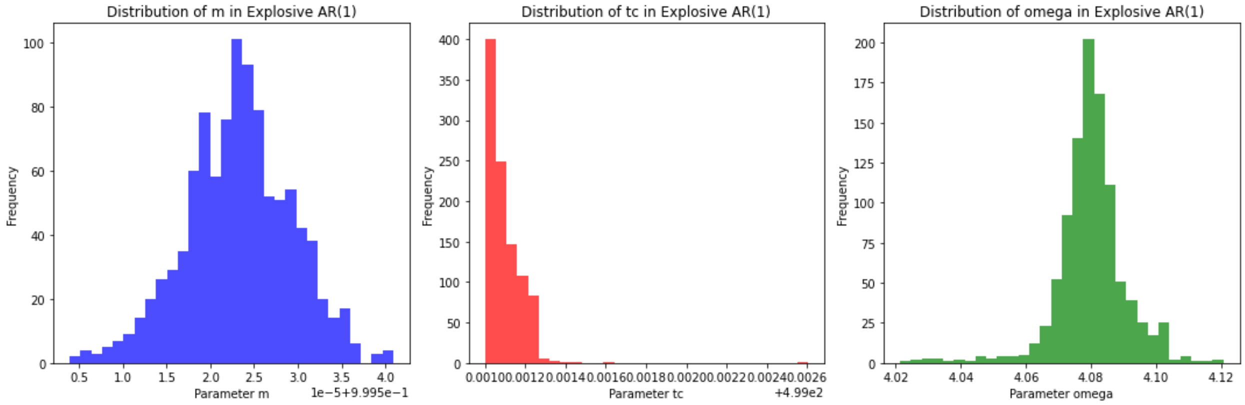


Figure 3.3: Explosive AR(1) distributions of m (blue), t_c (red) and ω (green)

The LPPLS model is very robust, being able to distinguish among stationary, nonexplosive, and explosive processes.

4 Topological Data Analysis

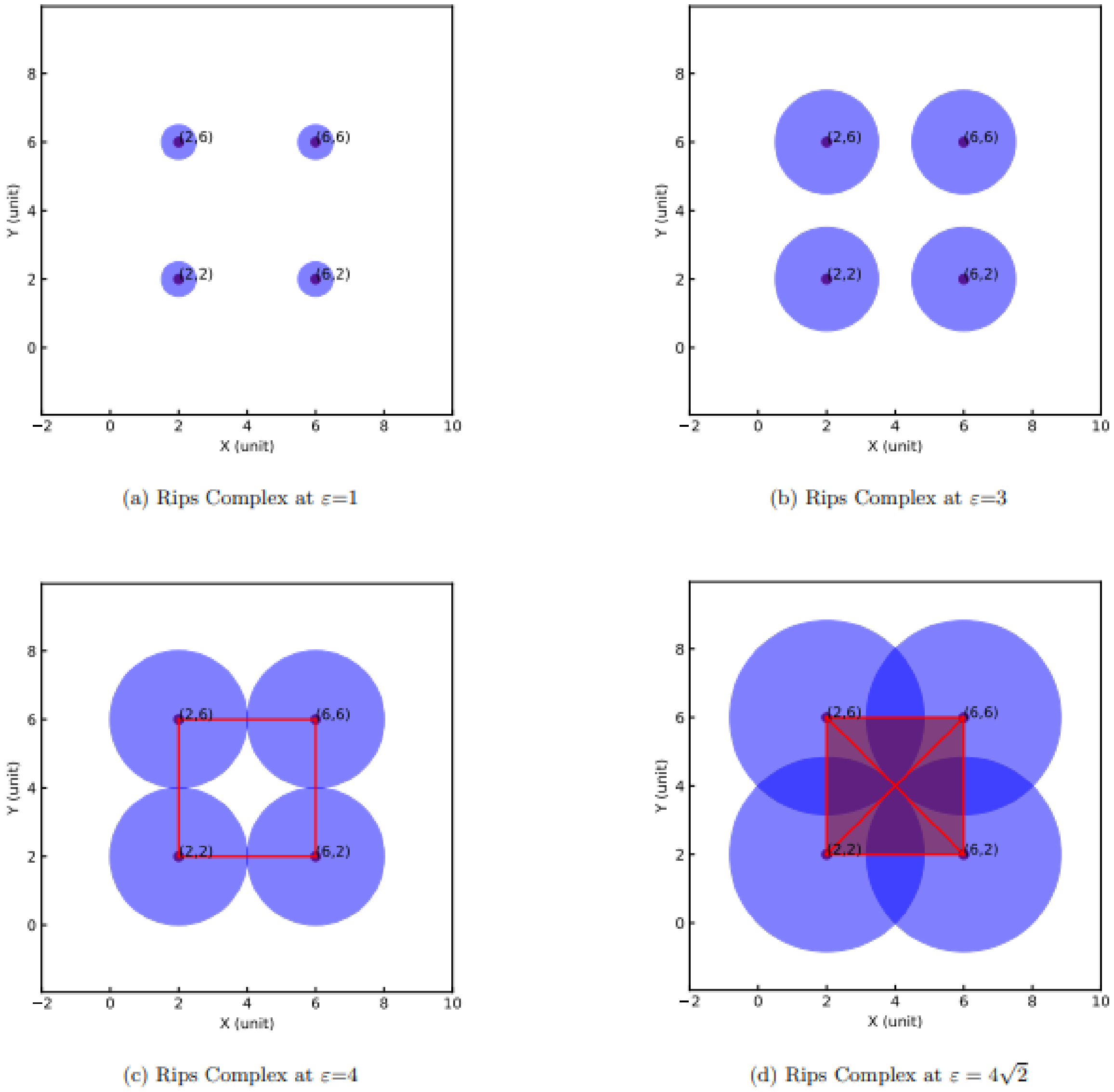
4.1 Introduction and literature review

Topological Data Analysis has drawn attention in recent years as a powerful tool in time series analysis and signal processing, particularly through the transformation of a time series into a point cloud in a higher-dimensional Euclidean space. The shape of the point cloud can be characterized by persistent homology, that is the connections among a given subset of the point cloud that persist at different resolution levels. The study of persistent homology in a sliding window of points can give early warning signals of potential shifts between regimes in a time series. TDA has been applied in different fields, such as finance, climatology, biology, and biomedicine (e.g. Sheffer et al., 2009; Lenton, 2011; Thompson & Sieber, 2011), specifically to detect critical transitions. A growing field of research pertains to the application of TDA to financial time series, particularly to detect both positive and negative bubbles (Philips & Yu, 2011; Gidea & Katz, 2017; Gidea et al., 2018; Akingbade et al., 2023; Rai et al., 2024). TDA performs very well when facing transition to endogenous bubbles, less so when the bubble has mainly exogenous nature (Song & Zhu, 2020). The evidence on the ability of TDA to detect bubbles has been mostly empirical. Some research tries to explain it by assuming the fact that time series follow the already mentioned log-periodic power law behavior before crashing. Some works lead positive TDA results back to the analogy between markets undergoing a crash and bifurcations in dynamical systems, while others explain it with drift changes or volatility shifts.

4.2 Methodology

Let us consider a univariate financial time series of length n. It is possible to transform this series into a multidimensional object by selecting a window length m and creating overlapping rolling windows of the same size. Each of these windows can be treated as a point in the space $\mathbb{R}^m$, resulting in a high-dimensional point cloud that represents the evolving structure of the time series.

To extract information from this point cloud, we apply a Vietoris–Rips filtration. More precisely, we select a second window of size $k \gg m$, and for each point in this window, we center a ball of radius ϵ. As ϵ increases, these balls start intersecting and generate simplices: connected pairs, triangles, tetrahedra, and so on. The presence of loops (1-dimensional homologies, which are the only ones we focus on) formed during this process is especially informative in financial contexts, as it may reflect coordinated or cyclical behavior among agents. As ϵ grows further, these features eventually disappear, marking their "death." The crucial concept in persistent homology is the *lifespan* of such features. The difference between the scale at which a feature appears (birth) and the scale at which it disappears (death) defines its persistence. Robust features persist over wide ranges of ϵ, while short-lived features are typically treated as noise. These topological features are encoded in a *persistence diagram* $\mathcal{P}_k$, a scatterplot in $\mathbb{R}^2$ where each point (b_α, d_α) represents the birth and death of a feature α of dimension k. The farther a point lies from the diagonal $b = d$, the more persistent the associated feature.



(a) Rips Complex at ε=1

(b) Rips Complex at ε=3

(c) Rips Complex at ε=4

(d) Rips Complex at $\varepsilon = 4\sqrt{2}$

Figure 4.1: Vietoris-Rips filtration (adapted from Rai et al., 2024)

In the example above, four initially disconnected points (a) form edges as the radius grows (b), eventually generating a loop (c) which collapses as more triangles are filled in (d). The resulting persistence diagram is shown below:

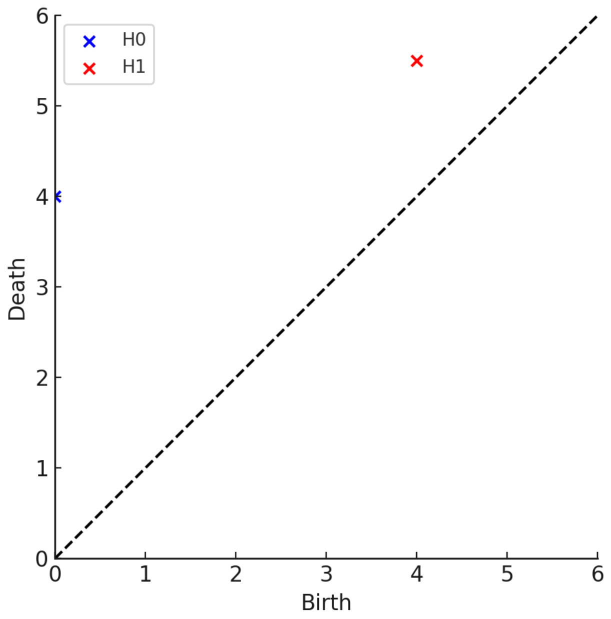


Figure 4.2: Corresponding Persistence Diagram

To enable quantitative analysis, the persistence diagram is transformed into a *persistence landscape*, which maps each topological feature into a piecewise-linear function. Given a persistence diagram P_k, the landscape consists of a sequence of functions $\lambda_k(x)$ constructed as follows:

$$\lambda_k(x) = \max \left\{ f_{(b_\alpha, d_\alpha)}(x) \;\middle|\; (b_\alpha, d_\alpha) \in P_k \right\}_{[k\text{-th largest}]},$$

where $f_{(b_\alpha, d_\alpha)}(x)$ is the triangular tent function associated with each point:

$$f_{(b_\alpha, d_\alpha)}(x) = \begin{cases} x - b_\alpha, & \text{if } x \in \left(b_\alpha, \frac{b_\alpha + d_\alpha}{2}\right] \\ d_\alpha - x, & \text{if } x \in \left[\frac{b_\alpha + d_\alpha}{2}, d_\alpha\right) \\ 0, & \text{otherwise.} \end{cases}$$

Each triangle has as its base the interval between birth and death, and its peak at the midpoint. A visualization of this transformation is given below:

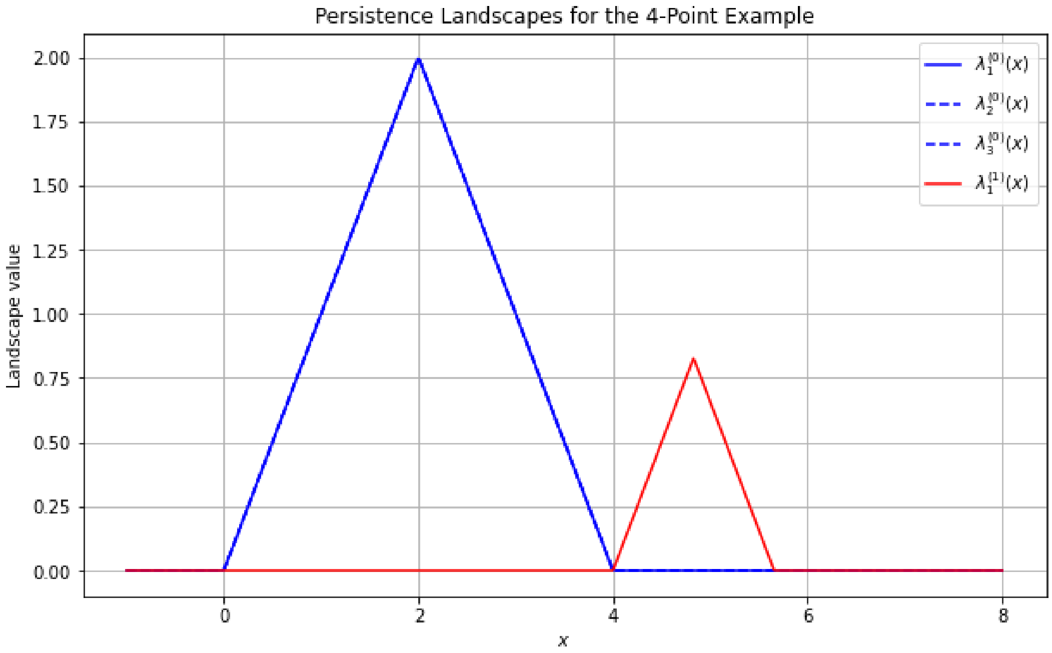


Figure 4.3: Persistence Landscape. The blue traingle represente the 0-dimensional homology, while the red represent the 1-dimnesional homology

Once the persistence landscape has been constructed, its L^p norm can be used to summarize the information into a single scalar indicator. The most common choice is the L^1-norm, which is computed as:

$$\|\lambda\|_{L^1} = \int_{-\infty}^{\infty} |\lambda_k(x)| \, dx.$$

This norm captures the total topological activity in the signal, giving greater weight to persistent features. In financial applications, it has been empirically shown that the L^1-norm increases significantly during the formation of bubbles, making it a useful early warning indicator of speculative regimes.

5 Permutation Entropy

5.1 Introduction and Literature Review

Permutation Entropy (PE), introduced by Bandt and Pompe (2002), is a non-parametric and model-free method used to quantify the complexity of time series by examining the order relations between values. Unlike classical entropy measures, PE is based on the ordinal structure of the data and is therefore robust to noise and invariant under monotonic transformations. Its utility has been demonstrated in distinguishing between deterministic, stochastic, and chaotic dynamics. More recently, PE has been applied in the context of financial markets, especially in emerging markets (Hou et al., 2023), as a tool to detect structural changes or episodes of abnormal behavior in asset prices. In this thesis, we evaluate its effectiveness in detecting speculative dynamics in the S&P 500 and the NYMEX Natural Gas futures market.

5.2 Methodology

Given a time series $\{x_t\}_{t=1}^{T}$, the Permutation Entropy is computed via the following steps:

For a fixed embedding dimension m and time delay τ, we generate a sequence of vectors:

$$\mathbf{v}_i = [x_i, x_{i+\tau}, x_{i+2\tau}, \ldots, x_{i+(m-1)\tau}], \quad \text{for } i = 1, \ldots, T - (m-1)\tau.$$

Each vector $\mathbf{v}_i$ is then mapped to its corresponding *ordinal pattern* π, which is the permutation of indices that reorders the elements of $\mathbf{v}_i$ in increasing order. In case of ties (equal values), the original order of appearance is preserved. For example, for the vector

$$\mathbf{v}_i = [4.1, 2.0, 3.5],$$

the ordinal pattern is

$$\pi = (1, 2, 0),$$

since $2.0 < 3.5 < 4.1$.

We then compute the relative frequency p_j of each of the $m!$ possible ordinal patterns:

$$P = \{p_1, p_2, \ldots, p_{m!}\}, \quad \sum_{j=1}^{m!} p_j = 1.$$

The permutation entropy is defined as the Shannon entropy of this distribution:

$$H = -\sum_{j=1}^{m!} p_j \log p_j.$$

To make the measure scale-invariant, H is normalized by the maximum entropy $\log(m!)$, leading to:

$$H_{\text{norm}} = \frac{H}{\log(m!)}.$$

The normalized Permutation Entropy H_{norm} lies in the interval $[0, 1]$. A value close to 1 suggests that the time series exhibits randomness similar to a white noise or random walk. Conversely, values closer to 0 suggest increased determinism or the presence of temporal structure, such as causality or regime shifts.

In this study, we use $m = 2$ and $\tau = 1$ as standard parameters following the literature, ensuring simplicity and comparability with prior applications.

To evaluate the statistical significance of the observed permutation entropy, we perform a one-sided Monte Carlo test against the null hypothesis that the series is generated by a random walk. The procedure is as follows:

1. Simulate $N = 1000$ independent random walks, each with the same length T as the original time series.

2. Compute the normalized PE for each simulated series, obtaining an empirical distribution $\{PE_{\text{rw}}^{(i)}\}_{i=1}^{N}$.

3. Compute the PE of the observed time series, denoted PE_{real}.

4. Calculate the empirical p-value as:

$$p = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(PE_{\text{rw}}^{(i)} \leq PE_{\text{real}}),$$

   where $\mathbb{I}(\cdot)$ is the indicator function.

If $p < 0.05$, we reject the null hypothesis and conclude that the time series exhibits statistically significant deviation from random walk behavior, which may indicate the presence of nonlinear structure or speculative dynamics.

6 Data Description

6.1 Main variables

The main variables object of the thesis are the S&P 500 index and the continuous future of natural gas, as listed on the New York Mercantile Exchange (NYMEX). All the data are taken from Bloomberg, as daily close prices. In the case of the S&P 500, daily prices are taken as adjusted for dividends. The daily S&P 500 for years 2021-2022 has this path.

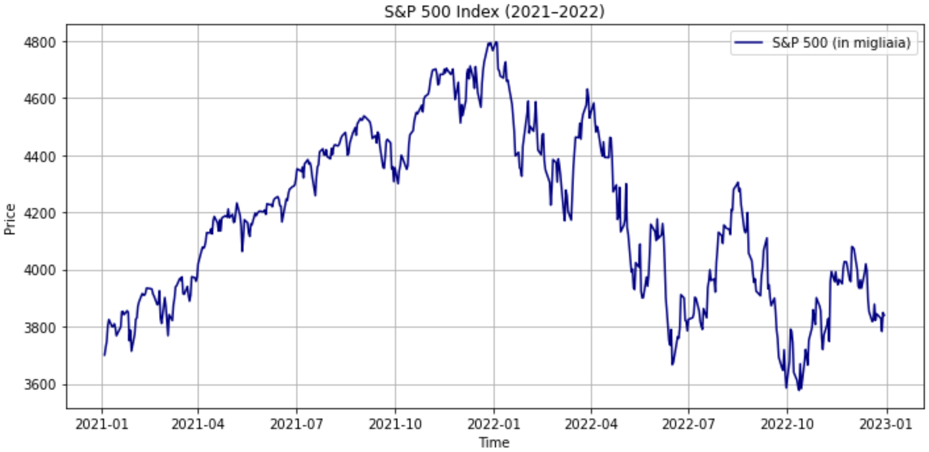


Figure 6.1: S&P daily price for 2021-2022

with this subsequent main descriptive statistics.

Table 6.1: Descriptive statistics for the S&P 500 index

Variable	Count	Mean	Std Dev	Min	25%	Median	75%	Max
S&P 500	504	4.186	0.302	3.577	3.919	4.181	4.437	4.797

The continuous natural gas future has had the following trajectory.

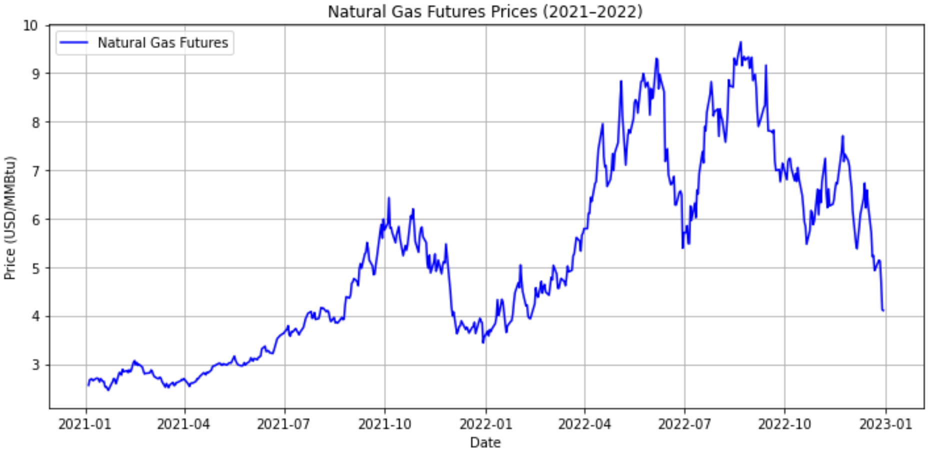


Figure 6.2: Natural Gas daily price for 2021-2022

with these descriptive statistics.

Table 6.2: Descriptive statistics for Natural Gas Futures

Variable	**Count**	**Mean**	**Std Dev**	**Min**	**25%**	**Median**	**75%**	**Max**
Natural Gas Futures	504	5.164	1.963	2.456	3.599	4.871	6.705	9.647

6.2 Control variables

Apart from the two main variables, in order to enhance the detection of bubbles, control variables are also considered. The first subgroup of control variables is strictly macroeconomic. They are the US stock market volatility index (VIX), the US Dollar Index (DX), WTI crude oil, and the yield to maturity of US 3-month Bills and 10-year Treasuries. All of them are considered as daily close prices and are taken from Bloomberg.

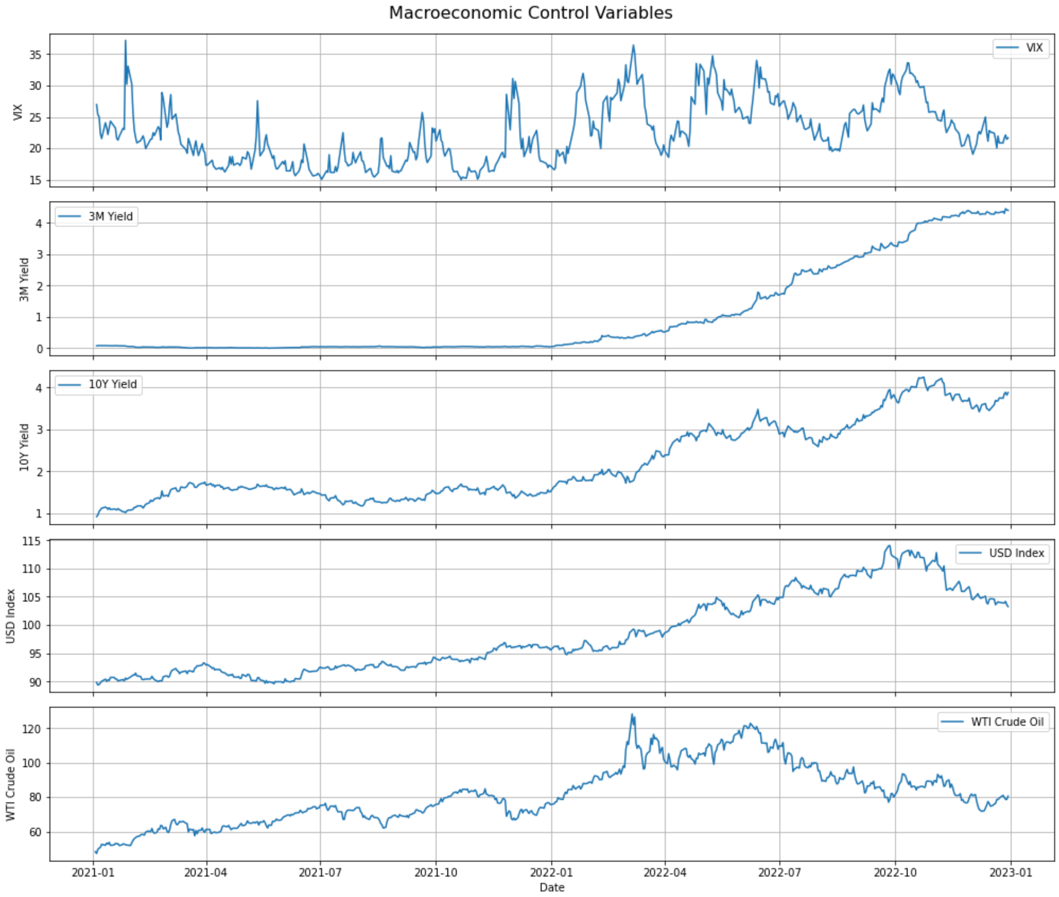


Figure 6.3: From above: VIX, 3M yield, 10y yield, Dollar Index and WTI during 2021-2022. Daily data

The main descriptive statistics for these control variables are the following.

Table 6.3: Descriptive Statistics of macroeconomic control variables

Variable	**Count**	**Mean**	**Std. Dev.**	**Min**	**25%**	**Median**	**Max**
VIX	504	22.64	4.92	15.01	18.58	21.89	37.21
3M Yield	504	1.05	1.46	0.01	0.05	0.09	4.46
10Y Yield	504	2.19	0.92	0.92	1.48	1.71	4.25
USD Index	504	98.21	6.94	89.41	92.41	96.01	114.05
WTI Crude Oil	504	81.67	17.32	47.28	68.61	79.68	128.26

These statistics reveal several important dynamics. The natural gas market displays significant price volatility, with values ranging from 2.5 to nearly 10 USD—more than triple its minimum. Conversely, the rapid increase in short-term interest rates (3M yield) from near-zero levels to above 4% highlights the aggressive monetary tightening conducted by the Federal Reserve over the period. Beyond these macroeconomic indicators, the analysis also incorporates four sector-specific exchange-traded funds (ETFs), particularly relevant in the context of the Topological Data Analysis (TDA) framework. These include:

- **XLK** – Technology Select Sector SPDR Fund
- **XLF** – Financial Select Sector SPDR Fund
- **XLY** – Consumer Discretionary Select Sector SPDR Fund
- **XLU** – Utilities Select Sector SPDR Fund

These ETFs are used to isolate and capture sectoral dynamics within the broader market. Each represents a distinct segment of the S&P 500 and is used to explore whether the early signals of financial bubbles are more pronounced in specific sectors.

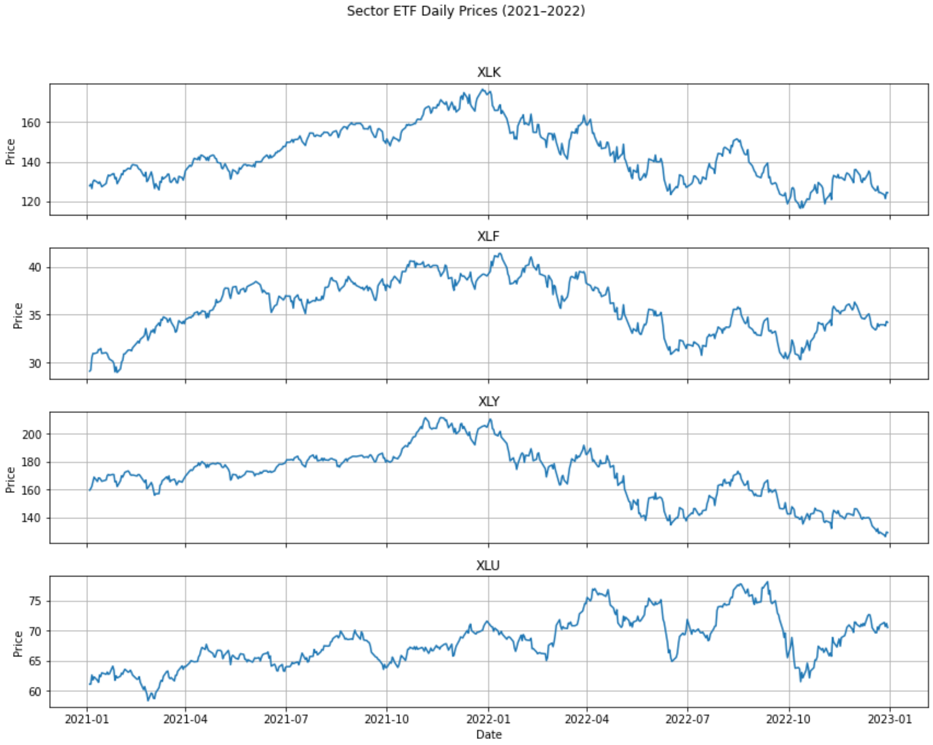


Figure 6.4: Daily price trajectories of sector ETFs for the 2021–2022 period. From top to bottom: XLK (Technology), XLF (Financials), XLY (Consumer Discretionary), and XLU (Utilities)

Table 6.4: Descriptive statistics for sector ETFs (2021–2022)

ETF	**Count**	**Mean**	**Std Dev**	**Min**	**25%**	**Median**	**75%**	**Max**
XLK (Tech)	504	144.02	13.93	116.56	132.53	142.10	154.45	176.65
XLF (Financials)	504	35.77	2.93	28.95	33.52	35.96	38.21	41.42
XLY (Cons. Disc.)	504	169.94	19.57	126.26	155.62	171.66	181.88	211.42
XLU (Utilities)	504	68.09	4.11	58.36	65.20	67.45	70.71	78.12

The different price paths and descriptive statistics reveal heterogeneous behavior across sectors. The Technology ETF (XLK) exhibits the highest average price, with significant variation, reflecting growth-oriented volatility. Conversely, the Utilities ETF (XLU) shows more stable dynamics, with the lowest standard deviation and narrower price range. Financials (XLF) and Consumer Discretionary (XLY) present intermediate patterns, with XLY displaying greater volatility due to its sensitivity to cyclical consumption. These sectoral differences are crucial for understanding how financial bubbles may emerge or be anticipated differently across market segments.
Regarding the application of multivariate TDA, three different stock indexes are employed together with the S&P 500. These indexes are:

- **RUT** – the Russell 2000 index, composed by the smallest 2000 American publicly listed companies
- **DJI** – the Dow Jones Industrial Average, made by 30 among the largest US listed companies. It is the only price weighted index still tracked worldwide.
- **IXIC** – the Nasdaq Composite, mainly composed by technology stocks.

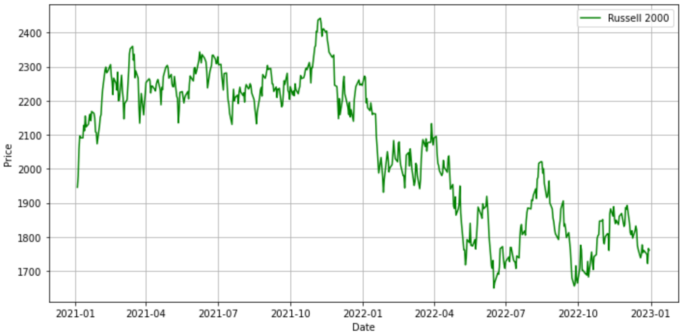


Figure 6.5: Russell 2000 daily adjusted close price 2021-2022

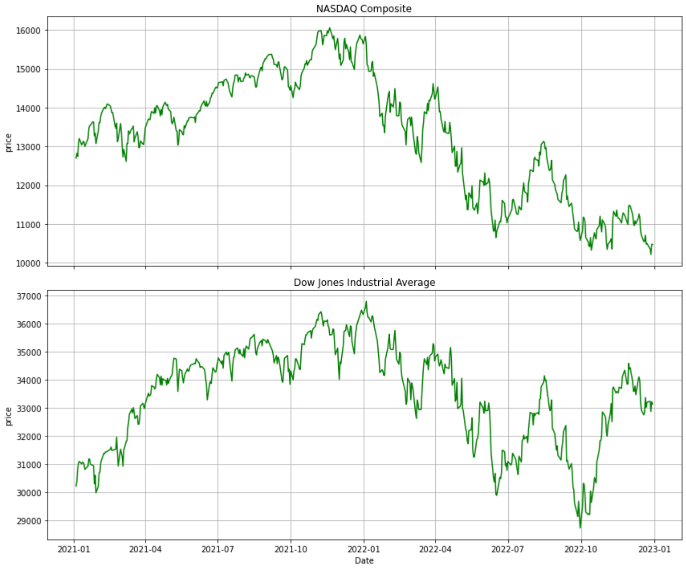


Figure 6.6: From above: NASDAQ Composite and Dow Jones Industrial Average adjusted close price 2021-2022

Table 6.5: Descriptive Statistics of Major U.S. Stock Indices

Statistic	**Russell 2000**	**NASDAQ Composite**	**DJIA**
Count	504	504	504
Mean	2,066.43	13,303.63	33,477.47
Standard deviation	209.31	1,554.59	1,770.00
Minimum	1,649.84	10,213.29	28,725.51
25th percentile	1,866.02	11,872.63	32,134.02
Median (50%)	2,131.45	13,542.12	33,891.35
75th percentile	2,243.04	14,542.46	34,798.00
Maximum	2,442.74	16,057.44	36,799.65

The descriptive statistics reveal distinct characteristics across the three indices. Notably, the NASDAQ Composite exhibits the highest standard deviation among the three, confirming its greater volatility compared to the Dow Jones Industrial Average and the Russell 2000. This heightened variability reflects the tech-heavy composition of the NASDAQ and may suggest a higher susceptibility to speculative episodes. Conversely, the Dow Jones, with the lowest volatility, appears more stable, aligning with its traditional and value-oriented constituents. The Russell 2000 presents an intermediate profile, indicative of its focus on smaller, potentially more sensitive firms.

7 Empirical Results

Once that both the models and the data used have been described, it is possible to show how the different approaches can detect both the insurgence of bubbles and, if possible, the sharp decline after the asset reaches its peak. For each asset, the outputs of the four models will be presented, together with a summary table comparing how early (or late) each model can detect the start and the peak of the bubble.

7.1 S&P 500 Application

7.1.1 Random Coefficient Autoregression

The in-sample period goes from July 2020 to December 2020 (128 observations), while the out-of-sample goes from January 2021 to December 2022 (503 observations), embedding both the ascending and descending phases of the bubble.

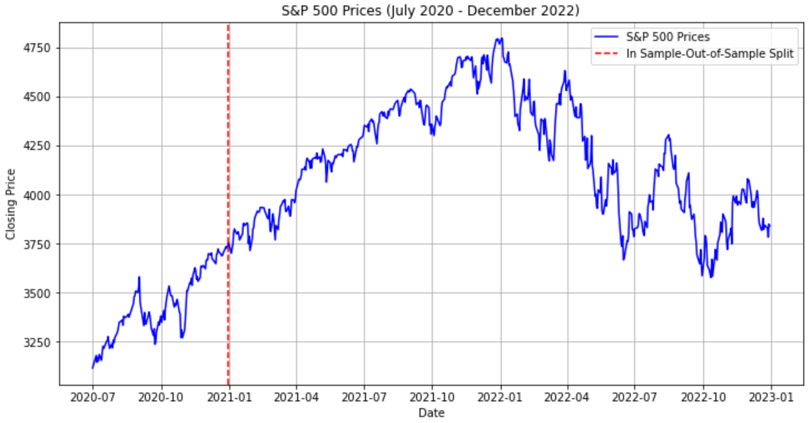


Figure 7.1: S&P 500 Daily Prices

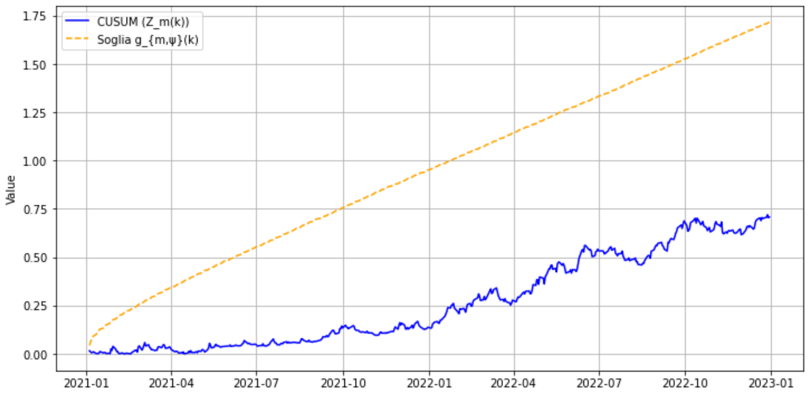


Figure 7.2: Univariate RCA. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

First, the univariate model has been tested, but no regime shift is detected. Since the weighted CUSUM does not reach the dynamic boundary function, the explosive regime is not detected, which is coherent with the results of Horvath & Trapani (2023).
Secondly, a bivariate RCA is applied to the same time series, using the VIX as exogenous covariate. The VIX is a measure of expected volatility, based on the weighted average of implied volatility of call and put OTM options of the S&P 500. In this case, the speculative regime is detected.

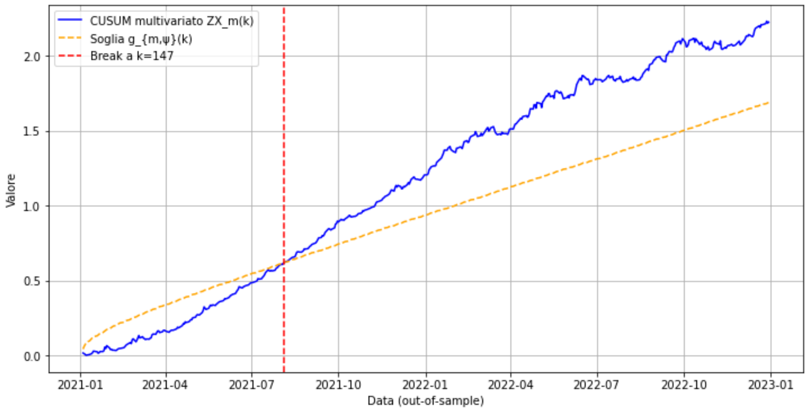


Figure 7.3: S&P 500 and VIX RCA. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

The break is detected on the 4 August 2021, 108 trading days before the peak of the bubble (31 December 2021). This would provide a useful signal for asset managers or speculators to capitalize on the upward trend or to reallocate toward a more conservative asset allocation.

In the third application of RCA the price of oil (West Texas Intermediate) is added. The detection of the speculative regime happens even earlier.

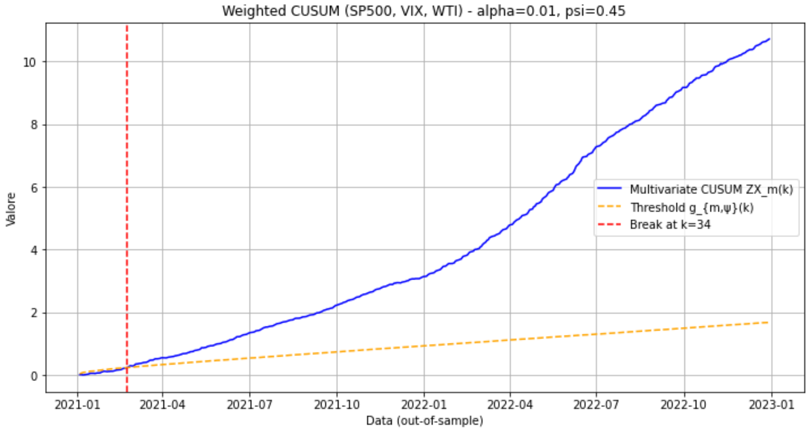


Figure 7.4: S&P 500, VIX and WTI RCA. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

The bubble is detected on 23 February 2021, almost one year before its tipping point.

The model is very accurate not only in terms of detection of the beginning of the positive phase of the bubble, i.e. the ascending phase, but can also, although with some delay, detect the shift from the positive explosive phase to the subsequent crash. At 1% significance level, using an in-sample span from April to October 2021 and an out-of-sample from November 2021 to December 2022, the shift toward the crash has been detected on 18th March 2022. This detection allows a potential speculator to sell short the S&P 500, potentially gaining almost 20% if the position is kept until the bottom is reached (20th September 2022).

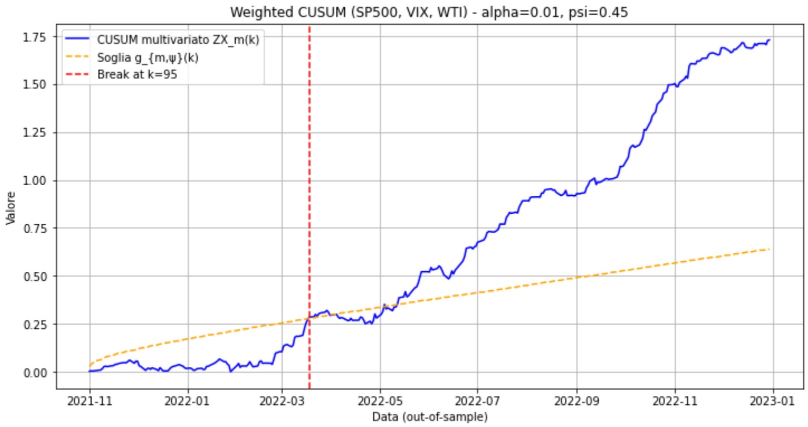


Figure 7.5: S&P 500, VIX and WTI RCA with crash detection. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

However, the results exhibit sensitivity to the choice of parameters ψ, α and $c_{\alpha,\psi}$. The estimation of the variances of the innovations $\epsilon_{i,1}$ and $\epsilon_{i,2}$, and δ is also subject to a certain degree of uncertainty, since the two types of innovations are not directly observable.

7.1.2 LPPLS model

The LPPLS is now applied to the S&P 500 daily data. The in-sample goes from May 2020 to July 2021, the out-of-sample goes from August 2021 onwards. The JLS model, however, only describes prices' behavior until the peak of the bubble is reached, but gives no information on the development of the crash. The results of the fitting is the following:

Table 7.1: LPPLS with Taboo Search, Levenberg–Marquardt and OLS

Parameter	**Value**
Taboo Search Solution (non linear)	[536.9, 0.99, 14.23586066]
Score Taboo	0.098
t_c	594.58 (16/12/2021)
m	0.998
ω	17.32
A	8.52 ($ 5047.03)
B	-0.000936
C_1	4.079×10^{-5}
C_2	2.747×10^{-5}
$\dfrac{m\,\lvert B\rvert}{\omega\,\sqrt{C_1^2+C_2^2}}$	1.088

It is important to highlight that all four conditions outlined in Chapter 3.4 are satisfied; this confirms that the conditions of the bubble are effectively developing in the in-sample data. The fitting of the model on S&P 500 gives the following result:

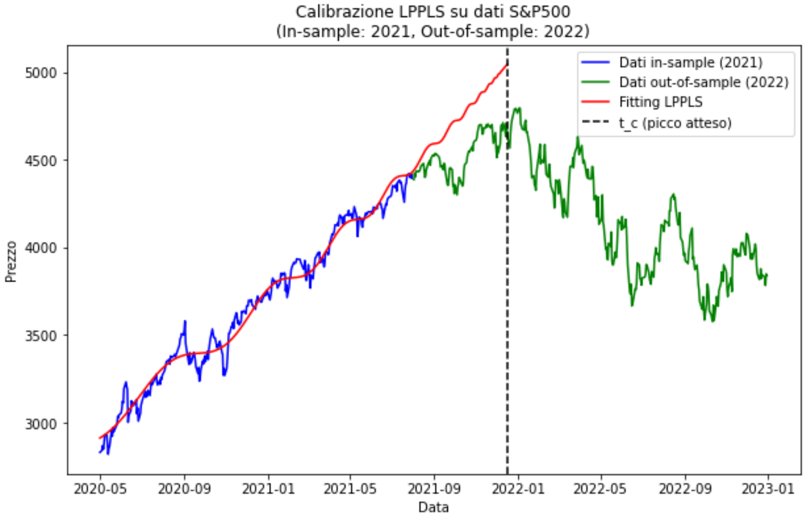


Figure 7.6: LPPLS fitted on S&P 500

The model misses the data of the tipping point by only 15 calendar days, while the real historical maximum of the bubble is about 7 % lower than what the model forecasts; nonetheless, the results demonstrate a high degree of accuracy. However, the model is quite unstable; moving the beginning or the end of the in-sample can bring completely different results, which is consistent with the lack of asymptotic properties of the model. Sornette et al. (2015) and Shu & Song (2024) have proposed a confidence indicator for the short term to detect positive bubbles in real time:

1. For a given instant t_2, different fitting windows are defined. In particular, for the short term indicator, we begin by fitting the model in an interval of 200 observation between t_1 and t_2, decreasing this interval by 5 observations each time until we reach a distance of 50 observations between t_1 and t_2. In this way, for each feasible t_2, we get 30 fitting windows.

2. The LPPLS is fitted on each of these windows

3. If the filtering conditions imposed on the Levenberg-Marquardt optimization are respected, a particular fitting is considered as bubbly. The

short-term confidence indicator is, for a given t_2, the ratio of bubbly windows over 30.

We also smooth the confidence indicator using a moving average. The result for the short-term indicator is the following:

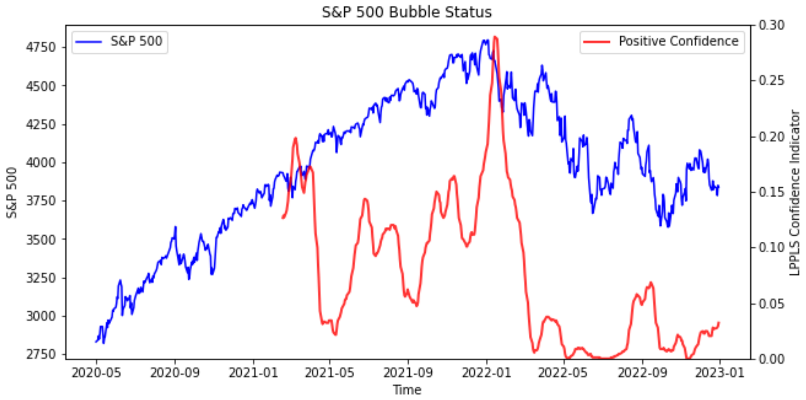


Figure 7.7: S&P 500 Short Term Confidence Indicator

Although it is quite precise in terms of detecting the peak of the bubble, which is missed only by a few weeks, this confidence indicator also shows a false signal towards the mid of 2022. Due to its ”sloppiness” and instability, the LPPLS model has been thoroughly criticized; also, the quite strict assumptions on the behavior of market operators leave no room for this model to spot an exogenously driven bubble or a bubble that is transmitted from one asset class to another. Furthermore, many studies show that the log-periodic component is statistically insignificant, as proven by Feigenbaum (2001a, 2001b) and that a Hidden Markov model can explain much more accurately and easily the shift between non-bubbly and bubbly market regimes (Chang & Feigenbaum, 2006 & 2008).

7.1.3 Topological Data Analysis

First, we simulate 3000 AR processes (1000 stationary with $\phi_1 = 0.8$, 1000 random walks and 1000 explosives with $\phi_1 = 1.1$), each with 1000 observations (300 observations are used for the in-sample, 700 for the out-of-sample). Each time series is treated as a multidimensional object by using a rolling window of three observations to mark the coordinates of each point. Then, to compute the persistence diagram, a rolling window of 5 points is used. The L1 norm is computed and, if the L1 norm surpasses the threshold of $\mu + 4\sigma$, the time series is counted as a bubble. With this methodology, 87% of explosive processes are counted as bubbles, whereas only 15% of random walks and 0.8% of stationary processes are classified as such. Therefore, TDA is able to distinguish among stationary, non stationary and non explosive, and non stationary and explosive regimes. The methodology that is now applied is adapted from Rai et al., 2024. I take adjusted close daily prices for four US stock indexes (S&P 500, Dow Jones Industrial Average, NASDAQ Composite and Russell 2000) from July 2020 to December 2020 as the in-sample (132 observations) and the years 2021-2022 as the out-of-sample (521 observations). Each point of the point-map is formed by the vector of four prices for each day, while the rolling window of points is formed by 30 days. The in-sample is used to determine μ and σ, while the thresholds to mark the structural break of the L^1 norm are $\mu + 4\sigma$,$\mu + 5\sigma$, and $\mu + 6\sigma$.

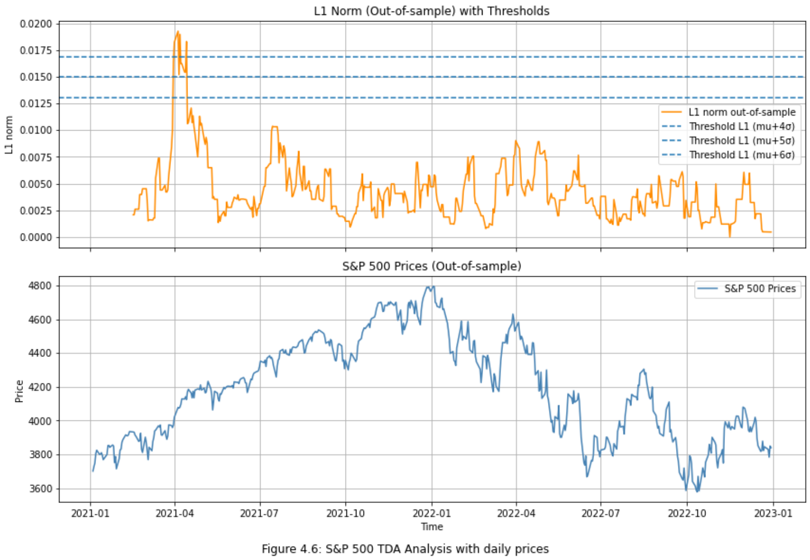


Figure 7.8: Topological Data Analysis using major US equity indices: S&P 500, Dow Jones Industrial Average, Nasdaq Composite, and Russell 2000. The L^1 norm of the persistence landscape indicates a structural break around April 2021, aligning with the early stages of the bubble. No signal is detected for the subsequent crash.

The conclusions to be drawn are as follows: the model detects a structural break as early as the beginning of April 2021, which almost coincides with the analysis of multivariate RCA in Chapter 2. However, after having marked the transition toward a speculative regime, TDA does not detect the crash, since the L^1 norm remains stationary from July 2021 onward. Instead of considering only US stock market, we can incorporate data such as the VIX and the 3-month Treasury Bill rate. Furthermore, in order to capture in a better way the fact that only a small number of sectors actually drive the returns of the S&P 500, I included in the second attempt the SPDR sectorial ETFs covering IT, financials, consumer discretionary, and utilities (since these last ones were highly hit by the spike in commodity prices), while I excluded the entire time series data of all US stock indexes. Finally, different thresholds have been used to test the ”noisy” nature of L^1 norm signals, using the same in-sample and out-of-sample. The results are much more accurate than before:

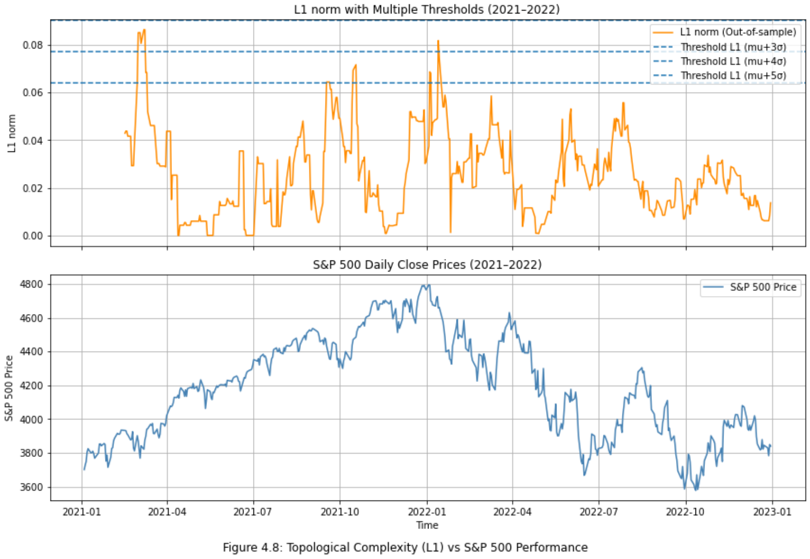


Figure 7.9: Topological Data Analysis using sectorial ETFs (XLK, XLF, XLY, XLU), VIX, and 3-month US Treasury Bills. The L^1 norm shows a structural break early in 2021 and again in early 2022, corresponding to the formation and collapse of the S&P 500 bubble.

Although the lower threshold yields mixed results, the $\mu+4\sigma$ threshold not only marks the starting point of the bubble (which is now well before April 2021) but also the starting point of the crash, which is indicated slightly after the beginning of 2022. However, this approach has also some evident flaws: since this approach is computationally intensive, it is difficult to scale. The chosen threshold is arbitrary and results could greatly vary if we choose a different one. Also, the conformation of the L^1 norms heavily depends on arbitrary choices, such as the usage of one time series or of multiple ones on the same timespan; particularly, if each point is represented by multiple prices from a single time series, instead of using the price of different sequences referring to the same date, the norm is much more noisy, leading to less easily interpretable results.

7.1.4 Permutation Entropy

Now Permutation Entropy is applied to the S&P 500.

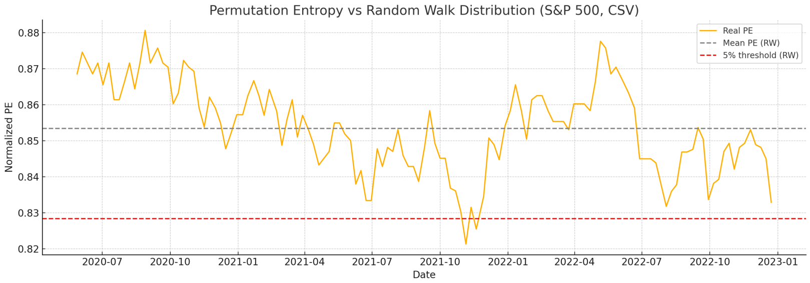


Figure 7.10: Normalized Permutation Entropy of the S&P 500 Index (2021–2022)

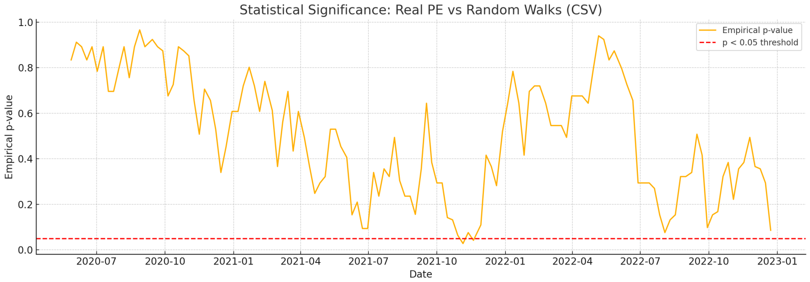


Figure 7.11: Empirical p-value of PE compared to simulated Random Walks (S&P 500)

As far as it concerns the S&P 500, the detection of the bubble only happens roughly a month before its peak.

7.1.5 Models comparison

Table 7.2: Comparison of Bubble Detection Models on the S&P 500

Method	**Bubble Onset**	**Days Before Peak**	**Crash Detected**
RCA (S&P 500 + VIX)	04 Aug 2021	108	No
RCA (S&P 500 + VIX + WTI)	23 Feb 2021	212	Yes (18 Mar 2022)
LPPLS (Fitting)	16 Dec 2021	15	No
LPPLS – Confidence Indicator	Dec 2021	20–40	Yes
TDA Setup 1 (S&P500, DJIA, Nasdaq, Russell)	01 Apr 2021	190	No
TDA Setup 2 (Sector ETFs + VIX + 3M Yield)	10 Feb 2021	220+	Yes
Permutation Entropy (S&P 500)	01 Dec 2021	22	No

The table provides a comparison of the different bubble detection methods applied to the S&P 500. While RCA and TDA setups involving exogenous variables provide earlier and often more informative signals, models like LPPLS and Permutation Entropy offer more lightweight but delayed or partial detection. The choice of method thus depends on the specific application, computational constraints, and desired robustness of the signal.

7.2 Natural Gas Futures Market Application

7.2.1 Random Coefficient Autoregression

After the analysis of the US stock market, we pass to the examination of the continuous Natural Gas NYMEX futures, to test whether CUSUM methodology is effective in the commodities futures market. The price series of the futures is split as follows, using the same span for the in-sample and out-of-sample adopted for the S&P 500.

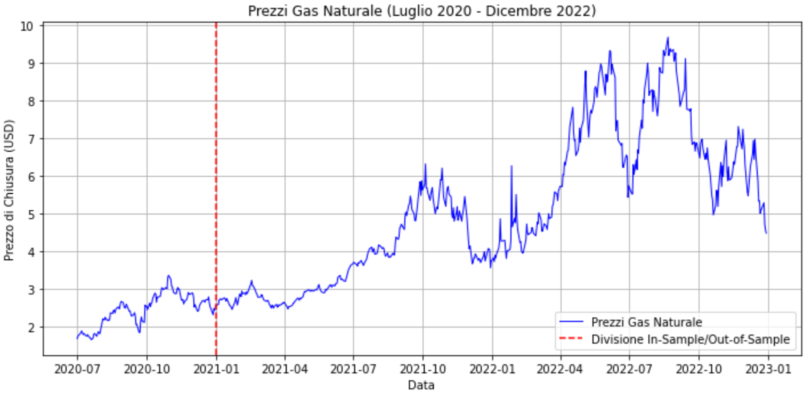


Figure 7.12: NYMEX Natural Gas daily prices

In the case of the natural gas market, the bubble is "double-peaked" (10th June and 26th August 2022), but it will only be examined whether the model catches the drop after the second peak. The univariate case does not catch the switch between the random walk of the in-sample and the explosive pattern in the out-of-sample, but if we add WTI, VIX, the Dollar Index, and the 10y Treasury yield as exogenous covariates, the results are the following.

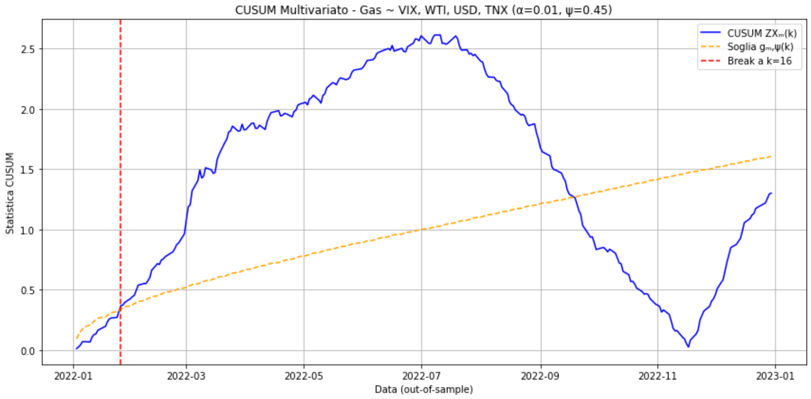


Figure 7.14: NYMEX Natural Gas decline with WTI,VIX,DIX and 10y YTM. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

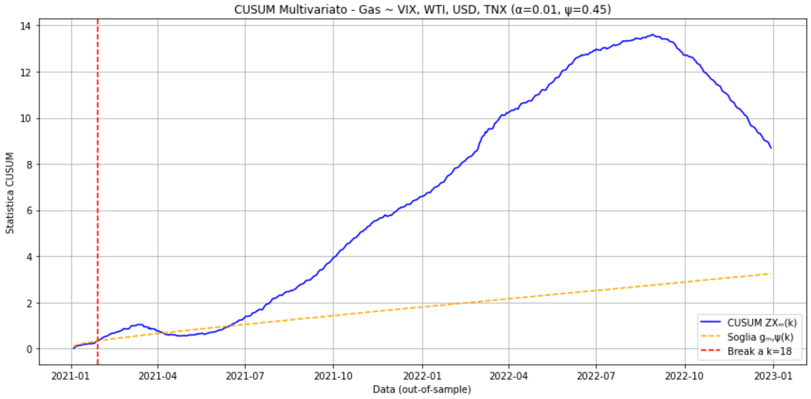


Figure 7.13: NYMEX Natural Gas with WTI, VIX, DIX and 10y YTM. The sum of the absolute value of the errors is in blue, while the dynamic threshold is in yellow

The results obtained using the CUSUM method are of particular interest, since the transition to a speculative bubble is identified at the very beginning of the out-of-sample period. Shifting the beginning of the out-of-sample window by a year, it is possible to see whether the model catches the shift between the bubble and the crash. In the case of the decline, the result is disappointing, since the detection occurs prematurely relative to the actual price decline. In conclusion, the CUSUM test statistic works better for the stock market than for the commodities future market, since the first is much more efficient, i.e. incorporates information from other asset classes, in a much faster and effective way than futures.

7.2.2 LPPLS model

The application of LPPLS to Natural Gas reveals a much less accurate fitting. Considering an in-sample going from January 2021 to April 2022, the following results were obtained.

Table 7.3: LPPLS Calibration Results for Natural Gas

Parameter	**Value**
Taboo Search Solution (non-linear)	[553.8, 0.696, 6.18586066]
Taboo Score	2.047
t_c	593.299 (17/08/2022)
m	0.6149
ω	6.917
A	1.961 ($ 7.1085)
B	-0.015
C_1	-2.916×10^{-3}
C_2	5.286×10^{-3}
$\dfrac{m\,\lvert B\rvert}{\omega\,\sqrt{C_1^2+C_2^2}}$	0.229

The fitting of the model yields the following result.

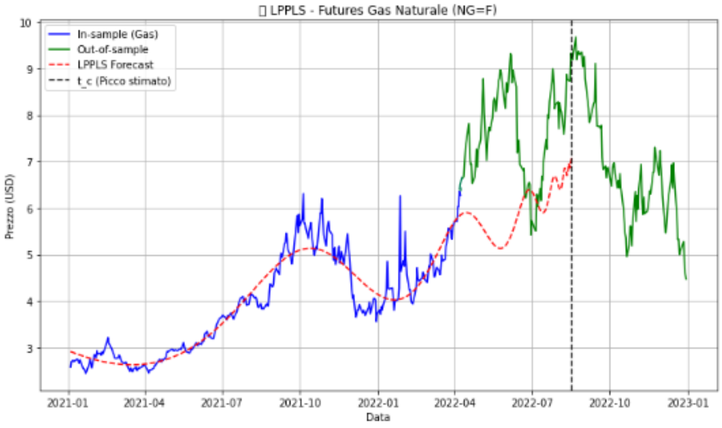


Figure 7.15: LPPLS fitting to NYMEX Natural Gas

Only the second peak is predicted, which is coherent with the assumption of the model, but the gap between the predicted and actual prices is substantial, being the actual 35% higher than the predicted. Using the LPPLS as a confidence indicator, the results appear to be more precise.

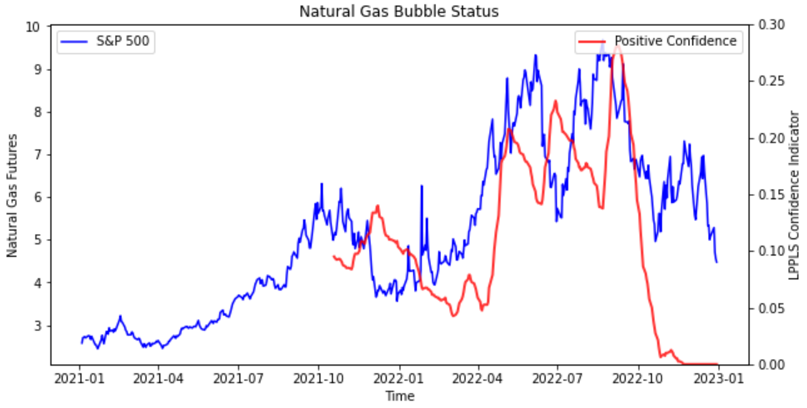


Figure 7.16: Natural Gas Short term confidence indicator

The confidence indicator is much more stable in terms of detecting potential bubbles, and can be successfully used not only to ride the bubble but also, and even more profitably, to sell short both assets.

7.2.3 Topological Data Analysis

Using only the series of prices from NYMEX, we consider a rolling window of three daily prices to get each point, and then a rolling window of 10 points to compute the persistence landscape and its L1 norm. The results, together with the threshold of extreme events, are the following.

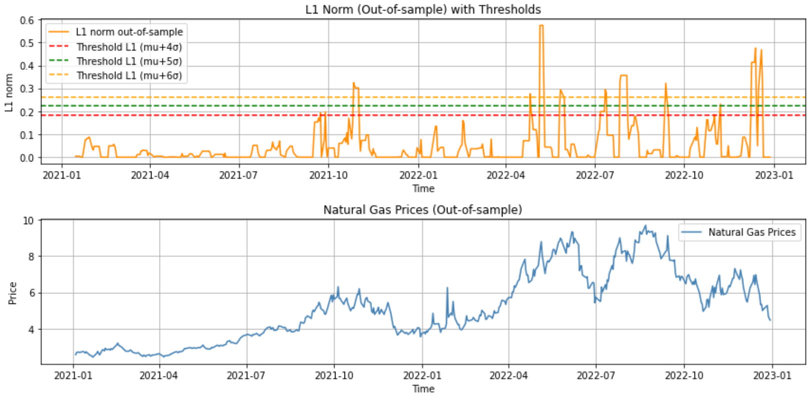


Figure 7.17: Topological Data Analysis on Natural Gas futures. The L^1 norm of the persistence landscape is shown along with the thresholds. A significant deviation occurs only near the price peak, limiting its forecasting usefulness

The analysis indicates that TDA fails to detect the shift toward a bubble unless the prices are in close proximity to the peak; therefore, the opportunity to ”ride” the explosive phase presents only a narrow window, although this indicator could be useful in a short-selling strategy. Also, the presence of a false signal at the end of 2022 deserves attention. Therefore, we can conclude that TDA is less efficient as a bubble indicator for the natural gas market than in the case of US stock markets. This could be explained by the fact that the S&P 500 is driven by few fundamental long-term factors, such as the leverage in the corporate sector, the growth of total factor productivity and the general openness of the global trade system, while the natural gas futures market depends also on short-term factors, such as climate shocks, seasonality and OPEC meetings, since oil-exporting countries are almost always also the major exporters of natural gas.

7.2.4 Permutation Entropy

The application of Permutation Entropy to the NYMEX futures yields the following result.

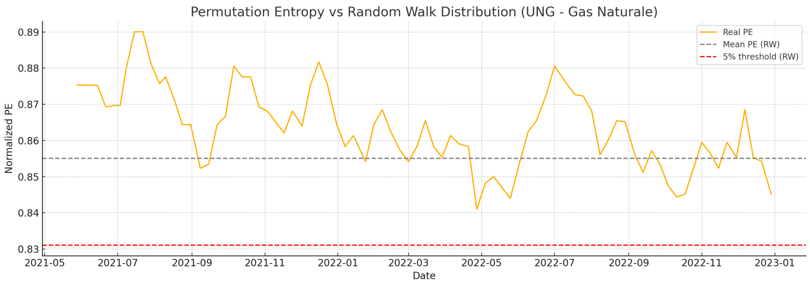


Figure 7.18: Normalized Permutation Entropy of Natural Gas Futures (2021–2022)

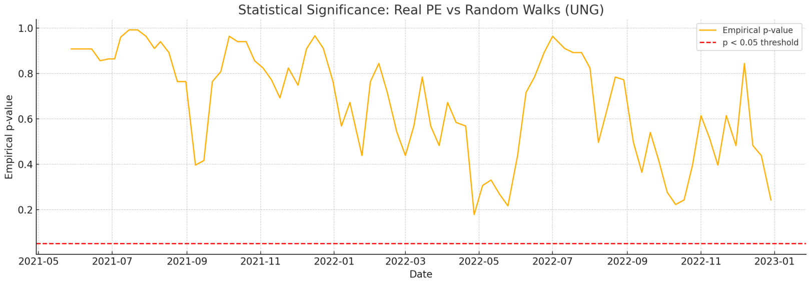


Figure 7.19: Empirical p-value of PE compared to simulated Random Walks (Natural Gas)

Regarding the natural gas, no bubbly pattern is detected.

7.2.5 Models comparison

Table 7.4: Comparison of Bubble Detection Models on Natural Gas Futures

Model	**Onset**	**Days Before Peak**	**Crash Detected**
RCA (Univariate)	–	–	No
RCA (NG + WTI + VIX + DXY + 10Y)	04 Jan 2021	599	No
LPPLS (JLS)	17 Aug 2022	9	No
LPPLS (Conf. Ind.)	mid-2022	30–60	Yes
TDA (Daily Prices)	Aug 2022	0	No
PE (Permutation Entropy)	–	–	No

This summary shows that the RCA model using a full macro setup is the only method detecting the bubble onset significantly in advance, though it fails to capture the collapse. The LPPLS fitting is weak, but its confidence indicator improves performance. TDA and PE display limited predictive power in the natural gas futures market, likely due to its higher exposure to exogenous and seasonal shocks.

8 Conclusion

It has been proven that bubbles can be detected before they crash using different techniques, both econometric and non-econometric ones, contrary to what is stated by the EMH. The cycles of credit and debt, whose importance was claimed by Minsky, are not really necessary to analyze in order to detect the majority of endogenous bubbles. In fact, most of the assumptions of behavioral finance, particularly the presence of feedback loops and the 'cluster' behavior of traders, have been proven correct. Also, these features of the market can, by changing their structural dynamics, cause a bubble on their own, without having to process extraordinarily negative information coming public. Particularly, the RCA model, conjugated with the CUSUM test statistics, has proven to be not only the most accurate method in terms of detecting both the bubble and the crash, but also the most stable with respect to data and parameters modifications. The LPPLS, in its confidence-indicator usage, has been proven accurate, but somehow noisy, with respect to the previous model. Topological Data Analysis approach yields mixed results: the noisy nature of the L^1 makes it difficult to be reliably used as a single tool to detect bubbles, although it may offer additional insights when combined with more statistically rigorous approaches. In conclusion, permutation entropy appears to be the least effective among the examined approaches.

9 Acknowledgements

I would like to thank LUISS Guido Carli University for the wonderful learning experience I have lived, both in my bachelor and in my master's degree. I would also like to thank professor Paolo Santucci de Magistris for the help he has given me during the process of writing the thesis and for the passion he has passed on to me towards statistics, econometrics and quantitative finance. I would also like to thank my parents and my aunt for the economic and emotional support they have given me during this tough, and at the same time rewarding, journey.

Bibliography

[Fama et al., 1969] Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). *The Adjustment of Stock Prices to New Information.* International Economic Review, 10(1), 1–21.

[Fama, 1970] Fama, E. F. (1970). *Efficient Capital Markets: A Review of Theory and Empirical Work.* The Journal of Finance, 25(2), 383–417.

[Fama, 1976] Fama, E. F. (1976). *Foundations of Finance: Portfolio Decisions and Securities Prices.* Basic Books.

[Jensen, 1978] Jensen, M. C. (1978). *Some Anomalous Evidence Regarding Market Efficiency.* Journal of Financial Economics, 6(2–3), 95–101.

[Lo & MacKinlay, 1988] Lo, A. W., & MacKinlay, A. C. (1988). *Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test.* The Review of Financial Studies, 1(1), 41–66.

[Brock, 1982] Brock, W. A. (1982). *Asset Prices in a Production Economy.* In J. J. McCall (Ed.), The Economics of Information and Uncertainty (pp. 1–46). University of Chicago Press.

[Tirole, 1982] Tirole, J. (1982). *On the Possibility of Speculation under Rational Expectations.* Econometrica, 50(5), 1163–1181.

[Blanchard & Watson, 1982] Blanchard, O. J., & Watson, M. W. (1982). *Bubbles, Rational Expectations and Financial Markets.* In P. Wachtel (Ed.), Crises in the Economic and Financial Structure (pp. 295–315). Lexington Books.

[Diba & Grossman, 1982] Diba, B. T., & Grossman, H. I. (1988). *Explosive Rational Bubbles in Stock Prices?* The American Economic Review, 78(3), 520–530.

[Lucas, 1978] Lucas, R. E. Jr. (1978). *Asset Prices in an Exchange Economy.* Econometrica, 46(6), 1429–1445.

[Bem, 1965] Bem, D. J. (1965). *An Experimental Analysis of Self-Persuasion.* Journal of Experimental Social Psychology, 1(3), 199–218.

[Kahneman & Tversky, 1979] Kahneman, D., & Tversky, A. (1979). *Prospect Theory: An Analysis of Decision under Risk.* Econometrica, 47(2), 263–291.

[Grossman & Stiglitz, 1980] Grossman, S. J., & Stiglitz, J. E. (1980). *On the Impossibility of Informationally Efficient Markets.* The American Economic Review, 70(3), 393–408.

[De Bondt & Thaler, 1985] De Bondt, W. F. M., & Thaler, R. (1985). *Does the Stock Market Overreact?* The Journal of Finance, 40(3), 793–805.

[Shiller, 1990] Shiller, R. J. (1990). *Speculative Prices and Popular Models.* Journal of Economic Perspectives, 4(2), 55–65.

[De Long et al., 1990] De Long, J. B., Shleifer, A., Summers, L. H., & Waldmann, R. J. (1990). *Noise Trader Risk in Financial Markets.* Journal of Political Economy, 98(4), 703–738.

[Jegadeesh, 1993] Jegadeesh, N., & Titman, S. (1993). *Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency.* The Journal of Finance, 48(1), 65–91.

[Hirshleifer & Subrahmanyam, 1998] Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). *Investor Psychology and Security Market Under- and Overreactions.* The Journal of Finance, 53(6), 1839–1885.

[Shiller, 2003] Shiller, R. J. (2003). *From Efficient Markets Theory to Behavioral Finance.* The Journal of Economic Perspectives, 17(1), 83–104.

[Soros, 2014] Soros, G. (2014). *Fallibility, Reflexivity and the Human Uncertainty Principle.* Journal of Economic Methodology.

[Minsky, 1992] Minsky, H. P. (1992). *The Financial Instability Hypothesis.* Levy Economics Institute, Working Paper No. 74.

[Phillips et al., 2011] Phillips, P. C. B., Wu, Y., & Yu, J. (2011). *Explosive Behavior in the 1990s Nasdaq: When Did Exuberance Escalate Asset Values?* International Economic Review, 52(1), 201–226.

[Genoni et al., 2023] Genoni, G., Quatto, P., & Vacca, G. (2023). *Dating Financial Bubbles via Online Multiple Testing Procedures.* Finance Research Letters, 58, 104238.

[Whitehouse et al., 2023] Whitehouse, E. J., Harvey, D. I., & Leybourne, S. J. (2023). *Real-Time Monitoring of Bubbles and Crashes.* Oxford Bulletin of Economics and Statistics, 85(3).

[Phillips et al., 2015a] Phillips, P. C. B., Shi, S., & Yu, J. (2015). *Testing for Multiple Bubbles: Historical Episodes of Exuberance and Collapse in the S&P 500.* International Economic Review, 56(4), 1043–1078.

[Phillips et al., 2015b] Phillips, P. C. B., Shi, S., & Yu, J. (2015). *Testing for Multiple Bubbles: Limit Theory of Real-Time Detectors.* International Economic Review, 56(4), 1079–1134.

[Phillips et al., 2011b] Phillips, P. C. B., Shi, S., & Yu, J. (2011). *Testing for Multiple Bubbles.* Social Science Electronic Publishing.

[Horváth & Trapani, 2022] Horváth, L., & Trapani, L. (2022). *Change-point Detection in Heteroscedastic Random Coefficient Autoregressive Models.* Journal of Business & Economic Statistics, 41(4), 1300–1314.

[Horváth & Trapani, 2023] Horváth, L., & Trapani, L. (2023). *Real-Time Monitoring with RCA Models.* arXiv preprint arXiv:2312.11710.

[Fremdt, 2015] Fremdt, S. (2015). *Detection of Changes in Dependent Data Using Two-Sample U-Statistics.* Electronic Journal of Statistics, 9(2), 3084–3114.

[Kirch & Stoehr, 2022a] Kirch, C., & Stoehr, C. (2022). *A New Approach to Page-CUSUM for Open-End Scenario for Structural Break Detection in Time Series.* Journal of Time Series Analysis, 43(4), 637–662.

[Kirch & Stoehr, 2022b] Kirch, C., & Stoehr, C. (2022). *Sequential Change Point Detection in High-Dimensional Time Series.* Electronic Journal of Statistics, 16(2), 5200–5243.

[Astill et al., 2023] Astill, S., Taylor, A. R., Kellard, N., & Korkos, I. (2023). *Using Covariates to Improve the Efficacy of Univariate Bubble Detection Methods.* Journal of Empirical Finance, 70, 342–366.

[Cvijovic & Klinowski, 1995] Cvijovic, D., & Klinowski, J. (1995). *Taboo Search: An Approach to the Multiple Minima Problem.* Science, 267(5198), 664–666.

[Sornette & Johnasen, 1997] Sornette, D., & Johansen, A. (1997). *Large Financial Crashes.* Physica A, 245(3–4), 411–422.

[Johansen, Ledoit & Sornette, 2000] Johansen, A., Ledoit, O., & Sornette, D. (2000). *Crashes as Critical Points.* International Journal of Theoretical and Applied Finance, 3(2), 219–255.

[Filimonov & Sornette, 2013] Filimonov, V., & Sornette, D. (2013). *A Stable and Robust Calibration Scheme of the Log-Periodic Power Law Model.* Physica A, 392(17), 3698–3707.

[Jiang et al., 2010] Jiang, Z.-Q., Zhou, W.-X., Sornette, D., Woodard, R., Bastiaensen, K., & Cauwels, P. (2010). *Bubble Diagnosis and Prediction of the 2005–2007 and 2008–2009 Chinese Stock Market Bubbles.* Journal of Economic Behavior & Organization, 74(3), 149–162.

[Zhou et al., 2008] Zhou, W.-X., Sornette, D., Hill, R. A., & Dunbar, R. I. M. (2008). *Discrete Hierarchical Organization of Social Group Sizes.* Proceedings of the Royal Society B, 275(1636), 265–271.

[Zhou & Sornette, 2003] Zhou, W.-X., & Sornette, D. (2003). *Evidence of a Worldwide Stock Market Log-Periodic Anti-Bubble Since Mid-2000.* Physica A, 330(3–4), 543–583.

[Sornette et al., 2009] Sornette, D., Woodard, R., & Zhou, W.-X. (2009). *The 2006–2008 Oil Bubble: Evidence of Speculation, and Prediction.* Physica A, 388(8), 1571–1576.

[Shu & Song, 2024] Shu, M., & Song, R. (2024). *Detection of Financial Bubbles Using a Log-Periodic Power Law Singularity (LPPLS) Model.* Working Paper.

[Shu & Zhu, 2020] Shu, M., & Zhu, W. (2020). *Detection of Chinese Stock Market Bubbles with LPPLS Confidence Indicator.* Physica A, 557.

[Feigenbaum, 2001a] Feigenbaum, J. A. (2001). *More on a Statistical Analysis of Log-Periodic Precursors to Financial Crashes.* Quantitative Finance, 1(5), 527–532.

[Feigenbaum, 2001b] Feigenbaum, J. A. (2001). *A Statistical Analysis of Log-Periodicity in the S&P 500 Index from 1980 to 2000.* Physica A, 294(3–4), 465–474.

[Chang & Feigenbaum, 2006] Chang, G., & Feigenbaum, J. A. (2006). *A Bayesian Analysis of Log-Periodic Precursors to Financial Crashes.* Quantitative Finance, 6, 15–36.

[Chang & Feigenbaum, 2008] Chang, G., & Feigenbaum, J. A. (2008). *Detecting Log-Periodicity in a Regime-Switching Model of Stock Returns.* Quantitative Finance, 8, 723–738.

[Scheffer et al., 2009] Scheffer, M., et al. (2009). *Early-Warning Signals for Critical Transitions.* Nature, 461(7260), 53–59.

[Lenton, 2011] Lenton, T. M. (2011). *Early Warning of Climate Tipping Points.* Nature Climate Change, 1(4), 201–209.

[Thompson & Sieber, 2011] Thompson, J. M. T., & Sieber, J. (2011). *Climate Tipping as a Noisy Bifurcation: A Predictive Technique.* IMA Journal of Applied Mathematics, 76(1), 27–46.

[Gidea & Katz, 2017] Gidea, M., & Katz, Y. (2017). *Topological Data Analysis of Financial Time Series: Landscapes of Crashes.* Working Paper.

[Gidea et al., 2018] Gidea, M., Goldsmith, D., Katz, Y., Roldan, P., & Shmalo, Y. (2018). *Topological Recognition of Critical Transitions in Time Series of Cryptocurrencies.* Working Paper.

[Akingbade et al., 2023] Akingbade, S. W., Gidea, M., Manzi, M., & Nateghi, V. (2023). *Why Topological Data Analysis Detects Financial Bubbles?* Working Paper.

[Rai et al., 2024] Rai, A., Sharma, B. N., Luwang, S. R., Nurujjaman, M., & Majhi, S. (2024). *Identifying Extreme Events in the Stock Market: A Topological Data Analysis.* Working Paper.

[Song et al., 2022] Song, R., Shu, M., & Zhu, W. (2022). *The 2020 Global Stock Market Crash: Endogenous or Exogenous?* Physica A, 585, 126425.

[Bandt & Pompe, 2002] Bandt, C., & Pompe, B. (2002). *Permutation Entropy: A Natural Complexity Measure for Time Series.* Physical Review Letters, 88(17), 174102.

[Hou et al., 2023] Hou, Y., Liu, F., Gao, J., Cheng, C., & Song, C. (2023). *Characterizing Complexity Changes in Chinese Stock Markets by Permutation Entropy.* Entropy, 25(3), 514.