**Title**:

Machine Learning for detection and prediction of stock market bubbles.

**Student Name:**

Edmundo Bernardo Paul Fernandes

**Student ID:**

SBS23034

**GitHub:** [Young-Jedi79/Capstone: Capstone Project (github.com)](https://github.com/Young-Jedi79/Capstone)

A Thesis Submitted in Partial Fulfilment

of the requirements for the

Degree of

Master of Science in Data Analytics



February 2024

**Supervisor:** David McQuaid

**Abstract**:

This study deals with economic bubbles and analyses their possible causes and tools for the prediction of such bubble’s development. An economic bubble is the commonly used term for an economic cycle that is characterized by a rapid expansion followed by a dramatic crash. While some bubbles happen naturally as a part of the economic cycle, some also occur because of investor exuberance and serve as correctives. These typically happen in securities, stock markets, real estate, and various other business sectors because of certain changes in the way key players conduct business. The well-known and widely discussed bubbles in asset markets were analysed and compared trying to deﬁne the main features, causes and signals of such bubble’s creation: S&P 500, NDX, SSE, etc.

This study focuses on developed bubble detection methods namely, Time Series Analysis, Neural Networks, LSTM, Topological Data Analysis, Backward sup Augmented Dickey-Fuller (BSADF) method and Generalized Autoregressive Conditional Heteroskedasticity (GARCH). These tests are predominantly advanced stationarity and as such, based on the fundamental factors of the assets but rather focus on the time series of the asset prices. Nonetheless, many results are still inconclusive, and bubbles remain an interesting avenue for further research.

**Acknowledgments**

I would like to express my gratitude to James Garza and Russel Rhoads, for their support, availability, and trust, and for passing on their knowledge towards the subject. I would like to express my sincere gratitude to my tutor David McQuaid for his assistance and support for this master thesis. I would also like to thank all the teaching staff of CCT College for the quality of the lectures they delivered throughout this last 3 years. My last thanks go to my friends and family who helped and supported me not only throughout this work but through my whole studies.

**Summary**

**1.** [**Research Question and motivation**](#ResearchQuestionandmotivation)**…………..………………………………………..……………………...6**

1.1 [Motivation](#Motivation)………………………………………………………………………..…………………………………………..6

1.2 [Research Question](#ResearchQuestion)………………………….…………………………………..…………………………………….….6

1.3 [Primary Objectives](#PrimaryObjectives)……………………..……………………………………………..………………………………….7

1.4 [Secondary Objectives](#SecondaryObjectives)……….………….……………………………………..…………………………………………8

**2** **[Research Ethics](#ResearchEthics)……………….……………………………………………………….……………………………………8**

2.1 [Legal Disclaimer](#LegalDisclaimer)….…………………………………………………………………….……………………………………9

**3** [**Introduction**](#Introduction)**………………………………………………………………………………………………………………..9**

3.1 [Theoretical Bubble Models](#TheoreticalBubbleModels)…………………………………………………………………………….……………10

3.2 [Rational Bubble Models with Symmetric Information](#RationalBubbleModelswithSymmetricIn)…………………………………………………10

3.3 [Rational Bubble Models with Asymmetric Information](#RationalBubbleModelswithAsymmetricI)……………………………………….……….11

3.4 [Intrinsic Bubble Models](#IntrinsicBubbleModels)………………………………………………………………………………………………..12

3.5 [Stock Market Crash](#StockMarketCrash)……………………………………………………………………………………………………….12

**4** [**Relevance**](#Relevance)**…………………………………………………………………………………………………………………….12**

**5** [**Contribution**](#Contribution)**………………………………………………………………………………………………………………..14**

**6** [**Methodology**](#Methodology)**……………………………………………………………………………………………………………….14**

**7** [**Literature Review**](#LiteratureReview)**…………………………………………………………………………………………………………15**

7.1 [Empirical Bubble Literature](#EmpiricalBubbleLiterature)…………………………………………………………………………………………..24

7.2 [Literature Review on Detecting Financial Bubbles](#LiteratureReviewonDetectingFinancial)…………………………………………………………25

**8** [**Evaluation and justification of predictive models**](#Evaluationandjustificationofpredicti)**…………………………………………………………31**

8.1 [Log Periodic Power Law](#LogPeriodicPowerLaw)……………………………………………………………….……………………………….32

8.2 [Topological Data Analysis](#TopologicalDataAnalysis)……………………………………………………………………………………………..38

8.3 [Random Forrest Classifier](#RandomForrestClassifier)……………………………………………………………………………………………..44

8.4 [Backward Supremum Augmented Dickey-Fuller (BSADF)](#BackwardSupremumAugmentedDickeyFulle)………………………………………………46

8.5 [Generalized Autoregressive Conditional Heteroskedasticity (GARCH)](#GeneralizedAutoregressiveConditionalH)…………………………..49

**9** [**Conclusion**](#Conclusion)**……………………………………………………………………………………………………………………51**

**10** [**Future Research**](#FutureResearch)**…………………………………………………………………………………………………………52**

[**References**](#References)**……………………………………………………………………………………………………………………..53**

**List of Tables**

[Table 1: Theoretical Bubble Models](#Table1)…………………………………………………………………………..………10

[Table 2. Critical values of the SADF and GSADF tests against an explosive alternative](#Table2)………..28

[Table 3. Weibull distribution to drawdowns](#Table3)……………………………………………………………………….33

[Table 4. Weibull outliers](#Table4)………………………………………………………………………………………………….…35

[Table 5. First Derivative of the S&P 500 Index](#Table5)………………………………………………………………….…40

[Table 6. Crash Detections](#Table6)……………………………………………………………………………………………………40

[Table 7. Point clouds from two successive windows and their associated persistence diagram](#Table7)……………………..……………………………………………………………………………………………………...41

[Table 8. Landscape Distances](#Table8)……………………………………………………………………………………….……42

[Table 9. Betti Curve Distances of the S&P 500 Index](#Table9)…………………………………………………..………43

[Table 10. Crash Comparisons](#Table10)………………………………………………………………………………………………43

[Table 11. BSADF result](#Table11)………………………………………………………………………………………………………..48

[Table 12. GARCH model](#Table12)………………………………………………………………………………………………………50

**List of Figures**

[Figure 1: autoregressive specification for xt](#autoregressivespecificatiofor)……………………………………………………………………………………26

[Figure 2: recursive test statistic ADFr](#recursiveteststatisticADFr)……………………………………………………………………………………………….26

[Figure 3: recursive test statistic ADFr(continued](#recursiveteststatisticADFr2))…………………………………………………………………………….26

[Figure 4: The martingale null with an asymptotic drift](#Themartingale1)……………………………………………………………………27

[Figure 5: The martingale null with an asymptotic drift alternative hypothesis](#Themartingale2)………………………………..27

[Figure 6: Regression model](#Regressionmodel)………………………………………………………………………………………………………………27

[Figure 7: Asymptotic distribution](#Asymptoticdistribution1)…………………………………………………………………………………………………….28

[Figure 8: Asymptotic distribution](#Asymptoticdistribution2)…………………………………………………………………………………………………….28

[Figure 9: first crossing time](#Figure9)…………………………………………………………………………………………………………….28

[Figure 10: first crossing time (continued](#firstcrossingtime2))…………………………………………………………………………………………28

[Figure 11 ADF statistic sequence](#ADFstatisticsequence)…………………………………………………………………………………………………….29

[Figure 12 critical value cvBT](#cvBT)……………………………………………………………………………………………………………30

[Figure 13 critical value scvBT](#criticalvaluescvBT)………………………………………………………………………………………………………….30

[Figure 14 Localizing Parameter](#localizingparameter)…………………………………………………………………………………………30

1. **Research Question and motivation.**

**1.1 Motivation**

Stock market prediction is basically defined as trying to determine the stock value and offer a robust idea for the people to know and predict the market and the stock prices. It is generally presented using the quarterly financial ratio using the dataset.

Thus, relying on a single dataset may not be sufficient for the prediction and can give a result which is inaccurate. Hence, this research contemplates towards the study of machine learning with various datasets integration to predict the market and the stock trends.

The problem with estimating the stock price will remain a problem if a better stock market prediction algorithm is not proposed. Predicting how the stock market will perform is quite difficult. The movement in the stock market is usually determined by many factors which will be thoroughly explored in this paper.

* 1. **Research Question**

The hypothesis posits that machine learning models, when applied to the analysis of stock market, have the potential to provide valuable insights and predictive capabilities. It is proposed that by leveraging advanced machine learning techniques, historical market data, It is possible to develop models that not only elucidate past patterns but also make accurate predictions about stock market bubbles.

This hypothesis assumes that machine learning models, such as regression, classification, and clustering algorithms, can effectively capture the complex relationships between various influencing factors and stock market. The proposal aims to test whether these models can outperform traditional financial analysis methods in terms of accuracy, efficiency, and robustness, while capturing the complex, nonlinear relationships inherent in financial data, contributing to a more nuanced understanding of stock market bubbles.

The hypothesis further suggests that the predictive capabilities of machine models can lead to more informed decision-making.

Investors, financial analysts, and policymakers can benefit from the insights generated by these models, allowing them to proactively respond to market trends, and external economic factors.

Ultimately, this proposal hypothesis seeks to test the potential of machine learning in revolutionizing the understanding and identifying and predicting stock market investor bubbles, thereby improving investment strategies and outcomes.

Machine learning algorithms can be used to identify patterns in historical stock market data to predict future trends, identify influential factors that can drive stock market movements, such as economic indicators and central bank policy decisions.

Once these patterns, and influential factors have been identified, machine learning will be used to develop predictive models that will be used to inform investment decision making, in addition to improving investment performance, the models will also be used to reduce market volatility, for example using the model to develop strategies that can help to reduce risk and smooth out returns. By harnessing the computational power of these algorithms, the research aims to uncover patterns, trends, and anomalies in stock markets that might be challenging to identify through more traditional methods.

**1.3** **Primary Objectives**

This objective of this study is to create robust and reliable machine learning model algorithms that can accurately identify periods of excessive exuberance in the stock market,

indicating the presence of a bubble. The model should utilize a combination of technical and fundamental indicators to capture the complex dynamics of stock prices and identify bubble-like patterns.

Evaluate the performance of machine learning algorithms for bubble detection and prediction.

This section involves evaluating the accuracy and effectiveness of various machine learning algorithms in detecting and predicting stock market bubbles. The study compare’s the performance of different algorithms, such as support vector machines, random forests, and artificial neural networks, across various metrics, including sensitivity, specificity, and forecasting accuracy.

Analysing the factors that contribute to the formation and collapse of stock market bubbles: This study theoretically explores the relationships between economic indicators and behavioural biases to identify the key factors that contribute to irrational exuberance and the subsequent correction of asset price.

**1.4** **Secondary Objectives**

**To achieve the primary objective, the following objectives must be completed: Develop a real-time bubble detection and prediction system by** creating a practical system that can monitor stock market data and provide real-time alerts or signals indicating the potential presence of bubbles. The system should integrate the machine learning models developed in the primary objectives and utilize efficient data processing techniques to provide timely insights. **Conduct cross-market analysis of bubble detection and prediction by** expanding the scope of the study to evaluate the performance of machine learning algorithms across different stock markets, such as the U.S, European, and Asian markets. By **exploring the role of machine learning in risk management and financial regulation, t**he secondary objective examines the potential application of machine learning in enhancing risk management practices and improving financial regulations and discuss how machine learning models can be used to identify early warning signs of bubbles, assess systemic risk, and guide regulatory actions to mitigate the impact of bubbles on financial stability.

1. **Research Ethics**

Research ethics are foundational principles that guide the conduct of scientific inquiry, ensuring the integrity, dignity, and rights of all individuals involved in the research process. Participants were informed about the purpose, procedures, potential risks, and benefits of the study,

and their voluntary informed consent was obtained. Confidentiality and anonymity were maintained throughout data collection, storage, and analysis, with measures implemented to safeguard sensitive information. Furthermore, any potential conflicts of interest were disclosed and managed transparently. The principles of honesty, integrity, respect, and accountability have guided every aspect of this research endeavour, ensuring its validity, reliability, and ethical soundness. This study adheres strictly to ethical guidelines established by relevant regulatory bodies and academic institutions. All data collection procedures prioritize the privacy and confidentiality of individuals and organizations involved, ensuring compliance with data protection regulations. Furthermore, the research maintains transparency and integrity in reporting findings, avoiding any form of bias or manipulation of results. Ethical considerations also extend to the responsible dissemination of research outcomes, with proper citation and acknowledgment of sources to uphold academic integrity.

* 1. **Legal Disclaimer**

Yahoo!, Y!Finance, and Yahoo! finance are registered trademarks of Yahoo, Inc.

yfinance is not affiliated, endorsed, or vetted by Yahoo, Inc. It's an open-source tool that uses Yahoo's publicly available APIs, and is intended for research and educational purposes.

You should refer to Yahoo!'s terms of use ([here](https://policies.yahoo.com/us/en/yahoo/terms/product-atos/apiforydn/index.htm), [here](https://legal.yahoo.com/us/en/yahoo/terms/otos/index.html), and [here](https://policies.yahoo.com/us/en/yahoo/terms/index.htm)) for details on your rights to use the actual data downloaded. Remember - the Yahoo! finance API is intended for personal use only.

**3** **Introduction**

The stock market serves as a critical driver of investment, capital allocation, and economic growth in modern economies. Through history, bubbles have emerged in the stock market, presenting substantial risks to financial stability and investor confidence. A stock market bubble is characterized as an asset bubble or a speculative bubble, which is when prices for a stock or an asset rise exponentially over a period of time, well in excess of its intrinsic value. Eventually prices hit a wall and consequently rapidly crash as the bubble pops. As a result, understanding how to effectively manage and prevent speculative bubbles has become a vital concern for policymakers, regulators, and market participants.

A stock market bubble is driven by raw speculation. A bubble it begins to form when there is a gathering acceleration in price for an asset that far outstrips the assets intrinsic value. Meaning that investors are willing to pay more for a security or another asset.

Irrational exuberance is a phrase popularized by former Federal Reserve Chair Alan Greenspan to describe the collective enthusiasm among traders and investors that fuels rapidly increasing prices that outstrip underlying fundamentals.

This research aims to contribute to the field of managing and preventing bubbles in the stock market by investigating their formation, identifying effective prevention strategies, and evaluating their impact on public policy and market participants. A key aspect of this literature review will involve examining various predictive models proposed to identify bubbles, as well as strategies and approaches employed to prevent and correct them. By evaluating the strengths and limitations of existing methodologies, this research seeks to develop a comprehensive understanding of bubbles and refine the tools available for their identification and prediction.

* 1. **Theoretical Bubble Models**

It is necessary to understand the propagating mechanisms that can lead to explosive characteristics in asset prices and economists have developed a range of different theoretical models for this purpose.

This paper distinguishes between two broad categories of standard approaches to modelling bubbles which are rational bubbles that attempt to explain explosive behaviour in economic and financial variables in a setting in which all agents are assumed to be rational with regards to their expectations of future earnings and in which markets are predominantly efficient. The second bubble model is also referred to as irrational bubble models because they abandon rational investor behaviour to a varying degree, leading to rational traders interacting with non-rational traders which might cause the emergence of a bubble.

* 1. **Rational Bubble Models with Symmetric Information**

The first theoretical bubble models developed were rational bubble models. The classic rational bubble has a longstanding tradition in theoretical literature, with seminal papers by Samuelson (1958), Diamond (1965), Blanchard and Watson (1982), Tirole (1982, 1985) and Froot and Obstfeld (1991), to name a few.

A diagram of a model

Description automatically generated

Table 1: Theoretical Bubble Models

Rational bubble models generally build on the present-value model of equity prices, also known as the market fundamentals model or simply the standard model. In this model, an asset’s value is determined solely by the value of the discounted fundamentals. In its most basic version for stock prices, the current stock price is equal to the present value of the expected future dividends. The standard present-value model builds on four main assumptions (Gurkaynak 2008). First, all investors are rational and risk neutral. This implies that the model does not incorporate risk premia.

Second, the discount rate used to determine the present value of the future payments is constant over all periods, and third, the dividend generating process is not expected to change. The fourth assumption relates to the information available to all investors. As the name suggests, in rational bubble models with symmetric information all investors possess the same information. LeRoy and Porter et al. (1981) and Shiller et al. (1981)) have demonstrated that such simple present-value models are unable to explain fluctuations in stock prices.

Pioneers on rational bubble models with symmetric information include Blanchard (1979), Blanchard and Watson (1982), Flood and Garber (1980), and Tirole (1985). Their models build on the so-called “no bubble condition” which states that in the present-value model of equity prices the value of the bubble must be always zero. The problematic aspect of this approach is that the fundamental component cannot be directly observed so assumptions must be made to determine it. One potential approach is to assume that dividends follow a random walk with drift while the bubble process is characterized by an explosive autoregressive process (Diba and Grossman 1988). This implies that whenever a bubble is not present, the price also follows a random walk with drift. Put differently, the fundamental component in the rational bubble model converges, while the bubble component is non-stationary.

* 1. **Rational Bubble Models with Asymmetric Information**

In rational bubble models with asymmetrical information the assumption of the present-value model that all investors have the same information is relaxed. This means that investors are asymmetrically informed, but at the same time they still share a common prior distribution (Brunnermeier 2008). Asymmetric knowledge can mean that it may not be common knowledge whether there is a bubble in the asset price. Alternatively, it can also mean that not all investors know that the other investors know there is a bubble. Due to this lack of higher-order mutual knowledge bubbles can also be present in the price of a finitely lived assets under certain conditions. One condition that can induce bubbles in such a setting are short-sale restrictions. Allen et al. (1993) allows for the existence of a bubble in their model when common knowledge is absent and short-sale constraints bind. All agents know that the asset is overvalued but they do not know that other agents know this as well. Agents are willing to hold an overvalued asset because they hope to resell it at a higher price to another agent who may value the asset highly in certain states due to his information structure. Kindleberger and Aliber (2005) refer to this belief as the “greater fool theory”, because the investor is convinced, he can sell the asset to an even greater fool.

* 1. **Intrinsic Bubble Models**

Intrinsic bubbles constitute a special class of rational bubble model that was first introduced by Froot and Obstfeld (1991). Unlike in the bubble models mentioned in the subsections above, in intrinsic bubble models the bubble component is a deterministic function of the fundamentals and not a function of time. This bubble parametrization has a few advantages. First, the bubble can overreact to changes in fundamentals. Second, it does not have to explode in relation to the fundamental value, and third, it may even disappear entirely. In intrinsic bubble models, since the bubble is tied to the level of dividends, dividends are modelled explicitly. Due to this, the bubble process depends completely on the level of the dividends and does not take off on its own. When there is no bubble, prices are a linear function of dividends, and the price-dividend ratio is a constant. However, since the stock price is correlated with the fundamentals, it is more sensitive to dividend innovations than is justified by the linear pricing equation. Non-linearity in the relationship between stock prices and dividends is interpreted as an intrinsic bubble and the differences in the behaviour of the price-dividend ratio when a bubble is present compared to when it is not, is exploited to form a bubble test. Unfortunately, a non-linear relationship between stock prices and dividends can only be interpreted as a sign of bubbles because the model is assumed to be linear. If the “true” model is non-linear, this is no longer the case.

* 1. **Stock Market Crash**

A stock market crash is a sharp and quick drop in total value of a market with prices typically declining more than 10% within a few days. A crash is usually attributable to the burst of a price bubble and is due to a massive sell-off that occurs when a majority of market participants try to sell their assets at the same time.

Professor Didier Sornette who successfully predicted multiple financial crashes uses log-periodic power laws (LPPLs) to describe how price bubbles build up and burst. The LPPL fits the price movement leading up to a crash to a faster than exponentially increasing function with a log-periodic component (reflecting price volatility with increasing magnitude and frequency).

**4.****Relevance**

In the words of Shiller (1989), a crash is a time when “the investing public en masse capriciously changes its mind.” But, as with the more rational theories, this explanation again leaves unanswered the question why such tremendous capricious changes in sentiment occur. Other studies have argued that even though fundamentals appeared high in 1929, Irving Fisher (1930), for example, argued throughout 1929 and 1930 that the high level of prices in 1929 reflected an expectation that future corporate cash flows would be very high. Fisher believed this expectation to be warranted after a decade of steadily increasing earnings and dividends, rapidly improving technologies, and monetary stability. In hindsight, it has become clear that even though fundamentals appeared high in 1929, the stock market rise was clearly excessive. A recent empirical study (De Long and Shleifer, 1991)) concludes that the stocks making up the S&P500 composite were priced at least 30 percent above fundamentals in late summer, 1929. Eugene N. White (2006) suggests that the 1929 boom cannot be readily explained by fundamentals, represented by expected dividend growth or changes in the equity premium.

 A bubble begins to form when there’s a gathering acceleration in price for an asset that far outstrips the asset’s intrinsic value. That means people are willing to pay more and more for a security or another asset, above and beyond what’s expected based on things like demand, earnings, revenue, or growth potential.

Irrational exuberance is a phrase popularized by former Federal Reserve Chair Alan Greenspan to describe the collective enthusiasm among traders and investors that fuels rapidly increasing prices that outstrip underlying fundamentals. Whether you call it the crowd mentality, herd bias, the bandwagon effect, there is a self-perpetuating cycle where people want to buy an asset because its price is increasing, driving the price even higher and making even more people want to buy it.

The efficient markets hypothesis lost ground entirely after the burst of the internet bubble in 2000, providing one of the recent most striking episodes of anomalous price behaviour and volatility in the most well-developed capital markets of the world. The movement of Internet stock prices during the late 1990s was extraordinary in many respects. The Internet sector earned over 1000 percent returns on its public equity in the two-year period from early 1998 through February 2000. The valuations of these stocks began to collapse shortly thereafter and by the end of the same year, they had returned to pre-1998 levels, losing nearly 70% from the peak. The extraordinary returns of 1998 to Feb. 2000 had largely disappeared by the end of 2000.

Although in February of 2000, most internet-related companies had negative earnings, the Internet sector in U.S. was equal to 6% of the market capitalization of all U.S. public companies, and 20% of the publicly traded volume of the U.S. stock market (see, e.g., Ofek and Richardson, 2002; 2003)).

**5.** **Contribution**

The novelty of this work is the application of machine learning functions and tools to identify and predict stock market bubbles. The dataset contains historical stock data, and it has been extracted from finance.yahoo.com. As stocks are open, traded and closed daily the data collected backs to 1927. The empirical analysis is performed with Python software. The aim of this paper is to present novel tests for the detection and prediction of early causal diagnostic of positive and negative bubbles in S&P 500 historical dataset and the detection of End-of-Bubble signals with their corresponding confidence levels.

**6.** **Methodology**

**Data Collection**

This study collected historical financial data from Yahoo Finance focusing on historical stock market indices.

**Feature Engineering**

Key features were derived from the raw data to capture important characteristics associated with market bubbles. This includes computing daily, weekly, monthly, and yearly exploratory data analysis, moving averages, daily returns, volatility measures, and any other relevant indicators such as those used in Log Periodic Power Law modelling.

**Model Selection**

A diverse set of machine learning and statistical models were selected for bubble identification and prediction. Including the Log Periodic Power Law model, Topological Data Analysis techniques, Random Forest Classifier, BSADF, and GARCH model. Each model was chosen based on research of ability to capture different aspects of market behaviour and bubble dynamics.

**Model Training and Evaluation**

The selected models were trained using historical data, with appropriate validation techniques employed to assess their performance. The models were evaluated based on various metrics such as accuracy, precision, recall, F1-score, and log-likelihood, ensuring robustness and reliability in bubble detection and prediction.

**Ensemble Approaches**

Ensemble techniques were applied to explore the predictions of individual models, leveraging the strengths of each model to improve overall performance in identifying and predicting stock market bubbles.

**Cross-Validation and Sensitivity Analysis**

Cross-validation techniques were utilized to ensure the generalization of the models across different time periods and market conditions. Sensitivity analysis was conducted to assess the impact of parameter changes and model assumptions on the predictive accuracy of the models.

**Implementation and Deployment**

The final models were implemented in a practical setting, integrated into trading algorithms or risk management systems to provide monitoring and decision support in identifying and mitigating the risks associated with stock market bubbles.

**7.** **Literature Review**

A bubble is created when demand of the stock is so high that it exceeds the growth of the company. According to Robert Shiller,” a bubble is a social epidemic that involves extravagant expectations for the future”. It means that a “bubble” is inflated with utterly high hopes of investors for growth in stock market. So, when the expectations of investors do not match the growth of companies, Stock market bubble becomes a cause of grave stock market crash. Like, The Stock Market Crash of 1929 Aka Black Tuesday and stock market crash of 2008. The recent one, the China stock market crash on July 08, 2015, also share the same cause, Stock market bubble burst. Thus, it is very important to know what causes Stock market to create such bubble. Prime causes of Stock market bubble are Cognitive Bias, luring stock market growth, uneducated investors, and biased newsletters.

Financial bubbles can have profound effects on an economy and the livelihoods of its citizens. Consequences, such as financial and debt crises, as seen in the dot-com bubble of 1999-2001 and the U.S subprime mortgage crisis of 2007-2009, often originate from the abnormal escalation of asset prices in the stock market or the real estate market.

When a bubble bursts, it can lead to the collapse of major financial institutions, pushing countries to the brink of bankruptcy and triggering comprehensive financial economic crisis. Governments are forced to allocate substantial resources to bailout packages and recovery programs in the wake of such crisis.

Moreover, the social costs remain significant, and the restoration of public confidence in the market poses a particularly formidable challenge. Inexperienced investors, lacking the knowledge to manage risks, are among the most severely impacted (Galbraith et al. 2009). They often hold assets during the late stages of a bubble, making them particularly vulnerable to the adverse effects of the bubble’s burst. Therefore, for governments and market supervisors researching and forecasting the status of financial bubbles is extremely important.

Cognitive Bias is a psychological phenomenon that deviates people from taking rational decisions. Fickle-minded investors suffer from Cognitive Bias and are serious cause of inflation in stock market bubble. These investors believe in what they hear but refuse to take future analysis into consideration. For example, China stock market was clearly experiencing formation of a “bubble” in early 2015 when the demand of Chinese stock prices was extravagant. But the fickle-minded investors kept on investing in the stocks because they refused to analyse a clear bubble burst in future. Thus, the bubble burst on July 08, 2015, resulting in an all-time low in Chinese stock prices. Cognitive bias deviates investors from understanding the clear picture of stock market. Thus, it causes a stock market bubble.

At times, exorbitant stock market prices are alarm to alert investors about the risk of further investment in stock market. Luring performance of stock market is a cause of stock market bubble. It attracts the investors to make huge investments, resulting in higher growth of stocks than growth of the company. This is the phenomenon that is known as “bubble”. For example, the record-breaking performance of Chinese Stock market in 2015 resulted in the formation of Stock market bubble. It lured the investors to invest in the stocks even though the growth outperformed the growth of the company. Stock market growth looks fascinating, but it may deceive the investors. The growth of the stock market should be analysed like two-sided coin. Every aspect must be taken into consideration to prevent formation of bubble.

In theory, the concept of financial bubbles is often referred to by various terms, like asset price bubbles or speculative bubbles, and is an intriguing research topic. There are various perspectives on financial bubbles and their identification. However, attempting to classify and provide a clear-cut definition remains a controversial subject within the academic community. Generally, the classification of financial bubbles falls into two primary categories: classical bubbles and modern bubbles.

Classical bubbles are primarily driven by irrational investor behavior. Shiller (2002) posits that bubbles in the market are a psychological phenomenon. He suggests that the occurrence of these bubbles is a result of amplified feedback-trading tendencies, which are caused by the attention paid to them by news media. The reason for this is that, as more investors show interest in a particular asset, news media tend to expand their coverage of it, which in turn attracts even more potential investors. This leads to an increase in demand for the asset, which causes its price to rise, thereby attracting even more attention from the news media. This cyclical process reinforces the feedback-trading tendency in the market, ultimately leading to the occurrence of bubbles. This phenomenon is often referred to as ‘herd mentality’, and its consequences can lead to a severe market collapse, subsequently exerting a profound impact on the overall economy. Kindleberger et al. (2005) proposed an approach to understanding financial bubbles from the perspectives of irrational exuberance and psychological expectations. According to these authors, financial bubbles were created by the irrational exuberance and blind faith of investors, leading to a series of reckless investment decisions and ultimately culminating in a market collapse and asset value correction Stiglitz (1990)suggested that the phenomenon of a financial bubble occurs when investors believe that current prices already reflect high levels of expectations, and the fundamental factors supporting those prices are no longer in place. In other words, when investors believe that current prices no longer offer them the potential for future profits and this sentiment becomes widespread, a bubble begins to form. When investors have faith that the upward trend will continue and fear missing out on potential gains if they do not buy now, the bubble inflates. However, a bubble will be prone to burst when investors start to believe that prices can no longer rise further, demand wanes, and this can trigger a significant sell-off, causing prices to fall rapidly (Case and Shiller2003).

The second approach is the modern bubble, described by (Tirole 2008) as a situation in which the price of an asset exceeds its fundamental value. The fundamental value of an asset is typically based on its expected future cash flows, such as dividends, coupon payments, or rental income. According to the author, bubbles occur when investors are willing to pay a higher price for an asset that can be resold immediately than if they were obligated to hold onto it for a longer period.

This view recognizes that an asset’s perceived value is not always tied to its true value, but rather to its potential for short-term profits. In the derivatives market, a bubble is considered to exist if the market value of a derivative consistently exceeds the cost of creating similar derivatives. This means that the price at which the derivative is trading in the market is higher than the cost of creating a comparable derivative. An illustration of such a bubble can be observed in the price disparity in option pricing. Specifically, a bubble may occur when a combination of put and call options, designed to replicate the movements of a stock, is traded at a price differential compared to the price of the underlying stock. This price differential must also consider factors such as interest rates and the cost of borrowing the stock. Therefore, the existence of bubbles in various forms highlights the complexity of market dynamics and raises the challenges associated with maintaining economic stability.

Additionally, the concept of bubbles can be classified into two categories: rational bubbles and partially rational bubbles. The rational bubble theory proposes that investors knowingly purchase overpriced assets with the understanding that they can sell them at a profit in the future. This theory posits that, even when faced with prices that are clearly overvalued, expectations of future profits can drive investment behavior. In other words, investors in rational bubbles willingly engage in the bubble, motivated by the prospect of profiting from price increases before the bubble eventually bursts. Shiller (2005) introduced the concept of partially rational bubbles, which posits that stock prices are influenced by a combination of rational and irrational behavior among investors. He suggested that individual investors were prone to irrational exuberance, often driven by sensationalized media reports, but this does not mean that investors are consistently irrational or ‘crazy’. Rather, the stock market is influenced by social trends and short-term desires, which may either lead to the formation of bubbles or not. In essence, partially rational bubbles recognize that market behavior can be influenced by both rational and irrational elements that stem from societal trends and collective beliefs, and these factors contribute to the persistence of bubbles. This perspective provides a more nuanced understanding of the dynamics that drive financial bubbles.

Fama (2014), who advocates the Efficient Market Hypothesis (EMH), offers an alternative perspective on financial bubbles. According to Fama, the extreme price fluctuations of assets can be anticipated, and as a result, there are no bubbles in asset prices. While this stance is still subject to debate, it provides a framework for delving into the identification of factors or causes that could lead to the predictable formation of bubbles. Fama’s view challenges the perception that financial bubbles are irrational and uncontrollable and proposes that there may be underlying patterns and elements that can be used to predict or understand the emergence of bubbles.

This perspective has given rise to further research into the predictability of bubbles and the factors that contribute to their formation.

This study approach bubbles from the perspective of irrational bubbles, which are characterized by a sudden surge in prices within a short time frame followed by a rapid decline in the VNINDEX. Additionally, it also recognizes the presence of the external factors that influence market dynamics beyond investor behavior, as posited by Fama’s perspective.

In his recent review of the financial economic literature on bubbles, Gurkaynak (2008) reports that “foreach paper that finds evidence of bubbles, there is another one that fits the data equally well without allowing for a bubble. We are still unable to distinguish bubbles from time-varying or regime-switching fundamentals, while many small sample econometrics problems of bubble tests remain unresolved”. Similarly, the following statement by former Federal Reserve chairman Alan Greenspan (2002), at a summer conference in August 2002 organized by the Fed to try to understand the cause of the ITC bubble and its subsequent crash in 2000 and 2001, summarizes well the state of the art from the point of view of practitioners: “We, at the Federal Reserve recognized that, despite our suspicions, it was very difficult to definitively identify a bubble until after the fact, that is, when its bursting confirmed its existence. Moreover, it was far from obvious that bubbles, even if identified early, could be pre-empted short of the Central Bank inducing a substantial contraction in economic activity, the very outcome we would be seeking to avoid.” To break this stalemate, Sornette, Anders Johansen (from1995to2002), Wei-Xing Zhou (since2002 (now Professor at ECUST in Shanghai)) and the FCO group at ETH Zurich (since 2008) have developed a series of models and techniques at the boundaries between financial economics, behavioural finance, and statistical physics. Our purpose here is not to summarize the corresponding papers, which explore many different options, including rational expectation bubble models with noise traders, agent-based and references traders with Bayesian updates of their beliefs, models with mixtures of non-linear trend followers and non-linear value investors, and soon (see Sornette (2003b) and references there in for the period 2002 and the two recent reviews in Kaizoji and Sornette (in press), Sornette and Woodard(2009 and references therein). In a nutshell, bubbles are identified as “super-exponential” price processes, punctuated by bursts of negative feedback spirals of crash expectations. These works have been translated into an operational methodology to calibrate price time series and diagnose bubbles as they develop.

Kindleberger (2000) and Sornette (2003) have identified the following generic scenario developing in five acts, which is common to all historical bubbles: displacement, take-off, exuberance, critical stage, and crash.

Fama (1991), Campbell (2000) and Cochrane (2004, 2006) have discussed and documented interrelationships between equity prices, earnings and dividends, discounting factors, real growth in GDP, increases in productivity, inflation, and interest rates. More narrowly, Bordo and Wheelock (2004, 2006, 2007), Bordo, Dueker and Wheelock (2007), Bernanke (2002), Bernanke and Gertler (1999, 2001), Bernanke and Kuttner (2005), Bordo and Jeanne (2002), and Borio and White (2004) have offered comprehensive reviews of the relationships between macroeconomic fundamentals, asset prices and the role of monetary policy. Bordo and Wheelock (2004) identify several episodes of sustained above average increases in the U.S. stock market measured by a representative index during the 19th and 20th centuries and then evaluate the role of economic fundamentals in each episode. Two booms stand out in terms of their length and rate of increase. These are the booms of 1923-29 and the 1994-2000. The authors conclude that these booms occurred in periods of rapid real growth and productivity suggesting that booms are driven at least partly by fundamentals. Bordo and Wheelock (2006, 2007) and Bordo, Dueker and Wheelock (2007) examine the relationship between stock market booms and monetary policy both in the U.S. and nine other developed countries during the 20th century. They find evidence that stock market booms in the U.S. and several other developed countries typically build up during periods of above average growth in real GDP and below average inflation. They find little evidence that stock market booms are driven by excessive liquidity. They also find evidence that equity booms ended within few months of an increase in inflation and consequent monetary policy tightening. Their overall conclusion suggests that stock market booms reflect both real positive macroeconomic fundamentals and monetary policies targeting price stability.

Greenspan (1999, 2002, 2003, and 2004) saw a conundrum in the use of monetary policy to defuse an asset price boom, however, and expressed the view that stock market booms are more likely to occur when inflation is low. Greenspan has remarked that the bubble during the period 1996-2000 may have been the result of the Fed’s success in reducing inflation and achieving price stability. Economists such as Bullard and Schalling (2002), Cecchetti (1998), Cecchetti, Genberg, Lipsky and Wadhwani (2000), Cecchetti, Genberg, and Wadhwani (2002), Gogley(1999) Filardo (2000, 2001, 2004), Roubini (2006) and Selody and Wilkins (2007), among others, carefully explore the appropriate role of monetary policy when the Fed confronts stock market booms. Some argue that monetary policy should deflate asset bubbles while others propose that uncertainty about the size of the bubble and doubt about the Fed’s effectiveness in diffusing such a bubble, both lead to no action. When monetary policy is successful in keeping price inflation low, history demonstrates that price-earnings ratios increase. What is really needed to keep stock market bubbles from occurring is some degree of uncertainty about future economic fundamentals.

Such uncertainty moderate’s asset price booms. If there is strong consensus that economic fundamentals will remain robust and the Fed’s monetary policy will succeed achieving long-term price stability, under such conditions often asset prices increase rapidly. There is a large literature known as the “Great Moderation” that reviews the recent empirical evidence of decreased volatility in the real economy since around mid-1980s and the gradual increase in asset volatility. Stock and Watson (2003) document the decrease in real volatility and Campbell (2005) decomposes real stock returns into fundamental news and return news components to analyse the effects of the Great Moderation on each. It is not currently clear if monetary policy could or should target both real volatility (output a price) as well as asset volatility.

From the papers cited above a broad analytical agreement emerges about the inter relationships among economic fundamentals, monetary policy and equity booms supported by empirical evidence. First, most asset booms have occurred during periods of relatively rapid economic growth generally accompanied by increases in productivity growth. Such periods of rapid economic growth are associated with increasing employment causing the percent of unemployment to decline. This suggests that booms were driven at least to some extent by standard macroeconomic fundamentals. There is no evidence that the stock market has prospered during periods of prolonged recessions. Second, many booms have also occurred during periods of relatively low interest rates. Low interest rates reduce the cost of borrowing for firms and encourage them to borrow and invest. Low interest rates, other things equal, also reduce the cost of capital and increase firms’ earnings. During the period when various monetary aggregates were considered important policy variables, stock market increases were associated with rapid growth of the money stock and bank credit, reflecting either passive accommodation of booms by the banking system or expansion of the monetary base by means of an easy monetary policy. Since the mid-1980s, the role of liquidity has been replaced by successful Federal funds policies. Third, over the past 200 years in the U.S., stock market booms have occurred in periods of deflation (e.g., the late 1870s and early 1880s), in periods of inflation (e.g., the 1830s, 1840s, late 1890s, and early 1900s) and in periods of price stability (e.g., the 1920s and 1990s).

Two major stock market bubbles happened in the late 1990s and early in 2000 in the USA. The soaring market of the 1990s was seen by many economists as the harbinger of a new age sustained, rapid economic growth. The same situation was in 1920.

As in the 1990’s it was widely claimed that a new economy had taken root in the USA. In both periods, unemployment was low with stable prices in the twenties and low inflation in the nineties. Participation in the market increased, as investing in the market seemed safer, with reduced macroeconomic risk and the seeming abundance of high return opportunities (White 2006).

In both 1920 and 1990 the boom was explained by scientists as driven by technological change raising dividends. The idea of technological age played a key role in the mind of the 1990s’ bull market. The rapid changes in computer/information technology and biotechnology were heralded as placing the economy on a higher trajectory. The new era vision was supported by many economists. It was expected that technology would have an even greater impact on productivity growth. Like in the 1920s the conclusion for 1990 was fairly clear – the expected dividend growth was not a major factor driving the boom (Eatwell 2004). Several articles were published analysing and comparing the situations in 1920 and 1990 and trying to provide the explanations of stock market bubbles (White 2006; Eatwell 2004; Pastor, Veronesi 2004; Cochrane 2002; Caballero, Hammour 2002; Kraay, Ventura 2005; DeLong, Magin 2001). Different economists provide different explanations. Pastor and Veronesi (2004) studied the NASDAQ bubble and argued that the fundamental value of a firm increases with uncertainty about average future profitability, and this uncertainty was unusually high in the late 1990s. Authors stated, that the models which had been used to value technology stocks omitted an important determinant of the fundamental value, namely the uncertainty about a firm’s average future profitability, which can also be thought of as the uncertainty about the average future growth rate of the firm’s books value. According to Pastor and Veronesi (2004) the late 1990s witnessed high uncertainty about the average growth rates of technology firms, and that this uncertainty was partly responsible for the high level of technology stock prices. Cochrane (2002) suggested that a mechanism much like the transactions demand for money drove many stock prices above the “fundamental value”. Caballero and Hammour (2002) interpreted a stock market bubble as a high-valuation equilibrium with the low effective cost of capital based on optimism about the future availability of funds for investment. Authors showed in their investigation that such bubbles arise naturally when the expansion is concentrated in the “new economy” sector and when it is supported by sustained financial surpluses, both of which would constitute an integral part, as cause and consequence, of a “speculative growth” equilibrium. The high-valuation equilibrium may take the form of a stock market bubble. In contrast to classic bubbles on non-productive assets, the bubbles in the Caballero and Hammour (2002) model encourage real investments, boost long run savings, and may appear in dynamically efficient economies. In the case of the U.S. in the 1990s, the authors argue that at least two factors created the conditions for a speculative growth episode: the emerging information technology sector and conservative fiscal policy.

Both factors created favourable conditions for growth-saving feedback and for the possibility of a speculative equilibrium characterized by extreme stock market valuations and a potential crash. Kraay and Ventura (2005) have provided a joint account of some of the major US macroeconomic events of the past decade: large current account deficits and a steady decline in the net foreign asset position; the large boom and a subsequent crash in the stock market; and the emergence of large fiscal deficits. According to the conventional view, the evolution of the stock market and fiscal deficits are unrelated events, with the former driven by sharp swings in US productivity, and the latter by shifting US political considerations. Both, in turn, fuelled current S. Girdzijauskas, D. Štreimikienė.

Kraay and Ventura (2005) proposed two alternative views in which the stock market and the fiscal deficits are closely linked. Authors stated that the US economy contains “pockets” of inefficiency. This opens the possibility for asset bubbles to exist, which in turn provides a more plausible explanation for the large swings in equity values over the past decade. The appearance of a bubble in the US stock market in the second half of the 1990s accounts for much of the decline in US net foreign assets during this period. At the same time, the bubble raised welfare worldwide by eliminating inefficient investments. According to Eatwell (2004) the collapse of the stock market in 2000 was the result of a coordination failure or change in investor sentiment, and the rapid expansion of public debt since then served to displace inefficient investments in the same way that the bubble did. Viewed in this light, the large budget deficits of the Bush administration can be interpreted as a welfare improving response to this market failure. But there is also a more “cynical” interpretation, that is observationally equivalent to the “benevolent” view. Under this interpretation the expansion of public debt caused the collapse of the bubble, as the US government tried to appropriate the value of the bubble from its US and foreign owners. White (2006) provides his own comments on the 1920 and 1990 stock market bubbles and criticizes both fundamental approach in forward looking assets and waves of pessimism and optimism driving investors decisions and therefore creation of bubbles, however the author does not provide any reasonable explanations of stock market bubbles and puts more questions than answers.

Stock bubbles can be interpreted as a deviation of stock prices from their fundamental value, which adopting the Cochrane model, which uses the ratio of stock prices to dividends which is influenced by expected returns and dividend growth rates. In general, if there is no bubble, the ratio of stock prices to dividends will be balanced in the long term or stationary. However, when a bubble occurs, the ratio explodes. There are several other approaches to measuring bubbles, including rational bubbles (Blanchard and Watson, 1982, bubbles caused by nonfundamental factors. Through the equation Pt = P\*t+Ct, where asset prices deviate from the fundamental price of the bubble component. This bubble component grows at the level of long-term return on assets with a probability π and will crash (break) with a probability 1-π. Prior to the occurrence of the bubble burst phenomenon, based on the above theory it was stated that the return on assets at that time experienced faster growth than the average return journey.

Another approach to bubble measurement is Froot and Obsfeld's bubble model, which is called an intrinsic bubble, a bubble that has an asset dependency on its fundamentals. Through the equations P(Dt) = Ppvt + B(Dt) and B(Dt) = cDtλ, dividends are influenced by the present value of the stock price or fundamental and intrinsic bubble prices. The intrinsic bubble is influenced by c, which is a constant that has a positive value (c > 0) while λ is more than 1 (λ > 1). The intrinsic bubble is determined by the dividend growth trend, dividend log, and random events with zero conditional mean and variance. If the fundamentals remain unchanged, the bubble component remains constant at a constant value. If the fundamentals show increasing activity, then the bubble component and the price of the asset will be affected by a persistent deviation from the value of the fundamental.

Dating back at least to 17th century in the Netherlands, Tulipomania (1636-1637) is widely considered as the earliest example of a bubble. Following the Dutch tulipomania, there was the Mississippi Bubble in France and South Sea Bubble in England, during 1719-1720, with these three events becoming some of the most famous stock price bubbles. The twentieth century saw bubble episodes such as but not limited to German Stock price bubble in 1927, the Wall Street Crash in 1929, Japanese asset price bubble in 1980’s to 1990’s. Bitcoin has also been characterized as a more recent example of a bubble, where its price reached an all-time high of close to US$ 20.000 in December 2017.

Prior to the work of Philips et al. (2015, PSY) traditional bubble detection tests were typically based on ex post analysis, see Gurkaynak (2008). To address a need for real time monitoring and surveillance of asset prices, a series of papers by Philips et al. (2011. PWY), Philips and Yu (2011), Philips et al. (2015, PSY) and Philips and Shi (2020) develop novel right tailed unit root tests for market exuberance by identifying the originating and collapse dates of asset price bubbles. These right-tailed unit root test have been widely employed as early warning systems for bubble-like behaviour in a wide variety of financial markets.

**7.1** **Empirical Bubble Literature**

Empirical bubble literature is primarily concerned with mechanisms for econometrically detecting bubbles and measuring their extent.

Empirical bubble detection mechanisms are especially valuable regulators, since they enable them to understand, monitor, and control systemic risk in financial system.

The goal of empirical bubble literature is to develop mechanisms to identify the origination, termination, and extent of explosive behaviour in asset prices based on explicit quantitative measures. These mechanisms must be able to empirically separate the contribution of rational bubbles and market fundamentals to exuberance detected in the data. The exuberance detected should only be attributed to bubbles when all other phenomena that effect asset prices have been ruled out as possible explanations. Even though the literature on empirical bubble detection has made great strides over the past decades, it is still a notoriously challenging task to design appropriate tests for identifying bubble episodes. The reason for this lies in the fact that the determinants of the fundamental value are generally not observable, and it is therefore challenging to determine an asset’s fundamental value.

The housing market has attracted several tests for bubbles. Yiu et al. (2013) test for explosive behaviour in three Hong Kong residential property market segments (e.g., the overall market, the mass segment, and the luxury segment) between March 1993 and March 2011. They found several bubbles in all three segments. Jiang et al. (2015) identified episodes of bubbles in the Singapore real estate market between 2006Q4 and 2008Q1. Their findings suggest that the cooling measures, implemented by the Singapore government during 2009–2013 to control house price inflation, were effective. Shi et al. (2016) examine the house price–rent ratios for housing bubbles in Australia major capital cities for the period December 1995 to January 2016. Their results suggest a nationwide occurrence of speculative behaviour in all capital cities in the 2000s. Turning to New Zealand, Greenaway-McGrevy and Phillips (2016) investigate the evidence of bubbles in the New Zealand regional housing markets for the period 1993Q1–2014Q4 and propose a time-varying regression approach to measure potential bubble migration across markets. They find evidence of a New Zealand-wide housing bubble over the 2003–2008 period and that the housing bubble was transmitted from Auckland to the other regional centres. Pavlidis et al. (2016) proposes a panel setting for the PSY procedure to examine explosive behaviour in housing markets for a large set of 22 countries and finds strong evidence of explosive behaviour between the early 2000s and 2006/2007. Engsted et al. (2016) examine house price bubbles using price–rent ratios in 18 OECD countries from 1970 to 2013. Their results show that 16 of the 18 markets exhibit bubble-like behaviour.

Hu and Oxley (2018) investigate the presence of bubbles in the US housing market at the State level for the period January 1975 to December 2014.

They present empirical evidence to show a housing bubble that originates in the early 2000s and collapses in the mid-2000s in more than 20 States and the District of Columbia concluding that the bubbles of the 2000s were more widespread than the 1980s.

**7.2** **Literature Review on Detecting Financial Bubbles**.

Throughout history a multitude of research studies have been conducted with the objective of identifying bubbles in financial markets that are characterized by speculation, including but not limited to stock markets, foreign exchange markets, real estate markets and more recently, cryptocurrency markets. However, within the context of this study there will be a concise overview of studies conducted solely within the sphere of the stock market, with a particular emphasis on studies that utilize models on time-series data.

Shiller (1981) introduced a novel method called Variance Bounds Tests, which were applied to sample data of the S&P 500 price index from 1871 to 1979, revealing evidence of a bubble existence. However, Shiller’s approach is often deemed less reliable when applied to small sample sizes.

Phillips et al. (2011) proposed the Sup Augmented Dickey-Fuller Test (SADF), also known as the PWY method, to assess the presence of rational bubbles in financial markets. This approach is based on the null hypothesis of a unit root, analogous to the conventional Dickey Fuller-Test, but with a right tailed alternative hypothesis. Rejecting the null hypothesis in this test indicates the presence of explosive behaviour in the price series, thereby providing empirical evidence for the existence of a bubble. The right tailed SADF unit root tests are conducted using rolling window forms.

Homm and Breitung (2012) applied this test to detect stock market bubbles, and after a process of simulation and evaluation criteria comparison, the authors found that SADF test was the most optimal among the methods employed. The SADF test is effective when there is a single bubble event, but in practical applications, there may be multiple bubbles appearing in sufficient large samples. While this method successfully identified famous historical bubbles, the SADF test failed to detect the 2007-2008 debt crisis bubble.

Philips et al. (2015) developed the Backward sup ADF (BSADF) test as an improvement over the SADF method, also referred to as the PSY procedure, to overcome its limitations. The BSDAF test is an iterative application of the right tailed ADF test based on the rolling window SADF test that aims to detect explosive patterns in sample sequences. Compared to the SADF and GSADF tests to the S&P 500 index from 1871 to 2010, revealing that the BSADF successfully identified two bubble periods: the post 1954 war period, Black Monday in 1987, and the dot-com bubble in 2000.

Based on the findings listed above, it is evident that the BSADF test is an effective approach for detecting the presence of market bubbles. Extensive simulation studies show that the Pwy approach is especially effective in detecting a single bubble episode.

A sup Augmented Dickey-Fuller (SADF) method is utilized in testing the presence of explosive behaviour and such testing procedure is implemented as follows. For each time series xt, it is applied the ADF test for a unit root against the alternative of an explosive root (right tailed). The following autoregressive specification for xt is estimated by least squares:

A math equation with black text

Description automatically generated with medium confidence (1)

For some given value of the lag parameter J, where NID denotes independent and normally distributed. The null hypothesis of this test is  and the alternative hypothesis is . The above equation is estimated repeatedly using subsets of the sample data incremented by one additional observation at each pass in the forward recursive regression.

Thus, the SADF test is constructed by repeatedly estimating the ADF test. Let rw be the window size of regression. The window size rw (rw=r2-r1) expands from r0 to 1, where r0 is the smallest sample window width fraction and 1 is the largest window fraction (the full sample). The starting point r1 is fixed at 0, and the end point of each sample (r2) is, therefore, denoted by  The SADF statistic is defined as the sup value of the ADF statistic sequence: 

The SADF test statistic cannot locate the origination and collapse dates of a bubble. To identify the origin and the collapse dates, we can compare the recursive test statistic ADFr against the relevant right-tailed critical values. If re is the origination date and r1 is the collapse date, it is possible to construct estimates of the dates as follows:

A black text on a white background

Description automatically generated (2)

A black text on a white background

Description automatically generated (3)

The ADF statistic and its corresponding critical value re used for dating the origination and termination dates of a bubble.

The PSY approach extends the work of PWY by allowing for more flexible window widths in the recursive regressions on which the test procedures are used. The PSY test has shown great power in detecting the presence of multiple bubbles.

The martingale null with an asymptotic drift is specified as

A black and orange math symbols

Description automatically generated with medium confidence (4)

Where d is a constant, the localizing coefficient n is greater than ½ and T is the sample size.

The alternative hypothesis is a mildly explosive process:

 (5)

Where  . The following regression model is estimated:

A math equation with numbers and symbols

Description automatically generated (6)

Where  is an intercept and k is optimum lag length.

The backward sup ADF (BSADF) test relies on repeated estimation of the ADF test regression model on subsamples of the data in a recursive fashion. The window size rw expands from r0 to 1, where r0 is the minimum window size.

The termination date of a bubble is calculated as the first observation after (Tre)+ log(T) whose backward sup ADF statistic falls below the critical value of the backward sup ADF statistic.

The BSADF statistic and its corresponding critical value are used for dating the origination and termination dates of a bubble as follows:

A close up of a text

Description automatically generated (7)

A black text on a white background

Description automatically generated (8)

It is assumed that the duration of the bubble exceeds log(T), where 8 is a frequency dependent parameter. The fractional origination and termination points of a bubble are calculated according to the following first crossing time equations:

A close up of a logo

Description automatically generated (9)

 (10)

Where scv T r2 is the 100(1-BT)% critical value of the sup ADF statistic based on (TR2) observations.

This approach, therefore, has greater power in the detection of multiple bubbles. Phillips et al. (2015) also derive the asymptotic distribution of this test statistic under the null. The minimum window size r0 needs to be large enough to allow initial estimation but not too large to miss an early bubble episode. As recommended in PSY, the minimum window size r0 is set equal to .

The limit distribution of the BSADF statistics is identical to the case where the regression model includes an intercept and the null hypothesis is a random walk or unit root process without drift. The usual limit distribution of the ADF statistic is a special case of equation with r1 = 0 and r2 = rw = 1 while the limit distribution of the SADF statistic is a further special case of r1 = 0 and r2 = rw E(r0,1).

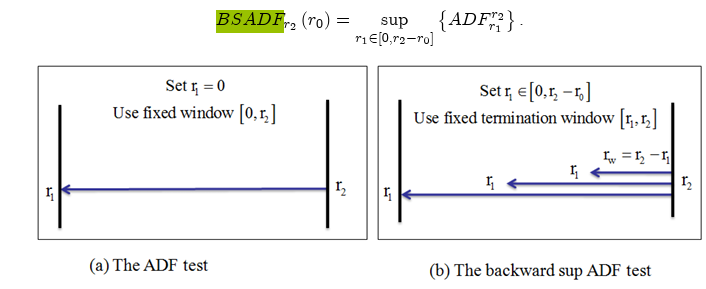
The asymptotic BSADF distribution depends on the smallest window size r0. In practice, r0 needs to be chosen according to the total number of observations T. If T is small, r0 needs to be large enough to ensure there are enough observations for adequate initial estimation. If T is large, r0 can be set to be a smaller number so that the test does not miss any opportunity to detect an early explosive episode. The asymptotic critical values are obtained by numerical simulations, where the Wiener process is approximated by partial sums of 2.000 independent N (0,1) variates and the number of replications is 2000. The finite sample critical value is obtained from 5000 Monte Carlo replications. The lag order k is set to 0. The parameters (d and n) in the null hypothesis are set to unity.

A table of values with numbers and letters

Description automatically generated with medium confidence

Table 2. Critical values of the SADF and GSADF tests against an explosive alternative

In particular, Backward SADF test performs a sup ADF test on a backward expanding sample sequence where the end point of each sample is fixed at r2 and the start point varies from 0 to r2-r0. The corresponding ADF statistic sequence is . The backward SADF statistic is defined as the sup value of the ADF statistic sequence over this interval:



(11)

Both ADF and BSADF detectors provide consistent estimates of the origination and termination dates of the bubble. When the point estimates re and rf are obtained as in PWY using the ADF test and the first crossing times, then  provided the following rate condition on the critical value cvBT holds A black text on a white background

Description automatically generated (12)

Consistency of (re,rf) was first proved in a working paper (Philips and Yu, 2009). When the point estimates re and rf are obtained from the BSADF detector using the crossing time criteria (9), (10), this model again has consistency  under the corresponding rate condition on the critical value scvBT A black text on a white background

Description automatically generated. (13)

Both strategies consistently estimate the origination and termination points when there is a single bubble episode in the sample period.

The rate conditions (12) and (13) require for consistency of (re,rf) that (cvBT, scvBT) pass to infinity and that their order of magnitude be smaller than . It is sufficient for consistency of (re,rf) that the critical values cvBT and scvBT used in the recursions expand slowly as  , for example at the slowly varying rate log (T). The probability of false rejection of normal behaviour then goes to zero. The upper condition that delimits the rate at which (cvBT, scvBT) pass to infinity ensures the successful detection of mildly explosive behaviour under the alternative. In effect, the critical values used in the crossing time (9) and (10) must not pass the infinity too fast relative to the strength of exuberance in the data which is governed by the value of the localizing parameter . (14)

**8.** **Evaluation and justification of predictive models.**

While a profound theoretical understanding of the bubble phenomenon and its causes is without doubt very important, it is, however, by no means sufficient. Academics and policy makers alike highlight the importance of sound empirical methods for detecting bubbles in asset prices. As Giglio et al. (2016) stress, bubbles should not be treated as a purely theoretical problem. Rather, the debate over the existence of bubbles is an inherently empirical question. The goal of empirical bubble literature is to develop mechanisms to identify the origination, termination, and extent of explosive behaviour in asset prices based on explicit quantitative measures. These mechanisms must be able to empirically separate the contribution of rational bubbles and market fundamentals to exuberance detected in the data. The exuberance detected should only be attributed to bubbles when all other phenomena that effect asset prices have been ruled out as possible explanations. Even though the literature on empirical bubble detection has made great strides over the past decades, it is still a notoriously challenging task to design appropriate tests for identifying bubble episodes. The reason for this lies in the fact that the determinants of the fundamental value are generally not observable, and it is therefore challenging to determine an asset’s fundamental value. Empirical evaluation of bubble phenomena requires some econometric technology. Gurkaynak (2008) provides a thorough survey of the literature on econometric tests for rational bubbles in the context of the present value of dividends model. He points out that while many studies have attempted to empirically separate the contribution of rational bubbles and market fundamentals to asset price movements, there is still no general agreement on which bubble detection methodology should ultimately be used. Furthermore, there is no consensus on whether bubbles are present in the different time series that have been analysed. He states: “For each paper that finds evidence of bubbles, there is another one that finds the data equally well without allowing for a bubble.”

**8.1** **Log Periodic Power Law**

Is an empirical method used in the context of bubbles in stock markets. Mudholkar, Srivastava, and Kollia (1996) introduced this model as a survival analysis technique to examine the hazard of risk of an event occurring over time. In the study of bubbles, the Weibull Exponential model is applied to investigate the duration and intensity of these bubbles. It enables researchers to analyse the probability of a bubble bursting or ending at different time points, providing insights into the dynamics and vulnerability of speculative market behaviour. The contribution of Mudholkar, Srivastava, and Kollia lies in their development and application of the Weibull model, offering a valuable tool for understanding the temporal aspects and risk factors associated with bubbles in stock markets.

LPPL model started by conducting a correlation analysis to assess the autocorrelation of daily returns across multiple datasets. The Autocorrelation measured the relationship between an observation and its lagged values, indicating the degree of similarity or dependence between consecutive data points.

Followed by plotting a histogram of daily returns, displaying how returns are distributed across different range of values assisting the understanding of central tendency, variability and shape of return distributions for each stock dataset. The implemented parameters were as follows: Figure size set to (10, 3) using plt.rcParams['figure.figsize'], ensuring that the histograms are displayed with appropriate dimensions. Each histogram was divided into 200 bins (bins=200), providing sufficient granularity to visualize the distribution of returns. The x-axis limits were set to (-0.2, 0.2) using plt.xlim(-0.2, 0.2), restricting the visualization to returns within this range. The parameter alpha=0.75 was used to set the transparency of the histograms, allowing for better visualization of overlapping bars.

Next stage was to formulate a drawdown analysis by Identifying the Peaks and Troughs in asset prices using the difference between consecutive price changes (ds['price'].diff(-1)). Peaks were defined as points where price changes from increasing to decreasing (pmin\_pmax == 1), while troughs were defined as points where price changes from decreasing to increasing (pmin\_pmax == -1). Drawdowns were computed as the percentage decline in prices from peaks to troughs ((np.array(ds['price'][pmin.index]) - np.array(ds['price'][pmax.index])) / np.array(ds['price'][pmax.index])).

The duration of each drawdown was computed as the number of business days between the corresponding peak and trough dates (np.busday\_count(p1.date(), p2.date())).

On the next step, the Weibull distribution was fit to the drawdown data to allow for the characterization of drawdown events in terms of their magnitude and frequency, providing insights into the underlying statistical properties of drawdowns. The drawdown data points were plotted against their corresponding ranks, with the size of each point proportional to the duration of the drawdown, allowing for the visualization of drawdowns of varying magnitudes and durations followed by fitting a dashed line representing Weibull distribution fit to the drawdown data. The curve represents how well the Weibull distribution align with the observed drawdown. The y-axis was scaled logarithmically to accommodate the wide range of drawdown ranks, ensuring that both small and large drawdowns are visible on the plot. The curve\_fit function from scipy.optimize was used to estimate the parameters of the Weibull distribution (chi and z) that minimize the difference between the observed drawdown data and the Weibull fit. The initial parameter values (init\_vals) were provided as starting points for the optimization algorithm. For a clearer understanding, the size of each data point was adjusted based on the duration of the drawdown, providing additional visual information about the persistence of each drawdown event.

A red dot line and black dotted line

Description automatically generated

Table 3. Weibull distribution to drawdowns

The next stage of the model was to use a method suggested by Jaccobson et al. (2009), which ensures that only extreme drawdowns are considered as crashes by calculating crash thresholds by identifying the drawdown value corresponding to the 99.5 percent of drawdown magnitudes, the method implies identifying crashes by selecting drawdown events that exceed the computed crash threshold by defining crash start and end dates, duration and rank based on drawdown magnitude. By printing crash details for each dataset alongside their respective titles, the code promotes transparency and facilitates the reproducibility of the analysis. This allows other researchers to verify the findings and conduct further investigations if needed.

The next stage of the analysis was to generate subplots for each dataset to visualize the financial data alongside identified crash events. Each subplot consists of two layers: the upper layer which displays the price data with crash events highlighted, while the lower layer displays the corresponding volatility. Matplotlib's GridSpec was employed to create a grid of subplots with varying heights. Allowing for better organization and presentation of the data. The upper subplot (price and drawdown) has a higher height ratio compared to the lower subplot (volatility), emphasizing the importance of price movements during crash events.

Crash events were identified based on the 99.5th percentile drawdown quantile, being highlighted using red shaded regions in both the price and volatility subplots.

Additionally, individual drawdown points were marked with red triangles on the price subplot for further clarity, ensuring that plots were well-sized and properly spaced, preventing overcrowding and ensuring readability.

Continued by displaying detailed crash information, including duration and rank which allows for a comprehensive analysis of market downturns, facilitating deeper insights into their characteristics and impacts.

To achieve this step, the code iterated through each dataset (datasets and dd\_df) along with the corresponding number of crashes (n\_crashes). For each dataset, it selected the top n crashes based on their rank (where n is the corresponding value from n\_crashes). It created a DataFrame (df\_c) containing information about each selected crash, including drawdown, start and end dates, duration, and rank. These crash datasets were appended to the crashes list and the start and end dates of crashes were converted to datetime objects and then to date objects for better readability.

This study incorporated the Weibull outliers, which focuses on extreme drawdown events that deviate significantly from the norm. This approach enables the identification and visualization of the most severe market downturns, providing insights into their timing and magnitude. To achieve the proposed step, matplotlib was used to create subplots for each dataset.

Each subplot consists of two panels: the upper panel shows the normalized price and drawdowns, with crashes highlighted, while the lower panel displaying the volatility.

This layout allows for a clear comparison between price movements, drawdowns, and volatility during crash periods. Crashes were highlighted using red vertical shaded regions in both the upper and lower panels.

This highlighting helps to visually identify the periods corresponding to crashes, facilitating their interpretation and analysis, the use of a grid layout (GridSpec) optimized space utilization, allowing for the simultaneous visualization of multiple datasets in a compact manner.

This efficient use of space ensured that all relevant information is presented clearly without overwhelming the viewer.

A graph of a crash

Description automatically generated

A graph of a graph of a graph

Description automatically generated with medium confidence

Table 4. Weibull outliers

**Justification**

The Weibull model facilitated comprehensive analysis and modelling of drawdown distributions and extreme events in financial datasets. By fitting Weibull distributions to drawdown data, insights into the statistical properties and characteristics of drawdowns were gained, enhancing understanding and aiding in risk assessment. Additionally, the determination of crash thresholds based on extreme drawdown quantiles enabled the identification of significant market downturns, allowing for proactive risk management and decision-making. Through systematic analysis and visualization of drawdowns and crashes, Log Periodic Power Law provided valuable tools for financial analysts and investors to assess market dynamics, identify potential risks, and inform investment strategies effectively.

**8.2** **Topological Data Analysis**

Topological Data Analysis, abbreviated TDA, is a suite of data analytic methods inspired by the mathematical field of algebraic topology, it uses tools from algebraic and combinatorial topology to extract features that capture the shape of data (Carlsson,2009).

Giotto-tda inherits the flexibility of scikit-learn, the most popular all purpose ML framework Pedregosa et al. (2011). To enable compatibility with scikit learn a Transformer Resampler Mixin base class was designed. It provides a resample method that modifies the number of samples in the target in accordance with how the input data is transformed.

Persistent Homology consists of four main points:

Homotopy.​

Algebraic topology aims to describe the connectivity of any arbitrary space. It does this by computing the homotopy, or number of “loops” in each dimension.

2. Approximation.

​ In computational topology, datasets can be interpreted as samples taken from an underlying topological space, and for any given margin of error a topology can be constructed to approximate the underlying space.

 3. Homology​.

Homotopy groups are extremely difficult to compute in high dimensions. Homology is a similar concept which can be easier to compute.

 4. Persistence.​

​​Persistence barcode plots show which topological features persist through many scales of the data and can be used to calculate similarity between different spaces.

**Homotopy:**

To describe a shape’s connectivity, you can count the number of “loops” in the shape, starting from an initial point called the basepoint. For example, the letters “O” and “P” are said to have the same connectivity because they each have one loop, whereas the letters “O” and “B” are said to have different connectivity because “O” has one loop and “B” has two loops. Algebraic topology takes this idea of classifying shapes based on how many loops they have and extends it to spaces of arbitrarily many dimensions (Carlsson, 2009).

**Approximation**

Real-world datasets are not explicit topological spaces. Rather, they are collections of points sampled​ from topological spaces, and the goal of topological data analysis is to analyse these point clouds and infer information about their underlying topological spaces. One can do this by using the point cloud to construct a “simplicial complex” which approximates the underlying topological space (Carlsson, 2009).

A diagram of a diagram of a molecule

Description automatically generated

The main idea behind turning point clouds into simplicial complexes is to put epsilon-balls, or error margins, around points and use the overlaps to determine the connections in the simplicial complex. The constructions generated using different values of epsilon will correspond to topological approximations of the point cloud at different levels of scale.

**Homology**

Unfortunately, homotopy groups are extremely difficult to compute in high dimensions. However, there is a similar concept, homology, which can be calculated on simplicial complexes via linear algebra (Carlsson, 2009). Like homotopy, homology also counts the number of loops of each dimension in a space, where loops are allowed to shift along the boundary of a higher-dimensional component on the space. In simplicial complexes, the lowest-level components are points, followed by segments, and then triangles, and then solid tetrahedrons, and so on. You can think of an nth level component as an n-dimensional triangle, or more formally, an n-simplex.

**Persistence**

When we’re interested in topological features which persist across many scales of the data, we need consider all values of epsilon for the simplicial complex we construct on our data.

This is the idea behind persistent homology: we can figure out which topological features persist over the full range of scale by making a plot that says whether a particular homology component was detected at some value of epsilon (Carlsson, 2009).

This study explores the evolution of daily returns of major US stock market indices during the technology crash of 2000, and the financial crisis of 2007-2009. Methodology is based on topological data analysis (TDA). The method is to use persistence homology to detect and quantify topological patterns that appear in multidimensional time series. Using a sliding window to extract time-dependent point cloud data sets, to which associate a topological space. The detection of transient loops that appear in this space and measure their persistence. This is encoded in real-valued functions referred to as a 'persistence landscapes'. Followed by quantifying the temporal changes in persistence landscapes via their Lp-norms.

The study suggests that TDA provides a new type of econometric analysis, which goes beyond the standard statistical measures. The method can be used to detect early warning signals of imminent market crashes. After performing TDA, it is believed that his approach can be used beyond the analysis of financial time series presented in this study.

The developed pipeline consists in: embedding the time series into a point cloud and constructing sliding windows of point clouds, building a filtration on each window to have an evolving structure encoding the geometrical shape of each window, extracting the relevant features of those windows using persistence homology, comparing each window by measuring the difference of those features from one window to the next, constructing an indicator of crash based on this difference.

**Time series as point clouds — Takens’ embedding**.

A typical starting point in a TDA pipeline is to generate a simplicial complex from a point cloud. Thus, the crucial question in time series applications is how to generate such point clouds? Discrete time series, like the ones considered for this study, are typically visualised as scatter plots in two dimensions. This representation makes the local behaviour of the time series easy to track by scanning the plot from left to right. But it is often ineffective at conveying important effects which may be occurring over larger time scales.

One well-known set of techniques for capturing periodic behaviour comes from Fourier analysis. The discrete Fourier transform of a temporal window over the time series gives information on whether the signal in that window arises as the sum of a few simple periodic signals.

For the purposes of this research there will be a different way of encoding a time-evolving process. It is based on the idea that some key properties of the dynamics can be unveiled effectively in higher dimensions. This section begins by illustrating a way of representing a univariate time series as a point cloud, i.e. a set of vectors in a Euclidean space of arbitrary dimension.

The procedure works as follows: By picking two integers d and τ. For each time tᵢ ∈ (t₀, t₁, …), collect the values of the variable y at d distinct times, evenly spaced by τ and starting at tᵢ, and present them as a vector with d entries, namely:



The result is a set of vectors in d-dimensional space! τ is called the time delay parameter, and d the embedding dimension. This time-delay embedding technique is also called Takens’ embedding after Floris Takens, who demonstrated its significance with a celebrated theorem in the context of nonlineardynamical systems. Finally, applying this procedure separately on sliding windows over the full time series leads to a time series of point clouds (one per sliding window) with possibly interesting topologies. The GIF below shows how such a point cloud is generated in 2 dimensions.

Methods and evaluation:

Embedding: Single Takens Embedding: The time series data is embedded into a higher-dimensional space using the Single Takens Embedding technique. It is performed with a fixed embedding dimension of 3 and a time delay of 2.

This step captures the dynamics of the time series data in a higher-dimensional space, making it amenable to topological analysis.

Sliding Window: Sliding windows are used to segment the embedded time series data into smaller chunks.

Each window is shifted by a stride of 4-time steps, with a window size of 31.

This step helps in analysing local topological features within different segments of the time series.

Visualization: The point cloud of the embedded data in a 3D space is visualized using a plot\_point\_cloud function.

This visualization provides insights into the structure and distribution of data points in the embedded space.

Derivative Analysis: The first derivative of the S&P 500 index data is computed to analyse the rate of change.

Absolute values of means of differences between consecutive windows are calculated.

These derivative values are plotted against time indices to visualize changes in the index over time.

A graph showing a number of blue lines

Description automatically generated

Table 5. First Derivative of the S&P 500 Index

To get an indication of where the first crashes happened, the following steps were conducted: Data preprocessing, embedding, sliding windows analysis, feature extraction, visualization, which is the next stage being evaluated for TDA.

Data Preprocessing: The TDA model preprocesses the dataset by converting the 'Date' column to datetime format, setting it as the index, and resampling the close prices. This ensures that the data is properly formatted and organized for further analysis.

Embedding: The model uses single-takens embedding to embed the time series data into a higher-dimensional space. This embedding technique preserves the topological properties of the data and allows for the detection of patterns and structures.

Sliding Window Analysis: Sliding window analysis is applied to the embedded data to capture local patterns and variations over time. This approach enables the model to analyse the data in smaller segments, allowing for the detection of changes in dynamics and behaviours.

Feature Extraction: The model calculates the absolute derivative of means from the sliding window data. This feature extraction step captures the rate of change in the data and provides insights into its dynamics and volatility.

A graph showing a number of data

Description automatically generated

Table 6. Crash Detections

Homology

Homology Dimensions: The model specifies homology dimensions (0-dimensional and 1-dimensional) to capture different topological features of the data. This allows for the detection of both connected components and loops in the underlying space, providing a comprehensive analysis of its topological structure.

Vietoris-Rips Persistence: The model utilizes the Vietoris-Rips persistence algorithm to compute topological features such as connected components, loops, and voids across various spatial scales. By considering different values of the radius parameter, the algorithm captures the persistence of these features and their evolution over different scales.

Parallelization: The model leverages parallel computing (n\_jobs=-1) to efficiently compute the persistence diagrams, enabling faster processing and scalability to large datasets. This parallelization significantly reduces the computation time, making the analysis more feasible for complex datasets.

Visualization: The model provides visual representations of the persistence diagrams, allowing for intuitive interpretation of the topological features detected in the data. The plot of persistence diagrams enables the identification of persistent features across different dimensions and scales, aiding in the understanding of the data's underlying structure.

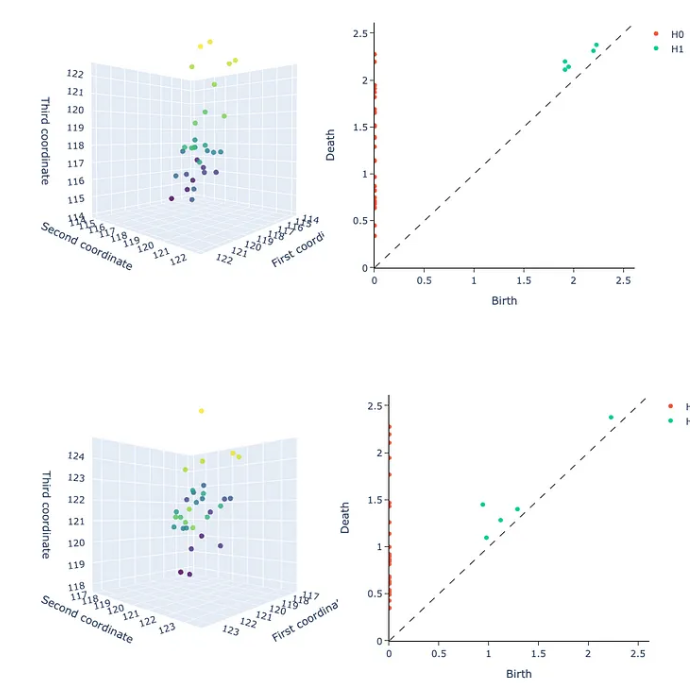


Table 7. Point clouds from two successive windows and their associated persistence diagram.

**Homological Derivative Model**

Homological Derivative: The model utilizes the concept of homological derivatives to analyse the evolution of topological features over time. By computing derivatives of topological summaries such as landscapes, it captures changes in the underlying topological structure of the data, providing insights into its dynamic behaviour.

Metric Parameters: The model specifies metric parameters such as the L^p norm (p=2), the number of layers (n\_layers=10), and the number of bins (n\_bins=1000) to define the landscape metric. These parameters influence the granularity and sensitivity of the landscape representation, allowing for fine-tuning based on the characteristics of the data.

Parallelization: The model leverages parallel computing (n\_jobs=-1) to accelerate the computation of homological derivatives, leading to faster processing and scalability. This parallelization enhances efficiency, particularly when dealing with large datasets or complex topological summaries.

Visualization: The model provides visualizations of landscape distances over time, allowing for the exploration of temporal patterns in the evolution of topological features. The plots enable the identification of trends, fluctuations, or anomalies in the topological structure of the data, facilitating interpretation and analysis.

A graph and chart of value

Description automatically generated with medium confidence

Table 8. Landscape Distances

**Betti Homological Derivative**

Betti Homological Derivative model offers a valuable approach to analysing the temporal evolution of topological features using Betti numbers. Its ability to quantify topological changes, control metric sensitivity, provide interpretability, and facilitate rigorous analysis justifies its use as a powerful tool for temporal topological data analysis.

Metric Parameters: The model specifies metric parameters such as the L^p norm (p=2) and the number of bins (n\_bins=1000) to define the Betti metric. These parameters influence the computation of distances between Betti numbers, affecting the sensitivity and granularity of the analysis.

Parallelization: Utilizing parallel computing (n\_jobs=-1), the model accelerates the computation of Betti homological derivatives, enhancing efficiency and scalability. This parallelization enables faster processing, particularly beneficial for handling large datasets or complex topological structures.

Visualization: The model provides visualizations of Betti homological derivatives over time, facilitating the exploration of temporal patterns in the evolution of Betti numbers.

These plots enable the identification of trends, fluctuations, or anomalies in the data's topological features, aiding interpretation, and analysis.

A graph of a graph of a graph

Description automatically generated with medium confidence

Table 9. Betti Curve Distances of the S&P 500 Index

**Crash Detection**

Crash Detection Comparison: The analysis compares two different metrics, Metric 1 and Metric 2, for crash detection based on specified threshold values. This comparison allows for evaluating the effectiveness of each metric in identifying potential crashes or significant changes in the data.

Time Range Selection: The analysis focuses on a specific time range from "1981-01-01" to "2022-01-01", enabling a comprehensive assessment of crash detection performance over a prolonged period. This extended timeframe enhances the robustness and reliability of the evaluation.

Distances Calculation: The distances used in the comparison are derived from two different sources: distances\_1 and distances\_2. These distances likely represent different aspects or characteristics of the data, providing diverse perspectives for crash detection analysis.

Visualization: The comparison results are visualized using a line plot, where each metric's distances are plotted over time. The plot distinguishes between crashes detected by Metric 1 only, Metric 2 only, and both metrics simultaneously, facilitating a clear interpretation of the comparison outcomes.

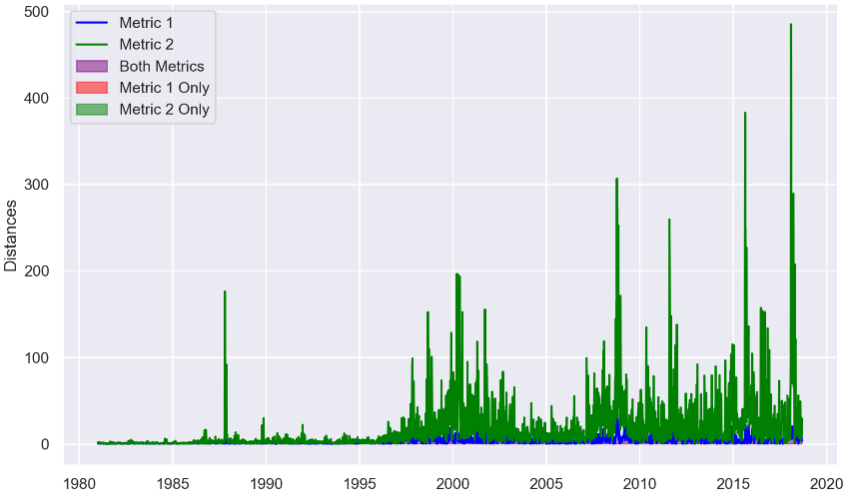


Table 10. Crash Comparisons

**Justification**

The model utilizes topological techniques to analyse the underlying structure and patterns within complex datasets. By leveraging concepts from algebraic topology, TPA offers a unique approach to understanding and interpreting high-dimensional data, by employing the Vietoris-Rips Persistence algorithm to compute persistent homology, a key aspect of TPA. This algorithm provides insights into the shape and connectivity of data points, allowing for the identification of topological features such as loops, voids, and clusters, creating powerful visualizations of persistence diagrams, which represent the lifespan of topological features across different scales. This visualization facilitates the interpretation of complex topological structures within the data, enabling researchers to gain deeper insights into its intrinsic properties. TPA allows for customization of parameters such as homology dimensions, enabling flexibility in the analysis based on specific research objectives and characteristics of the dataset. This parameterization enhances the adaptability and applicability of the TPA model to diverse datasets and research contexts.

**8**.**3** **Random Forrest Classifier**

The RF model was proposed by Breiman (2001) as an improved form of the decision tree. It has many applications in solving classification and regression problems and need to optimize a few parameters.

There are two parameters in the RF model that generally affect the performance of the model: the number of trees ntree and number of candidate variables randomly sampled at each split ntry. The value of ntry was suggested by p3/3, where p is the number of inputs (Dudek, 2015). For forecasting a time-series Yt, the autoregressive random forest (AR-RF) model will be used and denoted as AR-RF(p), where p is number of AR lags. In comparison with other machine learning models, RFs provide the greater precision with the ability to handle big data with numerous variables running into thousands.

Evaluation, result and conclusion:  
Model Accuracy: The accuracy of the model on the testing set is 1.00, indicating that all instances were correctly classified. However, accuracy alone may not provide a complete picture of the model's performance, especially when dealing with imbalanced datasets.

Classification:

Precision: Precision for class 0 (no market direction change) is 1.00, indicating that all instances predicted as class 0 were indeed class 0. However, precision for class 1 (market direction change) is 0.00, suggesting that none of the instances predicted as class 1 were correct.

Recall: Recall for class 0 is 1.00, indicating that all actual instances of class 0 were correctly classified. Recall for class 1 is also 0.00, meaning that none of the actual instances of class 1 were correctly classified.

F1-Score: The F1-score for class 0 is 1.00, reflecting the harmonic mean of precision and recall for class 0. However, the F1-score for class 1 is 0.00, indicating poor performance in correctly classifying instances of class 1.

Support: Class 0 has a high support (3414 instances), while class 1 has very low support (only 4 instances).

Feature Importances: The feature importances plot shows the importance of each feature in making predictions. In this case, the model seems to rely heavily on the moving average features, with 'MovingAverage\_200' being the most important feature.

Justification:

Accuracy: While the accuracy of 1.00 seems impressive, it is misleading due to the highly imbalanced dataset. The model achieves perfect accuracy by correctly classifying all instances of the majority class (class 0) but fails to correctly classify any instances of the minority class (class 1).

Precision and Recall: The model's poor precision and recall for class 1 indicate its inability to correctly identify instances of market direction change. This suggests that the model may not be suitable for practical use, especially in financial decision-making where correctly identifying market direction changes is crucial.

Feature Importance: The reliance of the model on moving average features suggests that these features are highly correlated with the target variable. However, the poor performance of the model indicates that other factors not captured by the provided features may be influencing market direction changes.

In conclusion, while the model achieves high accuracy, its poor performance in correctly classifying instances of market direction change, as indicated by precision, recall, and F1-score, undermines its practical utility.

**8.4 Backward Supremum Augmented Dickey-Fuller (BSADF) test**

This test not only estimate the Dicky fuller test for the whole sample but the econometric concept behind detecting bubbles with augmented Dickey Fuller has to do with the right tail of the tail distribution. Looking at the right-hand side of the distribution it is possible to see when return generating process is exposed, positive price increments lead to even higher positive price increments, and that by definition constitutes a bubble or explosive behaviour in price increments or returns, and that is the logic of the BSADF. The parameter that is quite important for the BSADF is the R 0 parameter, which is basically the smallest increment considered to estimate augmented Dickey fuller tests on and the most commonly it is defined as a proportion of the total sample which is corresponding to the length of the prices array and one of the most common specifications is to select the r0 as 10% of the length of the sample. To filter out short term dependence of returns which is again very important for stock markets, this test look at the augmented Dickey fuller test with 3 Lags, filtering out short-term autocorrelation and just leave the explosive behaviour that will be tested against the right tail critical value of the Augmented Dickey Fuller distribution and that has been estimated already using bootstrapping by Phillips et Al. (2015) and for r0 equal to 10%, the critical value of the right hand side of ADF distribution is 1.49. When encountering explosive behaviour, this test will compare ADF statistic with 1.49 and if it exceeds 1.49, reasonably if at 5% with 95% certainty there is a bubble. Rejecting the null hypothesis of no explosive behaviour in favour of the alternative Ho that there is bubble like behaviour in the prices dynamics. To make the tests more robust, natural log was applied to the prices to get log prices and then the differences of increments in log prices can be naturally interpreted as log returns, making the test more resilient, more robust to estimations in varying windows. Next stage was to calculate the array of backwards sub augmented Dickey fuller test statistics and those would be maxima of augmented Dickey fuller statistics estimated on variant subsamples defined by R0 and the number of observations n, followed by moving across all potential sub samples leading up to a particular date in the range and estimate all possible ADF test statistics using the number of lags specified and the usual T stat from ADF test. X or the independent variables, are the lagged log prices using a loop to augment the data frame of X with lagged increments of log prices. The dependent variable is the log price increments, so log returns was regressed on log prices, which is the variable of interest and the left increments which are control variables to filter out short term dependence and get left only with long term bubble like explosive behaviour. A regression was specified using the OLS function from the stats models API package and which was estimated based on dependent variable Y and independent variable X which is our X0. Lagged 0 prices and lagged increments.

Next stage was to fit the regression and the only parameter we are interested in is the ADF T stat, so append ADFS array which is augmented Dickey-Fuller statistics data frame with TSTAT which is a ratio between the parameter with index one which is the second parameter so constant and then X0, dividing it by bse which retrieves standard errors of coefficients.

Method applied:

The test iterates backward through the data, computing ADF statistics for multiple sub-samples of the time series.

The maximum ADF statistic of sub-samples was compared against the critical value of 1.49 to determine the presence of a unit root, indicating non-stationarity.

2. Data Preprocessing:

Logarithmic returns were computed to stabilize the variance and normalize the data for stationarity testing.

3. Model Implementation:

For each sub-sample, the BSADF test regresses the differenced log prices against lagged differences using Ordinary Least Squares (OLS).

The ADF statistic was then computed for each sub-sample, and the maximum value across all sub-samples was retained.

The pre-defined critical value was compared against the maximum ADF statistic to assess stationarity.

4. Results Interpretation:

The BSADF test results suggest that the null hypothesis of a unit root is rejected for the dates '2021-12-10' and '2021-12-15'. In simpler terms, rejecting the null hypothesis in the BSADF test implies that the time series has undergone significant structural changes or breaks. In this context, it suggests that current stock prices may not be sustainable.

These dates indicate that during those periods, the S&P 500 index was stationary, implying a lack of a unit root and mean-reverting behaviour.

However, non-rejection of the null hypothesis for other dates may suggest the presence of a unit root or non-stationarity during those periods.

A graph showing the growth of the stock market

Description automatically generated

Table 11. BSADF result

**Result**

Identifying the start and end dates of explosive behaviour in stock prices depends on the minimum duration of exuberance. In this context, an origin date was established when the time series of recursive test statistics BSADF, characterized by r values within the range of [r0,1], surpassed the critical value associated with the statistics. Similarly, a termination date was designated when the critical value of BSADF with r values spanning [r0,1] exceeds the corresponding test statistics. Philips and Shi (2020) noted that the origination of a bubble or crisis episode is determined as the point when the BSADF test statistic first surpasses its critical value, while the termination date corresponds to the point when the supremum test statistic subsequently falls below its critical value, establishing two distinct stopping times for the episode.

Table 11 shows the relationship between critical values from the Monte Carlo simulation (Table 2) and BSADF test statistics for each of the presented stocks. The x-axis represents the year while the y-axis represents the BSADF test critical value. The orange line is the critical value for the specific stocks for the specific date, the blue line is the BSADF test statistics, being possible to visualize where they intersect, suggesting potential bubble-like behaviour in analysed stocks.

**Justification**

The BSADF test provides valuable insights into the stationarity of the S&P 500 index, aiding in identifying mean-reverting behaviour and potential trading opportunities.

Rejection of the null hypothesis for specific dates suggests periods of stationarity, while non-rejection indicate non-stationarity.

Overall, the BSADF test offers a useful tool for assessing stationarity in financial time series data, contributing to informed investment strategies and risk management.

**8.5** **Generalized Autoregressive Conditional Heteroskedasticity (GARCH)**

The GARCH model effectively captures the time-varying nature of volatility in the S&P 500 closing prices by considering past squared residuals and lagged conditional variances.

Interpretable Model parameters:

mu: This parameter represents the mean or expected return of the asset. In this context, it signifies the average return rate of the asset over the given period.

omega: Omega is the constant term in the GARCH model and represents the long-term average variance of the asset returns. It's the coefficient of the lagged squared residual term in the volatility equation. It indicates the unconditional volatility of the asset returns.

alpha: Alpha is the coefficient of the lagged squared return in the volatility equation. It measures the impact of past squared returns on the conditional variance of the asset returns. A higher alpha indicates a stronger persistence in volatility, meaning past volatility shocks have a more significant effect on current volatility.

beta: Beta is the coefficient of the lagged conditional variance term in the volatility equation. It measures the speed of adjustment of the conditional variance towards its long-term average. A higher beta indicates a faster mean reversion of volatility towards its long-term average.

long-run volatility: This is calculated as the square root of omega divided by one minus alpha minus beta. It represents the long-term equilibrium volatility of the asset returns implied by the GARCH model.

log-likelihood: The log-likelihood value represents the maximized log-likelihood function of the GARCH model. It is a measure of how well the model fits the observed data. A higher log-likelihood indicates a better fit.

Findings:

1. mu: The estimated mean of the returns is 0.000775.

omega: The estimated constant term in the GARCH model, representing the long-term average variance, is 4e-06.

alpha: The estimated coefficient of the lagged squared residuals in the GARCH model is 0.2274. This parameter captures the short-term impact of shocks on volatility.

beta: The estimated coefficient of the lagged conditional variance in the GARCH model is 0.6966. This parameter captures the long-term persistence of volatility.

long-run volatility: The estimated long-run volatility, derived from the GARCH model parameters, is 0.0077. This represents the equilibrium level of volatility towards which the conditional volatility converges.

2. Log-Likelihood:

The log-likelihood of the GARCH model is 2483.6002. This value indicates how well the model fits the observed data. A higher log-likelihood suggests a better fit.

3. Plot of Realized and Conditional Volatilities:

The plot visually compares the realized (actual) volatilities with the conditional volatilities predicted by the GARCH model.

If the GARCH model is accurate, the conditional volatilities should closely track the realized volatilities.

Any significant deviations between the two lines indicate model inadequacy or missed dynamics in the volatility process.

The GARCH model parameters provided insights into the dynamics of volatility in the time series.

The estimated coefficients (alpha and beta) indicate the persistence and impact of shocks on volatility, respectively.

The long-run volatility gives an indication of the equilibrium level of volatility in the series.

The log-likelihood helps assess the goodness-of-fit of the model to the data.

A graph of a graph showing the number of the same numbers

Description automatically generated with medium confidence

Table 12. GARCH model

**Justification**

GARCH model first captured the conditional heteroskedasticity observed in financial returns, where volatility clusters in time. This was evident from the parameters, such as the significant values for omega (4e-06), alpha (0.2274), and beta (0.6966), which represented the persistence of volatility shocks. Secondly, the model's log-likelihood value of 2483.6002 indicated a good fit to the data, suggesting that the GARCH structure adequately explains the variance in returns. Additionally, the plot of realized volatility versus conditional volatility demonstrated that the model captures the dynamics of volatility over time. These applied methods indicate that the provided GARCH model is appropriate for capturing and predicting the volatility of the S&P 500 index, making it a valuable tool for risk management and financial analysis.

**9.** **Conclusion**

This study has investigated a range of machine learning and statistical models tailored to the identification and prediction of stock market bubbles.

Log Periodic Power Law model offered insights into the presence of speculative behaviour and potential bubble formations, serving as a vital tool for bubble detection. The model defined a crash as an observation in the empirical quantile of 99.5%. Crashes are often associated with drawdowns larger than 10% and all of the observations in the quantile are consistent with historical crashes and bubbles. The benefit of this definition is that it does not depend on any specific distribution, only the empirical distribution. By choosing a good value of E and reduce the presence of noise, it was possible to fit three parameter Weibull distribution.

Topological Data Analysis presented an innovative approach to discerning geometric structures within financial data, aiding in the identification of market anomalies and bubble patterns. Random Forest Classifier contributed a robust predictive framework for anticipating market movements and identifying key features indicative of bubble dynamics.

Additionally, the BSADF model provided statistical rigor in detecting structural breaks associated with bubble formation. This research suggests potential avenues for future research, including exploring online learning and adaptive models for dynamic market conditions. Cross-market dynamics analysis examining interconnections among different asset classes could yield insights into bubble occurrences. Further research could explore international spillovers and contagion effects by analysing transmission mechanisms for stock bubbles across markets and regions, considering factors like cross-border capital flows and information dissemination.

Finally, the GARCH model offered a powerful means to model volatility dynamics, crucial for understanding the risk associated with bubble phenomena. Together, these models offered a comprehensive toolkit for detecting, understanding, and predicting stock market bubbles, providing valuable insights for investors and policymakers in managing bubble-related risks.

**10.** **Future Research**

Several promising avenues for future research emerged from the models utilized in this study. Firstly, further exploration into ensemble methods such as combining the strengths of the Random Forest Classifier with deep learning techniques could enhance predictive accuracy and robustness, particularly in identifying complex patterns associated with stock market bubbles.

Additionally, extending the application of Topological Data Analysis (TDA) to other financial datasets and markets may uncover novel insights into market dynamics and bubble formation, offering new perspectives on risk assessment and market stability.

Moreover, integrating sentiment analysis of news and social media data into the GARCH model framework could provide a more comprehensive understanding of the behavioural factors influencing market bubbles. Finally, investigating the impact of regulatory policies and macroeconomic indicators on bubble formation and mitigation strategies could contribute to the development of proactive measures to prevent and manage future market bubbles.

**References**

Anh, L.H., Le Si Dong, Vladik Kreinovich and Nguyen Ngoc Thach (2017). Econometrics for Financial Applications. Springer.

Biljanovska, N., Lucyna Gornicka and Alexandros Vardoulakis (2019). Optimal Macroprudential Policy and Asset Price Bubbles. International Monetary Fund.

Bodie, Z. and Merton, R.C. (2000). Finance.

Brunnermeier, M. and Oehmke, M. (2012). Bubbles, Financial Crises, and Systemic Risk. [online] doi:https://doi.org/10.3386/w18398.

Caballero, R.J. and Krishnamurthy, A. (2005). Bubbles and Capital Flow Volatility.

Chen, S.-W., Hsu, C.-S. and Xie, Z. (2016). Are there periodically collapsing bubbles in the stock markets? New international evidence. Economic Modelling, 52, pp.442–451. doi:https://doi.org/10.1016/j.econmod.2015.09.025.

Cha Zhang and Ma, Y. (2012). Ensemble Machine Learning. New York: Springer.

Corchado, E., Abraham, A. and Witold Pedrycz (2008). Hybrid artificial intelligence systems : third international workshop, HAIS 2008, Burgos, Spain, September 24-26, 2008 : proceedings. Berlin: Springer.

Demirer, R., Demos, G., Gupta, R. and Sornette, D. (2018). On the predictability of stock market bubbles: evidence from LPPLS confidence multi-scale indicators. Quantitative Finance, 19(5), pp.843–858. doi:https://doi.org/10.1080/14697688.2018.1524154.

Demos, G. and Sornette, D. (2016). Birth or burst of financial bubbles: which one is easier to diagnose? Quantitative Finance, 17(5), pp.657–675. doi:https://doi.org/10.1080/14697688.2016.1231417.

Didier Sornette (2017). Why stock markets crash : Critical events in complex financial systems. Princeton, New Jersey: Princeton University Press.

Dunis, C.L., Middleton, P.W., Karathanasopolous, A. and Konstantinos Theofilatos (2016). Artificial Intelligence in Financial Markets Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics. London Palgrave Macmillan Uk :Imprint: Palgrave Macmillan

Economou, F., Konstantinos Gavriilidis, Gregoriou, G.N. and Vasileios Kallinterakis (2017). Handbook of Investors’ Behavior during Financial Crises. Saint Louis Elsevier Science.

Emmons, W.R. and Noeth, B.J. (2012). Household Financial Stability: Who Suffered the Most from the Crisis? [online] Stlouisfed.org. Available at: <https://www.stlouisfed.org/publications/regional-economist/july-2012/household-financial-stability--who-suffered-the-most-from-the-crisis>.

Evanoff, D.D., Kaufman, G.G. and Malliaris, A.G. (2012). New Perspectives on Asset Price Bubbles. New York Oxford University Press -03-01.

Frömmel, M. (2023). Finance 2: Asset Allocation and Market Efficiency. BoD – Books on Demand.

Flood, R.P. and Garber, P.M. (1994). Speculative Bubbles, Speculative Attacks, and Policy Switching. Cambridge, Massachusetts: Mit Press.

Fuller, W.A. (1976). Introduction to statistical time series. New York Etc.: John Wiley & Sons, Cop.

Garber, P.M. (2001). Famous first bubbles : the fundamentals of early manias. Cambridge, Massachusetts ; London, Uk: Mit Press.

Gerdesmeier, D., Reimers, H.-E., Roffia, B. and Diskussionspapiere, W. (2013). Testing for the existence of a bubble in the stock market. [online] Available at: https://www.econstor.eu/bitstream/10419/76657/1/749488018.pdf [Accessed 22 Feb. 2024].

Herzberg, A. (2014). Sustainability of External Imbalances. Springer.

Jagat Prirayani (2017). Econometric Modeling and Forecasting.

Jansen, S. (2020). Machine learning for algorithmic trading predictive models to extract signals from market and alternative data for systematic trading strategies with Python, second edition. Packt Publishing.

Johansen, A. and Didier Sornette (2002). Endogenous Versus Exogenous Crashes in Financial Markets.

JOHANSEN, A., LEDOIT, O. and SORNETTE, D. (2000). CRASHES AS CRITICAL POINTS. International Journal of Theoretical and Applied Finance, [online] 03(02), pp.219–255. doi:https://doi.org/10.1142/s0219024900000115.

Kaizoji, T. and Sornette, D. (n.d.). Market Bubbles and Crashes. [online] Available at: https://arxiv.org/ftp/arxiv/papers/0812/0812.2449.pdf [Accessed 25 May 2021].

Kindleberger, C.P. and Aliber, R.Z. (2015). Manias, Panics, and Crashes: A History of Financial Crises, Seventh Edition. Palgrave Macmillan Ltd.

Kyrtsou, C., Didier Sornette and Adcock, C. (2020). New Facets of Economic Complexity in Modern Financial Markets. Routledge.

Li, C. (2017). Log-periodic view on critical dates of the Chinese stock market bubbles. Physica A: Statistical Mechanics and its Applications, 465, pp.305–311. doi:https://doi.org/10.1016/j.physa.2016.08.050.

Mathias Kende (2018). The trade policy review mechanism : a critical analysis. Oxford ; New York, Ny: Oxford University Press.

Pierson, L. and Porway, J. (2017). Data science. Hoboken, Nj: John Wiley And Sons, Inc.

Razin, A. (2014). Understanding global crises : an emerging paradigm. Cambridge, Mass. ; London: The Mit Press, Cop.

Salge, M. (2012). Rational Bubbles. Springer Science & Business Media.

Shone, R. (2002). Economic Dynamics. Cambridge University Press.

Shu, M. and Zhu, W. (2020). Detection of Chinese stock market bubbles with LPPLS confidence indicator. Physica A: Statistical Mechanics and its Applications, 557, p.124892. doi:https://doi.org/10.1016/j.physa.2020.124892.

Tarlie, M.B., Sakoulis, G. and Henriksson, R. (2018). Stock market bubbles and anti-bubbles. International Review of Financial Analysis. doi:https://doi.org/10.1016/j.irfa.2018.07.012.

Vogel, H.L. (2018). Financial Market Bubbles and Crashes, Second Edition. Springer.

Vogel, H.L. (2021). Financial market bubbles and crashes : features, causes, and effects. Cham, Switzerland: Palgrave Macmillan.

Wang, J. and Wang, S. (2010). Business intelligence in economic forecasting. Hershey, PA: Business Science Reference.

William, G. (1991). Governing Big Cities.

Zhang, Q., Sornette, D., Balcilar, M., Gupta, R., Ozdemir, Z.A. and Yetkiner, H. (2016). LPPLS bubble indicators over two centuries of the S&P 500 index. Physica A: Statistical Mechanics and its Applications, 458, pp.126–139. doi:https://doi.org/10.1016/j.physa.2016.03.103..

Zhao, Z., Wen, H. and Li, K. (2021). Identifying bubbles and the contagion effect between oil and stock markets: New evidence from China. Economic Modelling, [online] 94, pp.780–788. doi:https://doi.org/10.1016/j.econmod.2020.02.018.