Quantitative Finance

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rquf20

Detecting bubbles via FDR and FNR based on calibrated *p*-values

Giulia Genoni, Piero Quatto & Gianmarco Vacca

To cite this article: Giulia Genoni, Piero Quatto & Gianmarco Vacca (2024) Detecting bubbles via FDR and FNR based on calibrated *p*-values, Quantitative Finance, 24:10, 1463-1491, DOI: 10.1080/14697688.2024.2406561

To link to this article: https://doi.org/10.1080/14697688.2024.2406561

Quantitative Finance, 2024
Vol. 24, No. 10, 1463–1491, https://doi.org/10.1080/14697688.2024.2406561

Detecting bubbles via FDR and FNR based on calibrated *p*-values

GIULIA GENONI†, PIERO QUATTO‡ and GIANMARCO VACCA *§

†Department of Economics, Università della Svizzera Italiana, Lugano, Switzerland
‡Department of Economics, Management and Statistics, Università degli Studi di Milano-Bicocca, Milano, Italy
§Department of Economic Policy, Università Cattolica del Sacro Cuore, Milano, Italy

(*Received 10 November 2023; accepted 12 September 2024*)

Detecting bubbles in asset prices is still an open question that has attracted considerable attention in recent years. This paper improves the bubble detection and dating approaches developed in recent years by Phillips and co-authors, proposing to assess the plausibility of its outcomes via the false discovery rate (FDR) and the false non-discovery rate (FNR) based on calibrated *p*-values. Calibrating *p*-values of unit root tests, applied sequentially to detect bubbles, allows recovery of their super-uniformity property, which is crucial for a valid implementation of the inferential procedure. The paper also develops original self-calibrated versions of both FDR and FNR for the specific problem of bubble testing. Calibrated *p*-values are implemented in an online false discovery-based approach which monitors bubbles in real time. The effectiveness of the proposed methods is investigated via a simulation study and an empirical application.

Keywords: Bubble dating; Unit root test; False discovery rate; False non-discovery rate; Online multiple tests

JEL Classification: C12, C58

1. Introduction

The problem of detecting and dating financial bubbles plays a crucial role in economics. A bubble occurs when asset prices deviate from fundamental values and show explosiveness (Stiglitz 1990). Similarly, sudden rises in asset prices followed by sharp collapses are identified as bubbles (Brunnermeier 2016, Su *et al.* 2017). Some crises start with the burst of a bubble, such as the German stock price bubble of 1927, the Wall Street Crash of 1929, the Japanese asset price bubble of the 1980s to 1990s, the dot-com bubble in 2001 and the USA mortgage crisis bubble of 2008. When bubbles emerge, market price indexes diverge from fundamental values, as shown by Jarrow *et al.* (2010) using the martingale theory (Jarrow and Protter 2010). This divergence can lead investors, financial institutions, and regulators to make serious errors in decision-making.

Several studies have been carried out on this topic (see, e.g. Gürkaynak 2008, Scherbina and Schlusche 2014, for an overview of recent literature on bubbles). Some of the earliest bubble detection methods include the variance-bound tests put forward by Shiller (1981), LeRoy and Porter (1981), Diba and Grossman (1984) and Hamilton and Whiteman (1985), who proposed stationarity tests to detect bubbles, and Campbell and Shiller (1987), who adopted unit root and cointegration tests. Evans (1991) pointed out that these tests suffer from a serious limitation. Indeed, due to their low power, they cannot detect periodically collapsing rational bubbles, as these processes can behave much like unit-root or even stationary linear autoregressive process provided that the probability of collapse of the bubble is not negligible. Examining the behavior of rational bubbles, as it emerges from the value theory of finance, Phillips and Yu (2011) and Phillips *et al.* (2011) concluded that the explosive behavior of stock prices can be modeled by measuring the change in the degree of non-stationarity in their generating process. The process is a random walk before the appearance of the bubble, becomes an 'explosive' process when the bubble period begins, and finally reverts to a random walk after the bubble collapses. Considering that, as suggested by Evans (1991), the explosive behavior is only temporary when bubbles periodically collapse, Phillips argued that they could be effectively detected and dated using the supremum of right-tailed unit root tests applied recursively on sub-samples of increasing length. Extensive simulation studies have shown that this approach is especially effective in detecting a single bubble episode, which was the core

*Corresponding author. Email: gianmarco.vacca@unicatt.it

of Phillips's seminal studies (Phillips and Yu 2011, Phillips *et al.* 2011). Later, he generalized this approach to the case of multiple bubbles and other variants (see Phillips *et al.* 2015, Phillips and Shi 2019, 2020), extensions and improvements were proposed over recent years (see Hu 2023, for a detailed survey on the latest contributions on this topic).

Therefore, the problem of effectively detecting and dating bubbles can be framed as sequentially identifying structural breaks in a time series as new data becomes available. However, since Phillips's approach hinges on multiple testing, it exhibits all the typical problems associated with this type of inference. As is well known, in a testing procedure based on a single hypothesis, a discovery, or rejection of the null hypothesis, is reported if the p-value is lower than some pre-selected threshold, say α. The latter guarantees that the probability of the discovery being false is less than α. In case of multiple hypotheses, this threshold does not provide a similar assurance, and the fraction of false discoveries may be arbitrarily large. In a multiple testing procedure, several error rates can be employed to measure and control false rejections of the null hypothesis. Among them, the most prevalent is the false discovery rate (FDR), which is the expected value of the false discovery proportion, or equivalently the proportion of true nulls that are wrongly rejected (see Barras *et al.* 2010, Bajgrowicz *et al.* 2016, for a discussion of the advantages of the FDR over other multiple testing control methods). Similarly, FNR is the expected value of the proportion of false alternative hypotheses that are classified as true by the test. These procedures can be applied either 'offline', that is using historical observations, or online, as new observations become available, by allowing users to quickly decide in real time what action to take in response to a stream of data.

In the paper, both these statistics are applied to assess the reliability of the outcomes of the Augmented Dickey Fuller (ADF) test used in Phillips's approach to date bubbles, with the aim to improve its performance. This target is achieved by calibrating the p-values of the unit root tests employed sequentially to the observations of a series. Indeed, simulation experiments prove that the p-values of the ADF test, when applied recursively, as required by the Phillips procedure, lack of validity, as they do not exhibit a super-uniform cumulative distribution function (CDF) under the null. This work addresses this drawback, that can hamper the validity of the inference, by calibrating the p-values using an unbiased estimate of their CDF under the null hypothesis. This estimate is obtained by using a training period of the sample that does not exhibit bubbles. The distribution of the p-values is then used to calibrate those of the ADF tests when applied sequentially for real time bubble detection of the remaining data sample. This approach is similar to that of Astill *et al.* (2018), who employed a training period to derive the critical values of the AHLST test proposed by Astill *et al.* (2017) for detecting bubbles' explosion. This approach shares also similarities with Whitehouse *et al.* (2023), who suggested using the critical value from a corrected AHLST test (derived from a bubble-free training period) to identify the end of explosive behavior and the beginning of the stationary regime. To account for potential heteroskedasticity in data that may affect the performance of the ADF test, the driving shocks of the data generating process are assumed to be conditionally heteroskedastic as well as possibly heavy-tailed.

The use of calibrated p-values introduces a significant improvement in the performance of Phillips's approach to bubble dating. Calibrated p-values are also used to implement both FDR and FNR, obtaining estimates of the latter that satisfy given thresholds. Simulation studies prove that the outcomes of these statistics outperform those based on original p-values.

Self-calibrated versions of both FDR and FNR, that require less computational effort, are also provided, and their asymptotic properties are investigated. Finally, calibrated p-values are further employed in an online FDR approach, known as LORD (Javanmard and Montanari 2018), which controls FDR by modifying the significance level according to the outcomes of the ADF test.

The performance of all the aforementioned methods is analyzed by using both simulated and real data. It emerges that the performance of Phillips's test (Phillips and Yu 2011) improves significantly if calibrated p-values are employed. In addition, it is evident from the results obtained that both offline and online approaches, based on false discoveries and that use calibrated p-values, are useful in validating the outcomes of Phillips's test. In fact, these methods allow for more accurate dating of the origination and ending of bubble periods, regardless of their structural characteristics of being less or more pronounced.

This result is even more significant when considering that the outcomes of Phillips's test are not robust with respect to data sampling frequency, as demonstrated in the empirical application.

The paper is organized as follows. Section 2 explains how the FDR and FNR approaches can be developed for multiple testing based on unit root tests, according to the approach proposed by Phillips and Yu (2011). The role of calibrated p-values in improving the performance of the ADF tests is here analyzed. Conservative estimates of both FDR and FNR are devised and their properties are investigated. An online version of the false discovery approach is also proposed for bubble dating. An analysis of the performances of the mentioned methods based on false discoveries is carried out in section 3 using simulated data. Section 4 offers an empirical analysis, based on the financial data employed in Phillips *et al.* (2011), which presents interesting results regarding the robustness of the methods considered in the paper. Section 5 provides some conclusions. Finally, appendix contains additional material pertaining to the application.

2. Bubble detection based on false discoveries

The aim of this section is to provides a solution to detecting bubbles following two different approaches. The first is based on the FDR and the FNR (Storey 2002, Efron 2010) computed with the p-values of the Augmented-Dickey-Fuller (ADF) test, according to the procedure proposed by Phillips and Yu (2011). The second is based on LORD, which is an online/real time testing procedure introduced by Javanmard

and Montanari (2018) that assures, under suitable conditions, guarantees the control for false discoveries.

2.1. The analytical framework: Phillips's test and its extensions

A series of papers (see Phillips and Yu 2011, Phillips *et al.* 2011, 2015, Phillips and Shi 2020) have proposed recursive procedures based on unit root tests for identifying and dating financial bubbles in real time

The basic concept underlying these procedures is that in the absence of bubbles, asset prices are unit root processes, while during bubble explosions they behave like mildly explosive processes. According to Phillips and Magdalinos (2007) and Magdalinos (2012), mildly explosive processes can be modeled using the following auto-regressive process

$$y_t = \theta y_{t-1} + \epsilon_t, \tag{1}$$

with a root slightly exceeding the unity,

$$\theta = 1 + \frac{c}{K_n}, \quad c > 0, \quad \frac{c}{K_n} \to 0 \quad \text{as } K_n \to \infty \tag{2}$$

where $\{K_n\}_{n\in N}$ is a sequence increasing for $n \to \infty$ such that $K_n = o(n)$ as n diverges.

Accordingly, the reference model for testing and dating bubbles can be specified as follows

$$y_t = \alpha_0 T^{-\gamma} + \theta y_{t-1} + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d.}(0, \sigma^2), \tag{3}$$

where T is the sample size, γ is a localization parameter controlling the magnitude of the intercept and drift as $T \to \infty$. When $\gamma > \frac{1}{2}$ the drift component turns out to be dominated by the stochastic trend.

The model (3) is usually complemented with transient dynamics explaining the switching from unit root to mildly explosive behavior. Thus, in case of a single explosive episode, starting at τ_o and ending at τ_f, Phillips and Yu (2011) proposed the following mechanism

$$y_t = y_{t-1} I_{t<\tau_o} + \delta_\tau y_{t-1} I_{\tau_o \le t \le \tau_f} + (\xi_{\tau_f} + y^*_{\tau_f}) I_{t>\tau_f} + \epsilon_t I_{t\le\tau_f}. \tag{4}$$

Here, I_A is the indicator function on A, $\delta_\tau = 1 + cT^{-\eta}$ with $c > 0$, $\eta \in (0, 1)$, $y^*_{\tau_f} = y_{\tau_f} + y^o$, with $y^o = Op(1)$ and $\xi_{\tau_f} = \sum_{k=\tau_f+1}^{t} \epsilon_k$ is a stochastic trend.

According to (4), y_t is a random walk in the pre-bubble period and becomes a mildly explosive process during the bubble's explosion. Once the bubble collapses, y_t switches to $y^*_{\tau_f}$, the value that prevailed prior to the emergence of the bubble and continues its random martingale behavior over the subsequent period. The model (4) can be extended to the case of multiple bubbles.

To identify the occurrence of a bubble, a sequence of ADF_0^r tests, based on an increasing fraction $r \in [0, 1]$ of the sample observations, is performed in the following re-parametrized version of the model (3)

$$\nabla y_t = \alpha + \delta y_{t-1} + \sum_{i=1}^{k} \psi_i \nabla y_{t-i} + \epsilon_t, \quad \epsilon_t \sim \text{i.i.d.}(0, \sigma^2),$$
$$t = 1, 2, \ldots, T. \tag{5}$$

Under the null of no bubble explosion, y_t is assumed to be a unit root process. In contrast, under the alternative it is a mildly explosive process

$$H_0 : \delta = 0$$
$$H_1 : \delta > 0. \tag{6}$$

Accordingly, the test is right-sided. The starting point of each recursive regression is $t_0 = [Tr_0]$, where $[\cdot]$ denotes the integer part of the argument and r_0 is a fraction of the sample size. The latter represents the smallest sample width fraction in the recursion and 1 the largest one as it corresponds to a regression based on the full sample.†

As for the unit root test, either the coefficient test statistic or the t-statistic, denoted by ADF_r^δ, ADF_r^t respectively, can be used for this purpose

$$\text{ADF}_r^\delta = \tau(\tilde{\delta}_\tau - 1), \quad \text{ADF}_r^t = \left(\frac{\sum_{j=1}^{\tau} \tilde{y}_{j-1}^2}{\tilde{\sigma}_\tau^2}\right)^{1/2} (\tilde{\delta}_\tau - 1), \tag{7}$$

where $\widetilde{\delta}_\tau$ is the least squares estimate of δ in (5) based on the first $\tau = [Tr]$ observations, $\widetilde{\sigma}_\tau^2$ is the estimate of σ^2 based on the same sample, and $\widetilde{y}_{j-1} = y_{j-1} - \tau^{-1} \sum_{j=1}^{\tau} y_{j-1}$.

The occurrence of a bubble is then detected via the Supremum Augmented Dicky Fuller test (SADF, Phillips and Yu 2011)

$$\text{SADF}(r_0) = \sup_{r\in[r_0,1]} \text{ADF}_0^r \to \sup_{r\in[r_0,1]} \times \frac{\frac{1}{2}[W(r)^2 - r] - \int_0^r W(s)\,\mathrm{d}s W(r)}{r^{1/2}\{r\int_0^r W(s)^2\,\mathrm{d}s - [\int_0^r W(s)\,\mathrm{d}s]^2\}^{1/2}}, \tag{8}$$

where $W(\cdot)$ is the standard Brownian motion.

The SADF test statistic cannot locate the origination and collapse dates of a bubble. The date stamping procedure is performed conditional on this global test rejecting the null hypothesis. If the global test fails to accept the null, the origination date of the bubble can be determined by looking for the infimum value r such that the p-value of the corresponding test leads to the rejection of H_0 at a given significance level α, that is

$$\widehat{r}_0 = \inf\{r \ge r_0 : p_r < \alpha\} \tag{9}$$

where p_r is the p-value of an ADF test based on a regression using Tr sample observations.‡

† r_0 depends on the sample size. It must be large if the sample size is small and vice-versa. This rule guarantees, on the one hand, enough observations for the initial estimation, while avoiding the possibility of missing the detection of a bubble originating at the beginning of the sample.

‡ Equation (9) can be also expressed as

$$\widehat{r}_0 = \inf\{r \ge r_0 : \text{ADF}_r > c_{1-\alpha}\}$$

where $c_{1-\alpha}$ is the $1-\alpha$ percentile of the limiting distribution of the ADF statistic

$$P(\text{ADF}_r \ge c_{1-\alpha} | H_0) = \alpha$$

Conditional on finding some origination date $\widehat{t}_0 = [T\widehat{r}_0]$, the ending of an explosive bubble can be determined by looking for the infimum value r_f, among all values that are greater than $\widehat{r}_0 + \beta\frac{\log T}{T}$ at level α, that leads to the acceptance of the null†

$$\widehat{r}_f = \inf\left\{r \geq \widehat{r}_0 + \beta\frac{\log T}{T} : p_r > \alpha\right\}. \tag{10}$$

Here, $\beta \log T$ is assumed to be the minimal duration of a bubble which depends on a parameter β, related to the data frequency. A drawback of the test is its inability to distinguish between two bubbles when they are not separated by a large interval.

In case of a single bubble, the authors proved that, under suitable conditions,‡ the following holds

$$(\widehat{r}_0, \widehat{r}_f) \xrightarrow{p} (r_0, r_f). \tag{11}$$

This procedure was later enhanced by Phillips *et al.* (2015), who suggested implementing a recursive evolving algorithm, the so called generalized supADF (GSADF), where the start point r_1 vary within a feasible range, depending on the end point r_2 of the sample

$$\text{GSADF}_{r_2}(r_0) = \sup_{\substack{r_1\in[0,r_2-r_0]\\ r_2\in[r_0,1]}} \text{ADF}_{r_1}^{r_2}. \tag{12}$$

This test also has the crucial limitation of not providing the origination and termination dates of the bubbles.

To achieve this goal, a further recursive test, the backward ADF test, BSADF test in short, must be implemented. The BSADF test is defined as the sup value of the ADF statistic sequence

$$\text{BSADF}_{r_2}(r_0) = \sup_{r_1\in[0,r_2-r_0]} \text{ADF}_{r_1}^{r_2}. \tag{13}$$

The BSADF test improves the detection of bubbles especially when the generating mechanism is not the same for all the sample observations. Under this new identification strategy, the origination and collapse dates of a bubble are defined as follows

$$\widehat{r}_0 = \inf\{r_2 \geq r_0 : \text{BSADF}_{r_2}(r_0) > c_{1-\alpha}\}, \tag{14}$$

$$\widehat{r}_f = \inf\left\{r_2 \geq \widehat{r}_0 + \beta\frac{\log T}{T} : \text{BSADF}_{r_2}(r_0) < c_{1-\alpha}\right\}, \tag{15}$$

where $c_{1-\alpha}$ is the $(1-\alpha)$ percentile of the limit distribution of the sup of an ADF test based on $[Tr_2]$ observations. Under suitable conditions (see footnote 4) the estimates of the bubbles dates prove to be consistent.

Over the last years, this approach has been widely employed to detect and date bubbles, and a significant amount of research has been carried out to either generalize or improve the approach and properties of Phillips's tests. In this regard, it is worth mentioning (Phillips and Shi 2018), who proposed a reverse regression for detecting the bubble's explosion and determining its origination and termination, as well as Phillips and Shi (2020), who suggested incorporating a bootstrap procedure in the GSADF test to address the multiplicity issue in recursive testing and conditional heteroskedasticity in data that can affect the performance of Phillips's test. To overcome this drawback, Harvey *et al.* (2016) developed a wild bootstrap procedure to correct the size of the test when data exhibit deterministic volatility. To cover the case of stochastic volatility, the procedure was extended by Hafner (2020) using spline-GARCH models. Harvey *et al.* (2019) proposed a weighted least squares variant of Phillips's test which proves to be robust to several different patterns of stochastic volatility, while Harvey *et al.* (2020) developed a sign-based variant of the GSADF test. To address the issue of bubble detection in presence of serially correlated innovations, Pedersen and Schütte (2020) proposed sieve bootstrap versions of both the Phillips and Yu (2011) and Phillips and Shi (2020) tests, while Lui *et al.* (2024) worked out a new heteroskedasticity and autocorrelation robust (HAR) test statistic.

2.2. *Bubble detection based on FDR and FNR with calibrated p-values*

In this section it is shown as the performance of the unit root tests, employed sequentially to detect bubbles, can be improved by calibrating p-values. Indeed, simulation experiments prove that the p-values of the unit root tests, when applied recursively, do not enjoy either the uniformity or the super-uniformity property required for a correct application of the multiple tests in real time. To overcome this drawback, the distribution of the p-values is determined from a sub-sample, or training period, ideally coinciding with the initial portion of the sample, unaffected by bubbles explosion. This distribution is then used to calibrate the p-values of the unit root tests when applied to the rest of the sample, the so called monitoring period. The results of the recursive application of the ADF test are then employed to determine FDR and FNR as well the so-called LORD tests. The latter is a class of online tests that can be effectively implemented to detect bubbles in real time.

This approach is similar in spirit to the ones implemented by Astill *et al.* (2017) and Whitehouse *et al.* (2023). Both used a training period: the former to determine the critical values of the Astill *et al.* (2017) test to detect bubbles, the latter to establish the termination date of the bubble. More precisely, to deal with the multiplicity issue in sequential testing and controlling the false positive rate (FPR), Astill *et al.* (2018) suggested an approach based on comparing the results from the sequential application of the AHLST test, proposed by Astill *et al.* (2017), in both a training period (TP) and a monitoring period

† Equation (10) can be also defined as

$$\widehat{r}_f = \inf\left\{r \geq \widehat{r}_0 + \beta\frac{\log T}{T} : \text{ADF}_r \leq c_{1-\alpha}\right\}$$

‡ It is required that

$$\frac{1}{c_{1-\alpha}} + \frac{c_{1-\alpha}}{T^{1-\frac{\eta}{2}}} \to 0 \quad \text{as } T \to \infty$$

where η is the parameter characterizing the specification of δ_τ in (4).

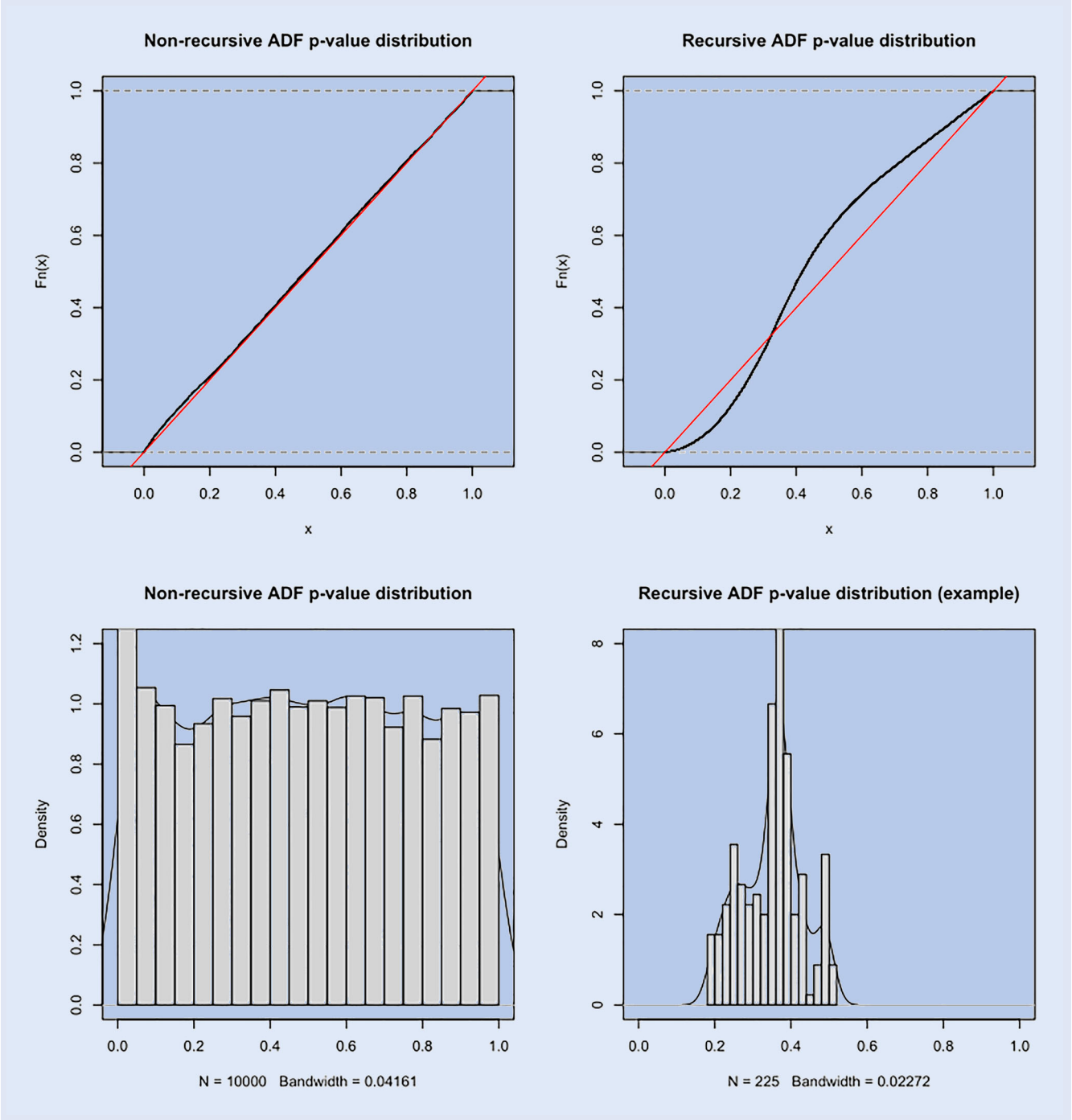

Figure 1. ECDF and histogram/kernel of the p-values distribution in a non-recursive and recursive application of the ADF test (left and right column, respectively). The red line denotes the uniform distribution.

(MP).† The TP must be characterized by the absence of explosive bubbles. The AHLST test is first applied sequentially over rolling sub-samples of a given length in TP to determine the upper tail critical value (cv), for a given significance level. Then, it is applied in real time to the first available window, including sub-samples of the same length, outside the TP, using cv as the critical value. Alternatively, they suggested comparing the number of contiguous rejections in both TP and MP, always using cv as threshold. They proved that both these approaches guarantee FPR to not exceed a specified level.

Whitehouse *et al.* (2023) refined their analysis to enhance real-time detection of the end of explosive behavior and the beginning of a stationary regime. They proposed a statistic based on the different signs of the mean of the first data differences in the explosive and stationary regimes. This statistic is computed in TP over all rolling samples of a given length and the minimum of these training statistics is taken as the critical value. Then, conditional on finding an explosive regime in MP, the statistic is computed over rolling samples and the collapse of the bubble is assumed to start at the first point where it falls below the critical value determined in the TP period.

† The AHLSL statistic rests on the Taylor expansion of the first differences of data and tests for the presence of an upward trend in them.

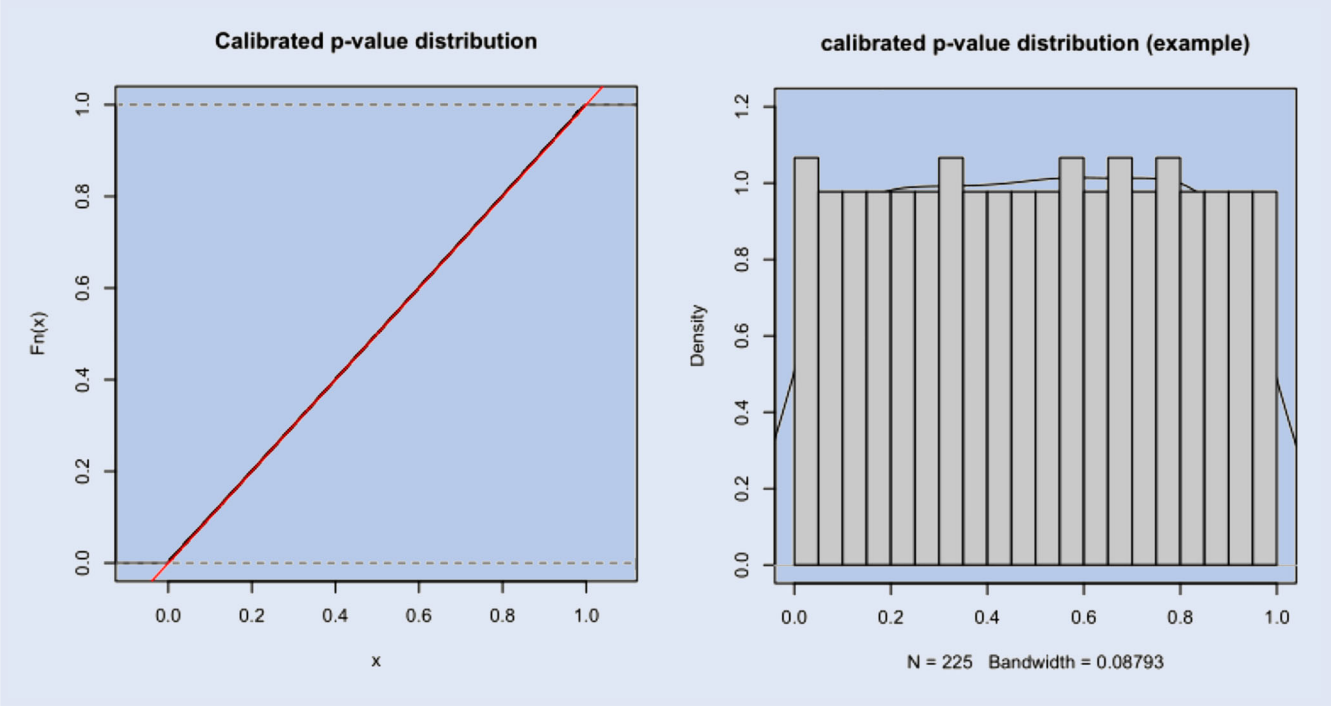

Figure 2. ECDF and histogram/kernel of the p-values distribution after calibration.

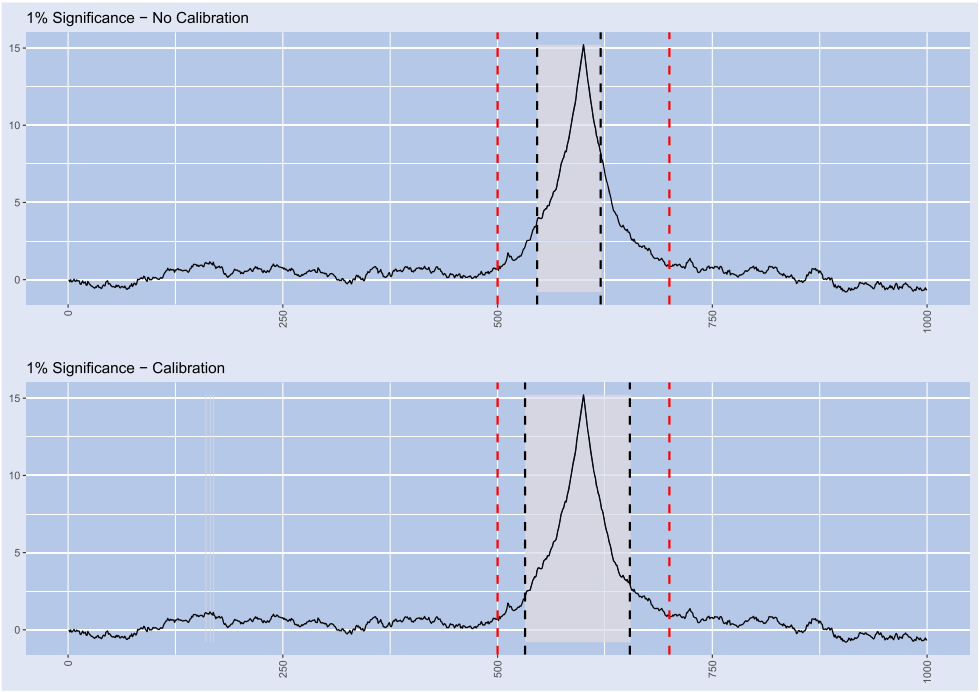

Figure 3. Dating of bubble, before and after p-values calibration, at significance $\alpha = 0.01$. Red vertical lines denote the real originating and ending date, while black vertical lines denote the estimates of these dates.

As explained before, the basic objective of the Philips's test (Phillips and Yu 2011) is to detect bubbles by performing a multiple procedure testing (MPT), which consists in the application of a series of sequential unit root tests. While the non-recursive application of the ADF test engenders p-values with a uniform distribution, this is not true when the test is applied recursively. Indeed, simulation experiments prove that the p-values of the ADF test do not enjoy the so-called super-uniformity property under the null hypothesis

$$P(p_t \leq u \,|H_0) \leq u, \quad u \in [0,1] \tag{16}$$

for any t, when the test is applied recursively. However, this property is necessary for a correct application of multiple tests in real time or online tests. The non-uniformity property of the p-values of the ADF test, when applied recursively, can be proved with the simulation experiment displayed in figure 1.

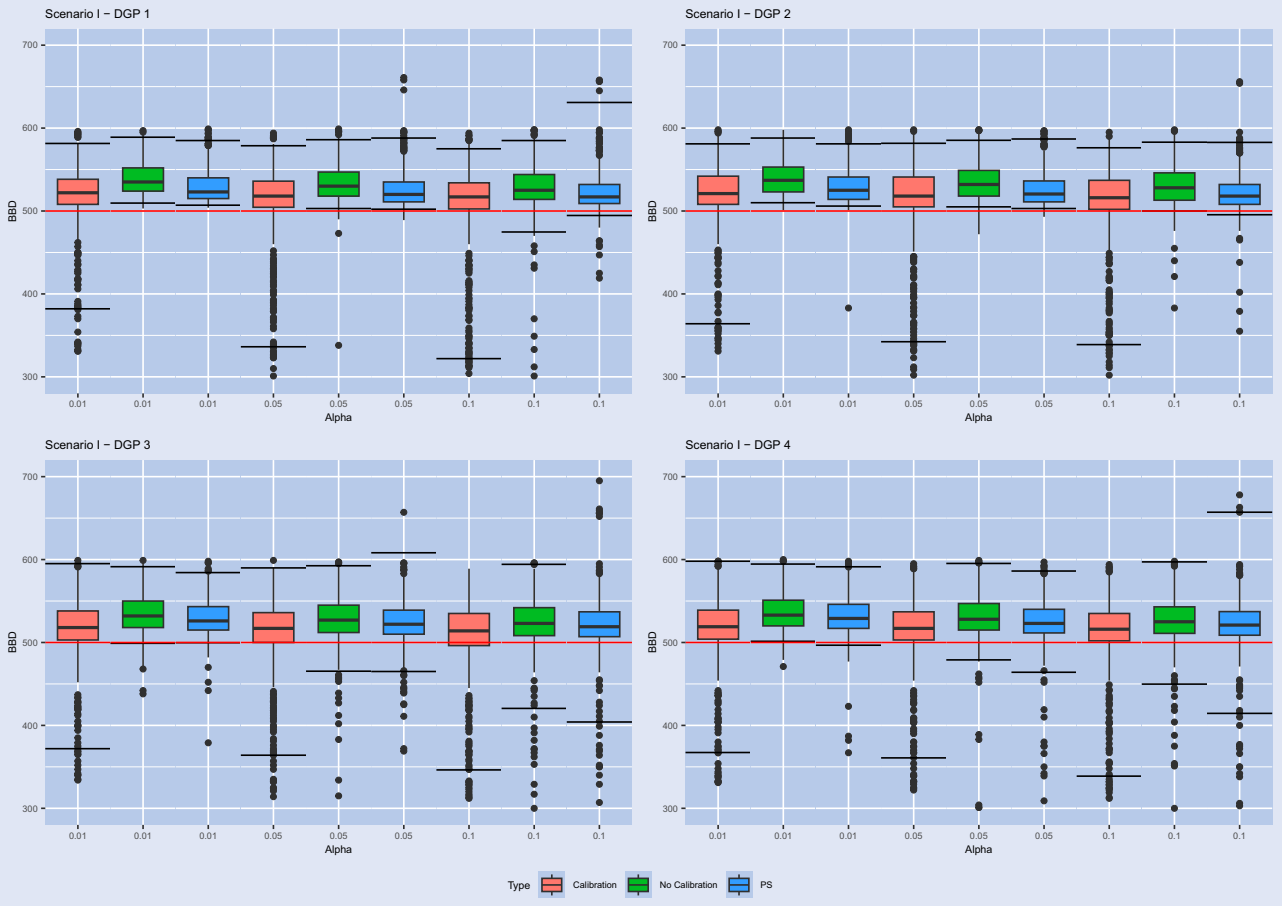

Figure 4. Comparison between BBD estimates obtained with the PH, CPH and PS (bootstrap) tests at different significance levels $\alpha = \{0.01, 0.05, 0.1\}$. Scenario I. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

It is based on $B = 10\ 000$ replications of a unit root process of length $T = 250$. The top panel of the left column in figure 1 shows the p-values empirical cumulative distributions (ECDF), while the bottom panel of the same column shows the kernel density functions/histograms obtained across replications of the ADF test on the entire series (i.e. through a non-recursive application of the test). The top panel of the right column of the same figure shows the ECDF, while the bottom panel of the same column depicts the kernel density of the p-values obtained across replications of the ADF test on subsequent sub-samples (i.e. obtained via a recursive application of the test). In the top panels, the p-value distribution is compared to that of a uniform.

Figure 1 highlights the lack of (super)-uniformity property of the p-values of the ADF test when the latter is applied sequentially. This drawback can be overcome by calibrating the p-values. Calibrated p-values, denoted as p_c, are estimated from the original ones, p, as follows

$$\hat{p}_c = \hat{F}_0(p), \tag{17}$$

where $\hat{F}_0$ is an unbiased estimate of the cumulative distribution function (cdf) F_0 of the p-values under the null (Van der Vaart 2000, chapter 19)

$$\hat{F}_0(\alpha) = \frac{\#(p_j \leq \alpha,\ [Tj] = 1, 2, \ldots, m_0)}{m_0}, \tag{18}$$

with m_0 denoting the number of p-values under H_0. The function $\hat{F}_0$ can be determined from the p-values of the test performed on initial recursive sub-samples whose length, according to Phillips's suggestion, must allow an initial estimation without missing early bubble episodes.

Calibrated p-values enjoy the super-uniformity property. This follows from the inequality

$$P(p_c \leq \alpha | H_0) = P(F_0(p) \leq \alpha | H_0) \leq \alpha \tag{19}$$

which holds, according to the probability integral transformation $p_c = F_0(p)$ (Van der Vaart 2000, chapter 21).

Figure 2, based on the aforementioned simulated data, shows the effects of p-value calibration. The left panel depicts the ECDF of calibrated p-values, while the right panel their kernel density.

Figure 2 proves that calibration allows to recover the super-uniformity property of the p-value CDF, making them usable in the sequential application of the ADF test.

This adjustment has relevant for implications the test's performance, when applied to dating bubbles, as proven by the following case study consisting in the simulation of a bubble according to the following DGP

$$y_t = \mu + \epsilon_t, \quad t = 1, 2, \ldots, T$$

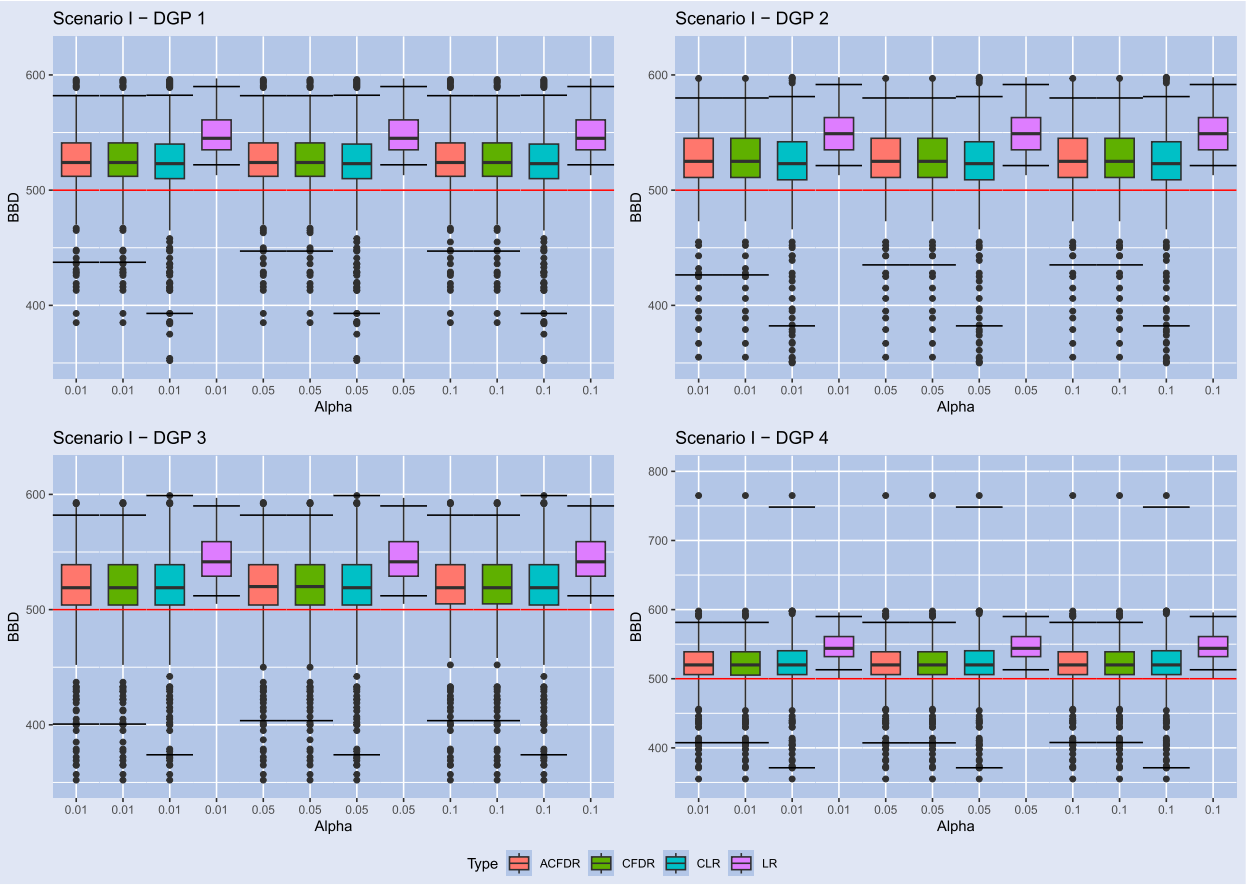

Figure 5. BBD estimates for (C)LR and (AC-C)FDR methods. Along the x-axis, α represents the significance level for the CPH test, and the threshold level for the FDR, (C)LR and (AC-C)FDR methods. Scenario I. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

$$\epsilon_t = \begin{cases} \epsilon_{t-1} + \eta_t & t = 2, \ldots, [\tau_1 T] \\ (1+\delta_1)\epsilon_{t-1} + \eta_t & t = [\tau_1 T]+1, \ldots, [\tau_2 T] \\ (1-\delta_2)\epsilon_{t-1} + \eta_t & t = [\tau_2 T]+1, \ldots, [\tau_3 T] \\ \epsilon_{t-1} + \eta_t & t = [\tau_3 T]+1, \ldots, T. \end{cases} \quad (20)$$

Here, $\delta_1 > 0$, $\delta_2 > 0$, ϵ_t is a possibly conditionally heteroskedastic process for which

$$\eta_t = \sigma_t z_t. \quad (21)$$

In (21), z_t is a zero mean stationary process, possibly heavy tailed like a Student-t, and $\sigma_t = f(\frac{t}{T})$ where f is a non-stochastic and strictly positive function. This allows the process to display possibly multiple one-time volatility shifts (not necessarily located at the same point in the sample), polynomially, possibly piece-wise trending volatility and smooth transition variance breaks, among others. Conventional homoskedasticity arises when $\sigma_t = \sigma$ for all t. The simulated series consists of $T = 1000$ observations, where the bubble regime origination date is set by design at $t = 501$ ($\tau_1 = 0.5$), the shift to the stationary period (i.e. the bubble burst) is set to $t = 601$ ($\tau_2 = 0.6$), while the date when the process returns to a unit root behavior is $t = 700$ ($\tau_3 = 0.7$).

The process has been initialized at $y_1 = c = 1$, with parameters $\mu = 0$, $\delta_1 = \delta_2 = 1/T^{0.5}$ and $\sigma_t = \sigma = 0.5$, and $z_t \sim$ i.i.d.$N(0,1)$.†

Figure 3 depicts the role played by p-value calibration via a simulated series according to the above specification for the process in (20). It shows the bubble together with the real originating and ending date (red vertical lines) and the estimates of these dates (black vertical lines), using original and calibrated p-values of Phillips's test (Phillips and Yu 2011) in the upper and lower panels, respectively. Comparing the two panels, it emerges that the bubble detection window is closer to the nominal one, highlighted in red, when calibrated p-values are employed.

A further improvement in detecting and dating bubbles is possible by controlling the compound error of this multiple testing procedure by implementing FDR and FNR that use calibrated p-values. FDR (FNR)-controlling procedures are designed to control the expected proportion of 'false (non-)discoveries', namely incorrect rejections (acceptances) of the null. The false discovery rate, defined as follows (Efron 2010, Storey 2002)

$$\text{FDR}(\alpha) = P(H_0 | p_r \leq \alpha) \quad (22)$$

† Whitehouse (2019) highlighted the role played by the initial condition on the limiting distributions of unit root tests. She demonstrated that they depend on the initial condition under locally explosive alternatives.

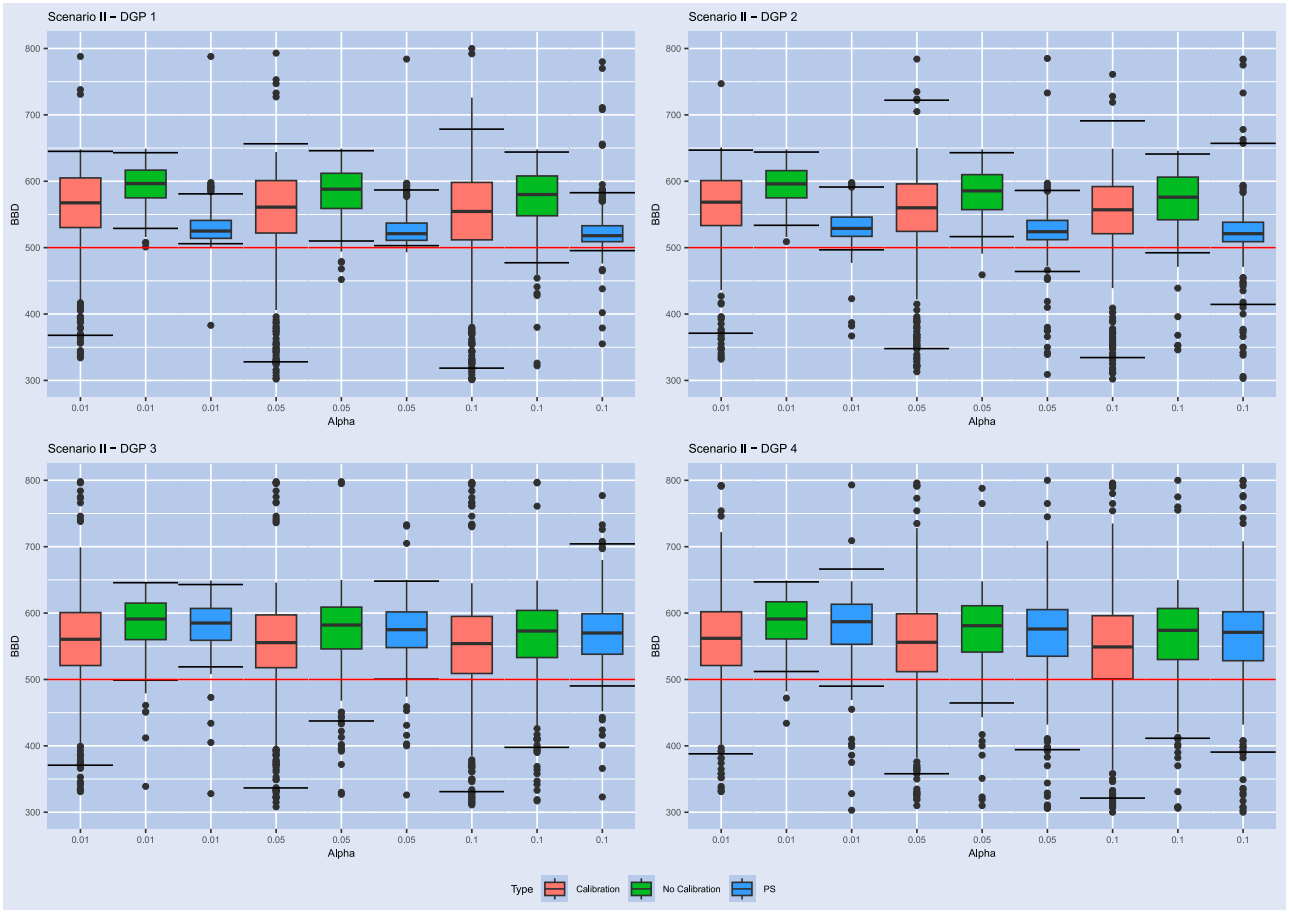

Figure 6. Comparison between BBD estimates obtained with the PH, CPH and PS (bootstrap) tests at different significance levels $\alpha = \{0.01, 0.05, 0.1\}$. Scenario II. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

in the test setting, provides the probability of an actual non-explosive behavior given that an explosive episode has been detected by the test at level α. Accordingly, a low value of $\text{FDR}(\alpha)$ would confirm the result of the test which suggests rejection of the null, while a high value of $\text{FDR}(\alpha)$ would support the null, despite the result of the test. In this case, a deeper analysis is necessary to better identify the origination date of the bubble. Its search can be carried out by looking for the infimum value, $\hat{r}_i^{\text{inf}}$, located on the right side of $\widehat{r}_0$ that, besides leading to the rejection of the null at level α, has the lowest $\text{FDR}(\alpha)$.

Similarly, the bubble ending date can be detected by looking for the infimum value, $\hat{r}_f^{\text{inf}}$, located on the left side of $\widehat{r}_f$ that, besides leading to the acceptance of the null, at level α, has a low false non-discovery rate defined as†

$$\text{FNR}(\alpha) = P(H_1 \,|\, p_r > \alpha) = 1 - P(H_0 \,|\, p_r > \alpha). \tag{23}$$

By making use of Bayes' theorem, (22) and (23) can be expressed as

$$\text{FDR}(\alpha) = \frac{P(p_r \leq \alpha | H_0) P(H_0)}{P(p_r \leq \alpha)} = \frac{F_0(\alpha)\pi_0}{F(\alpha)}, \tag{24}$$

$$\text{FNR}(\alpha) = 1 - \frac{P(p_r > \alpha | H_0) P(H_0)}{P(p_r > \alpha)} = 1 - \frac{(1 - F_0(\alpha))\pi_0}{1 - F(\alpha)}, \tag{25}$$

where $F(\cdot)$ is the CDF of the calibrated p-values, which is a mixture of the CDFs of the p-values under the null and the alternative hypotheses, $F_0(\cdot)$ and $F_1(\cdot)$, with weights $\pi_0 = P(H_0)$ and $\pi_1 = P(H_1)$

$$F(\alpha) = P(p_r \leq \alpha) = P(p_r \leq \alpha \,|\, H_0) P(H_0) + P(p_r \leq \alpha \,|\, H_1) \times P(H_1) = F_0(\alpha)\pi_0 + F_1(\alpha)\pi_1. \tag{26}$$

The distribution $F(\alpha)$ can be unbiasedly estimated by the empirical CDF of the calibrated p-values (Van der Vaart 2000, chapter 19)

$$\hat{F}(x) = \frac{\#[s \leq r : p_s \leq x]}{[Tr]}. \tag{27}$$

If the number of calibrated p-values under the null and alternative hypotheses, m_0 and m_1 respectively, were known, then unbiased estimates of the CDFs F_i, $i = \{0, 1\}$ could be obtained from the empirical CDF of the calibrated p-values as

† Similarly, FDR and FNR can be expressed in terms of the ADF statistic as

$$\text{FDR}(\alpha) = P(H_0 \,|\, \text{ADF}_r > c_{1-\alpha})$$

$$\text{FNR}(\alpha) = P(H_1 \,|\, ADF_r \leq c_{1-\alpha}) = 1 - P(H_0 \,|\, ADF_r \leq c_{1-\alpha})$$

follows

$$\hat{F}_i(\alpha) = \frac{\#(p_{ij} \leq \alpha,\ j = 1, 2, ..m_i)}{m_i}, \tag{28}$$

where p_{ij} denotes the jth p-value under H_i, with $i = \{0, 1\}$.

Otherwise, instead of calibrating the original p-values, a self-calibrated version of the FDR can be directly obtained from the original CDFs F_0 and F of the original p-values

$$\text{FDR}(\gamma_\alpha) = \frac{F_0(\gamma_\alpha)\pi_0}{F(\gamma_\alpha)}, \tag{29}$$

where γ_α is the percentile of these distributions satisfying

$$\begin{aligned}\gamma_\alpha &= F_0^{-1}(\alpha) = \inf\{\gamma \in (0, 1) : F_0(\gamma) \geq \alpha\} \\ &= \sup\{\gamma \in (0, 1) : F_0(\gamma) < \alpha\}.\end{aligned} \tag{30}$$

These percentiles can be consistently estimated as (Van der Vaart 2000, chapter 21)

$$\hat{\gamma}_\alpha = \sup\{\gamma \in (0, 1) : \hat{F}_0(\gamma) < \alpha\}. \tag{31}$$

Both (24) and (29) require an estimate of π_0. Should the numbers m_i, $i = \{0, 1\}$, of original/calibrated p-values under the null and the alternative hypotheses be known, then the probabilities π_i would tally with the proportions† $\hat{\pi}_i = \frac{m_i}{m} = \frac{m_i}{m_0+m_1}$. In general, a conservative estimate of the probability π_0 can be obtained as

$$\hat{\pi}_0(\lambda) = \frac{\#[s \leq r : p_s > \lambda]}{[Tr](1 - \lambda)} = \frac{1 - \hat{F}(\lambda)}{(1 - \hat{F}_0(\lambda))} \tag{32}$$

or, alternatively, by following the approach proposed by Storey (2002):

$$\hat{\pi}_0(\lambda) = \frac{1 - \hat{F}(\lambda)}{1 - \lambda}. \tag{33}$$

In both (32) and (33) $\lambda \in [0, 1]$ is a well-chosen constant, called tuning parameter.

The above estimates can be justified by the following identity

$$\begin{aligned}1 - F(\lambda) &= P(p_s > \lambda) = P(p_s > \lambda \,|\, H_0)\pi_0 \\ &\quad + P(p_s > \lambda \,|\, H_1)\pi_1 \geq (1 - F_0(\lambda))\pi_0,\end{aligned} \tag{34}$$

which entails the following inequality

$$\frac{1 - F(\lambda)}{1 - F_0(\lambda)} \geq \pi_0. \tag{35}$$

As for the optimal value to assign to λ, it can be determined in different ways. λ can be set equal to a low quantile of the distribution of the calibrated p-values, as suggested by Storey (2002), or it can be obtained by minimizing the mean squared error of the bootstrap estimates of the FDR. After computing B bootstrap estimates of the p-values, p^b, $b = 1, 2, \ldots, B$,‡ obtained via a re-sampling procedure which must take into account the data dependence structure, the bootstrap estimates $\widehat{\text{FDR}}_\lambda^b(\alpha)$ are determined for a fixed parameter λ and the mean squared error

$$\lambda = \underset{\lambda\in[0,1]}{\text{argmin}}\, \text{MSE}(\lambda) = \underset{\lambda\in[0,1]}{\text{argmin}} \frac{1}{B}\sum_{b=1}^{B}[\widehat{\text{FDR}}_\lambda^b(\alpha) - \text{FDR}(\alpha)]^2 \tag{36}$$

is computed for any λ which ranges within a given interval, say (0.5, 0.99) with step 0.01. In (36), since $\text{FDR}(\alpha)$ is unknown, it must be replaced with an appropriate estimate.

In the following application we will consider the point estimate of FDR at level α, as in Sala *et al.* (2014), instead of the minimum of all FDR estimates, as in Storey (2002), for every value of λ in the chosen interval. As demonstrated in Sala *et al.* (2014), this approach provides better results especially when p-values are dependent.§¶

With these elements at hand, an estimate of FDR, needed to implement the testing procedure, can be obtained as

$$\widehat{\text{FDR}}(\gamma) = \frac{\hat{\pi}_0\hat{F}_0(\gamma)}{\hat{F}(\gamma)}, \tag{37}$$

where γ is either equal to γ_α or α, depending on whether the use of the self-calibrated version of the FDR is made or disregarded.

It is worth noting that such an estimate is conservative, as it can be proved that $\mathbb{E}(\widehat{\text{FDR}}(\gamma)) \geq \text{FDR}(\gamma)$. The strong control of false discoveries is guaranteed if $\widehat{\text{FDR}}$ is below or equal to a given threshold, say α_1^*.

Similarly, an estimate of the FNR is given by

$$\widehat{\text{FNR}}(\gamma) = 1 - \frac{\hat{\pi}_0(\lambda)(1 - \hat{F}_0(\gamma))}{1 - \hat{F}(\gamma)}, \tag{38}$$

where, as in (36), the tuning parameter λ can be estimated by minimizing the mean square error of the bootstrap estimates

$$\lambda = \underset{\lambda\in[0,1]}{\text{argmin}}\, \text{MSE}(\lambda) = \underset{\lambda\in[0,1]}{\text{argmin}} \frac{1}{B}\sum_{b=1}^{B}[\widehat{\text{FNR}}_\lambda^b(\alpha) - \text{FNR}(\alpha)]^2, \tag{39}$$

with $\text{FNR}(\alpha)$ replaced by an appropriate estimate (see Shao and Tu 2012). As for γ, it can be set either equal to α or to γ_α, as defined in (30). This latter choice provides a self-calibrated estimate of FNR. The strong control of false non-discoveries is guaranteed if $\widehat{\text{FNR}}$ is below or equal to a given threshold, say α_2^*.‖

† Generally, the quantities m_i can be assumed as known only in simulation studies.

‡ The symbol r at the pedix of the p-value has been omitted for the sake of better readability.

§ In the optimization problem (36), Storey considered the MSE of the bootstrap estimates $\widehat{\text{FDR}}_\lambda^b(\alpha)$ with respect to the minimum of all FDR estimates computed for different values of λ ranging within a given interval, and with a given step size.

¶ The paper by Bajgrowicz and Scaillet (2012) provides an interesting study about the properties of the FDR approach under a wide range of p-values dependence structures, such as association, latent factor model, block dependence and alpha mixing. The latter form of dependence views p-values as a spatial process on [0,1] characterized by the property that p-values located in sub-intervals close to zero are independent from those located in sub-intervals close to one.

‖ Given the peculiarity of the multiple testing being studied, the threshold levels for $\widehat{\text{FDR}}$ and $\widehat{\text{FNR}}$ can be different.

Table 1. BBD descriptive statistics for (C)PH, PS, (C)LR and (AC-C)FDR methods. Column α represent significance level for (C)PH and PS tests, and FDR control threshold for (C)LR and (AC-C)FDR. Scenario I.

DGP	Type	α	Detections	% on CPH	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$	DGP	Type	α	Detections	% on CPH	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$
(1)												(3)											
$z_t \sim N(0,1)$	PH	0.01	463	–	539.4	21.3	509.6	524	535	552	589	$z_t \sim N(0,1)$	PH	0.01	461	–	535.4	25	499	518	532	550	589
$\sigma_t = \sigma\ \forall t$		0.05	468	–	534.7	22.7	503.7	518	530	547	586	$\sigma_t = \omega t/T$		0.05	461	–	529.1	28.4	468	512	527	545	586
		0.1	463	–	530.4	26	489	514	526	544	583.9			0.1	459	–	523.8	34.2	438.2	509	524	542	583.6
	CPH	0.01	452	–	520.5	37.3	411	509	522.5	539	579		CPH	0.01	464	–	516.9	42	399.6	504	518.5	538.3	582.9
		0.05	434	–	515.7	42.7	385.3	507	520	536.8	578.2			0.05	462	–	513.3	44.5	382.6	501	517	536.8	580.5
		0.1	409	–	514.7	43.2	383.2	505	518	536	575.8			0.1	453	–	511.9	45.9	379.6	500	516	536	578.7
	PS	0.01	481	–	529.8	20.4	507	515	523	540	585		PS	0.01	465	–	531.2	25.2	499.6	515	526	544	584
		0.05	487	–	527.7	27.9	502.2	511	520	535	587.9			0.05	463	–	525.3	29.9	463.8	510	522	539	582
		0.1	489	–	525.5	33.8	494.4	509	517	533	593			0.1	457	–	523.6	41.5	433.6	507	520	537	589.4
	CFDR	0.01	405	89.6%	526.2	30.6	437.4	512	524	541	582		CFDR	0.01	454	97.8%	517.4	40.9	400.7	504	519	539	582
		0.05	405	93.3%	526.4	30.3	447.1	512	524	541	582			0.05	454	98.3%	518.2	40.2	403.6	504.3	520	539	582
		0.1	405	99.0%	526.5	30.1	447.1	512	524	541	582			0.1	454	100.2%	518.4	39.9	403.6	505	519.5	539	582
	ACFDR	0.01	405	89.6%	526.2	30.6	437.4	512	524	541	582		ACFDR	0.01	454	97.8%	517.5	40.9	400.7	504	519	539	582
		0.05	405	93.3%	526.4	30.3	447.1	512	524	541	582			0.05	454	98.3%	518.2	40.2	403.6	504.3	520	539	582
		0.1	405	99.0%	526.5	30.1	447.1	512	524	541	582			0.1	454	100.2%	518.4	39.9	403.6	505	519.5	539	582
	LR	0.01	444	98.2%	548.7	18.4	522	535	545	561	589.9		LR	0.01	446	96.1%	545.4	21.6	512	529	541.5	559	590
		0.05	444	102.3%	548.7	18.4	522	535	545	561	589.9			0.05	446	96.5%	545.4	21.6	512	529	541.5	559	590
		0.1	444	108.6%	548.7	18.4	522	535	545	561	589.9			0.1	446	98.5%	545.4	21.6	512	529	541.5	559	590
	CLR	0.01	459	101.5%	522	36.2	417.4	510	523	540	582		CLR	0.01	465	100.2%	517.6	41.5	400	504	519	539	582
		0.05	459	105.8%	522	36.2	417.4	510	523	540	582			0.05	465	100.6%	517.6	41.5	400	504	519	539	582
		0.1	459	112.2%	522	36.2	417.4	510	523	540	582			0.1	465	102.6%	517.6	41.5	400	504	519	539	582
(2)												(4)											
$z_t \sim t_v$	PH	0.01	465	–	540.2	20.9	510	523	537	553	588	$z_t \sim t_v$	PH	0.01	454	–	536.9	22.6	501.3	520	533	551	587.4
$\sigma_t = \sigma\ \forall t$		0.05	470	–	535.1	22.3	505	518	532	549	585.3	$\sigma_t = \omega t/T$		0.05	458	–	531.7	26.4	484	515	528.5	547	589
		0.1	470	–	530.8	24.6	500	513	528	546	583			0.1	461	–	526.3	31.9	452.5	511	525	543	587
	CPH	0.01	457	–	520.6	41.4	396.8	508	522	543	580.6		CPH	0.01	465	–	518.8	42.3	404	505	520	540	586.4
		0.05	439	–	517.2	47.3	378.8	506	520	542	581.1			0.05	459	–	516.9	44.4	399.2	504	518	537	582.6
		0.1	420	–	513.4	49.1	371.8	504	517	539.3	575.5			0.1	454	–	515.6	47.5	392.3	503	517	537	583
	PS	0.01	479	–	530.1	24.2	506	514	525	541	581		PS	0.01	464	–	531.7	25.7	496.6	517	529	546	584.4
		0.05	485	–	526.8	23.7	503	511	521	537	584.6			0.05	465	–	526.7	31.4	478.2	512	524	541	582.4
		0.1	489	–	523.9	31.6	495.4	509	518	533	581.4			0.1	466	–	525	39.1	448.3	509	521	539	585.5
	CFDR	0.01	417	91.2%	525.4	31.5	436.4	511	525	545	579.6		CFDR	0.01	450	96.8%	518.9	37.8	407.5	505.3	520	539	580
		0.05	417	95.0%	525.5	31.2	445.4	511	525	545	579.6			0.05	450	98.0%	519	37.9	407.2	506	520	539	580
		0.1	417	99.3%	525.6	31.1	445.8	511	525	545	579.6			0.1	450	99.1%	519.3	37.6	407.7	506	520	539	580
	ACFDR	0.01	417	91.2%	525.4	31.5	436.4	511	525	545	579.6		ACFDR	0.01	450	96.8%	518.9	37.8	407.5	506	520	539	580
		0.05	417	95.0%	525.5	31.2	445.4	511	525	545	579.6			0.05	450	98.0%	519	37.8	407.2	506	520	539	580
		0.1	417	99.3%	525.6	31.1	445.8	511	525	545	579.6			0.1	450	99.1%	519.3	37.6	407.7	506	520	539	580
	LR	0.01	454	99.3%	550.4	19.2	521.3	535	549	563	591.7		LR	0.01	443	95.3%	546.9	20.6	513	532	544	561	590
		0.05	454	103.4%	550.4	19.2	521.3	535	549	563	591.7			0.05	443	96.5%	546.9	20.6	513	532	544	561	590
		0.1	454	108.1%	550.4	19.2	521.3	535	549	563	591.7			0.1	443	97.6%	546.9	20.6	513	532	544	561	590
	CLR	0.01	470	102.8%	521.8	37.8	393.4	509	523	542	580.3		CLR	0.01	463	99.6%	519.6	39.3	404	506	520	540.5	585.9
		0.05	470	107.1%	521.8	37.8	393.4	509	523	542	580.3			0.05	463	100.9%	519.6	39.3	404	506	520	540.5	585.9
		0.1	470	111.9%	521.8	37.8	393.4	509	523	542	580.3			0.1	463	102.0%	519.6	39.3	404	506	520	540.5	585.9

Table 2. BBD descriptive statistics for (C)PH, PS, (C)LR and (AC-C)FDR methods. Column α represent significance level for (C)PH and PS tests, and FDR control threshold for (C)LR and (AC-C)FDR. Scenario II.

DGP	Type	α	Detections	% on CPH	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$	DGP	Type	α	Detections	% on CPH	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$
(1) $z_t \sim N(0,1)$ $\sigma_t = \sigma\ \forall t$	PH	0.01	406	–	593.2	31.3	529	575	596.5	616.8	643	(3) $z_t \sim N(0,1)$ $\sigma_t = \omega t/T$	PH	0.01	403	–	585	39.1	499.1	560	591	615	643
		0.05	427	–	584.5	37.6	510	559	588	612	645			0.05	426	–	574	51.6	443.6	546.3	582	609	644
		0.1	430	–	576.4	43.1	483.6	548.3	580.5	608	643			0.1	427	–	564.9	58.2	411	535	574	604.5	644
	CPH	0.01	428	–	561.3	63.4	390	532	568.5	606	644		CPH	0.01	441	–	556	70.7	387	525	563	602	646
		0.05	406	–	558.6	64.9	382.6	528	565.5	603	642			0.05	433	–	553.8	72.5	383.8	522	559	600	736.6
		0.1	388	–	551.9	69.4	372.4	522	560.5	601.3	639			0.1	423	–	552.6	76.7	385.6	518.5	557	597.5	752.8
	PS	0.01	479	–	530.1	24.2	506	514	525	541	581		PS	0.01	454	–	582.6	34.2	519.7	559	585	607	643
		0.05	485	–	526.8	23.7	503	511	521	537	584.6			0.05	469	–	574.7	40.6	501	548	575	602	644.3
		0.1	489	–	523.9	31.6	495.4	509	518	533	581.4			0.1	477	–	569.8	45.2	491.8	538	570	599	646.1
	CFDR	0.01	382	89.3%	569.5	53.5	430.1	539	574	608.8	644		CFDR	0.01	419	95.0%	557.2	68	389.7	530	564	601.5	645.6
		0.05	382	94.1%	569.8	53.3	430.1	539	574	608.8	644			0.05	419	96.8%	557.5	67.7	395.7	530	564	601.5	645.6
		0.1	382	98.5%	569.8	52.5	430.1	539.3	574	608	643.5			0.1	419	99.1%	557.8	67.5	395.7	531	564	601.5	645.6
	ACFDR	0.01	382	89.3%	569.5	53.6	430.1	539	574	608.8	644		ACFDR	0.01	419	95.0%	557.2	68	389.7	530	564	601.5	645.6
		0.05	382	94.1%	569.8	53.3	430.1	539	574	608.8	644			0.05	419	96.8%	557.5	67.7	395.7	530	564	601.5	645.6
		0.1	382	98.5%	569.8	52.5	430.1	539.3	574	608	643.5			0.1	419	99.1%	557.8	67.5	395.7	531	564	601.5	645.6
	LR	0.01	361	84.3%	610.1	21.7	568	598	611	627	645		LR	0.01	367	83.2%	604.9	26.9	537.5	592	608	622	644.9
		0.05	361	88.9%	610.1	21.7	568	598	611	627	645			0.05	367	84.8%	604.9	26.9	537.5	592	608	622	644.9
		0.1	361	93.0%	610.1	21.7	568	598	611	627	645			0.1	367	86.8%	604.9	26.9	537.5	592	608	622	644.9
	CLR	0.01	443	103.5%	563.1	60.9	389.3	535	570	606.5	644		CLR	0.01	443	100.5%	557.1	69	387.3	526.5	564	602	646
		0.05	443	109.1%	563.1	60.9	389.3	535	570	606.5	644			0.05	443	102.3%	557.1	69	387.3	526.5	564	602	646
		0.1	443	114.2%	563.1	60.9	389.3	535	570	606.5	644			0.1	443	104.7%	557.1	69	387.3	526.5	564	602	646
(2) $z_t \sim t_v$ $\sigma_t = \sigma\ \forall t$	PH	0.01	414	–	594	28.8	533.7	575	596	616	644	(4) $z_t \sim t_v$ $\sigma_t = \omega t/T$	PH	0.01	399	–	587.2	37.9	511.9	561	591	617	645
		0.05	426	–	583.5	34.2	516.5	557.3	585.5	610	642.4			0.05	415	–	574.8	49.6	473.5	542	581	611.5	642.7
		0.1	427	–	572.6	43.7	498.3	542	576	606.5	641			0.1	419	–	566.6	55.5	438.3	532	575	607	643
	CPH	0.01	431	–	563.1	56.1	395.8	537	570	602.5	643		CPH	0.01	439	–	556.1	67.1	401.8	522	562	602.5	647
		0.05	417	–	556.3	65	369	531	563	598	644			0.05	425	–	553.2	69.5	392	515	557	600	645.8
		0.1	392	–	552.9	63.8	368.1	528	559	595.3	640.2			0.1	417	–	549.6	73.7	377.6	510	553	597	716.2
	PS	0.01	464	–	531.7	25.7	496.6	517	529	546	584.4		PS	0.01	418	–	581.7	45	494.4	554	587	613.8	641.6
		0.05	465	–	526.7	31.4	478.2	512	524	541	582.4			0.05	422	–	569.7	51.7	444.2	538	577	606.8	641.5
		0.1	466	–	525	39.1	448.3	509	521	539	585.5			0.1	431	–	566.8	57.6	445.3	532	572	603	674.3
	CFDR	0.01	382	88.6%	571	46.2	453.6	545	575	604	642		CFDR	0.01	413	94.1%	557.4	63.8	404.3	523	563	601	643.7
		0.05	382	91.6%	571.7	46.9	462.8	545.3	575	604	643			0.05	413	97.2%	557.6	63.7	404.3	526	564	601	643.7
		0.1	382	97.4%	571.5	45.8	464.3	546	575	604	642.5			0.1	413	99.0%	557.6	63.1	405.2	526	564	602	643
	ACFDR	0.01	382	88.6%	571	46.2	453.6	545	575	604	642		ACFDR	0.01	413	94.1%	557.4	63.8	404.3	523	563	601	643.7
		0.05	382	91.6%	571.7	46.9	462.8	545.3	575	604	643			0.05	413	97.2%	557.6	63.7	404.3	526	564	601	643.7
		0.1	382	97.4%	571.5	45.8	464.3	546	575	604	642.5			0.1	413	99.0%	557.6	63.1	405.2	526	564	602	643
	LR	0.01	371	86.1%	610.5	19.5	570.3	598	611	624.5	643		LR	0.01	353	80.4%	605.1	26.5	545.8	590	608	624	645.2
		0.05	371	89.0%	610.5	19.5	570.3	598	611	624.5	643			0.05	353	83.1%	605.1	26.5	545.8	590	608	624	645.2
		0.1	371	94.6%	610.5	19.5	570.3	598	611	624.5	643			0.1	353	84.7%	605.1	26.5	545.8	590	608	624	645.2
	CLR	0.01	440	102.1%	564.2	56.2	392.8	538	573	600.5	643		CLR	0.01	435	99.1%	558.1	65.7	403.7	525.5	564	603	645.2
		0.05	440	105.5%	564.2	56.2	392.8	538	573	600.5	643			0.05	435	102.4%	558.1	65.7	403.7	525.5	564	603	645.2
		0.1	440	112.2%	564.2	56.2	392.8	538	573	600.5	643			0.1	435	104.3%	558.1	65.7	403.7	525.5	564	603	645.2

As for the statistical properties of both FDR and FNR, it can be proved that if the p-values are stationary associated with

$$\sum_{i=1}^{m} \mathbb{C}\text{ov}(p_{i/m}, p_1) = o(m), \quad m \to \infty. \tag{40}$$

Then, in light of Yu (1993), stating that $\widehat{F}(\gamma) \xrightarrow{a.s.} F(\gamma)$, the following holds†

$$\widehat{\text{FDR}}(\gamma) = \frac{\hat{F}_0(\gamma)\pi_0}{\widehat{F}(\gamma)} \xrightarrow{a.s.} \frac{F_0(\gamma)\pi_0}{F(\gamma)} = \text{FDR}(\gamma), \tag{41}$$

$$\begin{aligned}\widehat{\text{FNR}}(\gamma) &= 1 - \frac{(1-\hat{F}_0(\gamma))\pi_0}{1-\widehat{F}(\gamma)} \xrightarrow{a.s.} 1 - \frac{(1-F_0(\gamma))\pi_0}{1-F(\gamma)} \\ &= \text{FNR}(\gamma),\end{aligned} \tag{42}$$

thanks to the continuous mapping theorem.

2.3. *Online control of false discoveries with calibrated p-values*

Calibrated p-values, as defined in (17), have been also used in the LORD approach proposed by Javanmard and Montanari (2018) and Fisher (2022). This is an online procedure that can be adopted to control FDR and which finds its root in Benjamini and Hochberg (1995). Unlike standard FDR control techniques, online procedures control FDR by choosing different significance levels α_j for testing sequentially the nulls H_j's. In an online testing procedure, the significance levels are indeed required to be functions of the prior outcomes, that is $\alpha_j = \alpha_j(R_1, \dots R_{j-1})$, where $R_j \in \{0, 1\}$ and $R_j = 1$ is to be interpreted as rejection of the null H_j.

In this regard, this testing procedure belongs to the class of the alpha-investing rules that were first proposed by Foster and Stine (2008). An alpha-investing method allows the significance threshold α_j for upcoming tests to depend on the test statistics observed so far. Such a method starts with an initial significance level, at most equal to α. This amount changes during the testing of different hypotheses according to a function $w(j)$. It increases each time a discovery occurs, it decreases otherwise.

More precisely, LORD is a generalized alpha-investing procedure that, given a sequence of p-values, $p_1, p_2, \dots$, generates a sequence of decisions R_j at level α_j

$$R_j = \begin{cases} 1 & \text{if } p_j \le \alpha_j(R_1, \dots R_{j-1}) = f(w(j)) \\ 0 & \text{otherwise.} \end{cases} \tag{43}$$

After each decision j, the rule updates $w(j)$ which is employed to determine the test levels α_j according to a given scheme. $w(j)$ is decreased by a quantity ϕ_j if H_j is accepted and increased by a quantity equal to $\psi_j - \phi_j$ if it is rejected, where $\phi_j, \psi_j : [0, 1]^{j-1} \to \mathbb{R}_0^+$. Thus, $w(j)$ is updated according to the following rule

$$\begin{aligned} w(j) = w(j-1) &- \phi_j(R_1, \dots R_{j-1}) \\ &+ R_j\psi_j(R_1, \dots R_{j-1}), \quad w(0) = w_0^*, \end{aligned} \tag{44}$$

where w_0^* is the initial condition and ϕ_j and ψ_j are

$$0 \le \psi_j \le \min(\phi_j + b_0, b_0), \quad \phi_j \le w(j-1), \quad \alpha_j = \phi_j, \tag{45}$$

for a given constant $b_0 > 0$.

Several versions of the LORD method are available depending on the choice of α_j. Here, we have considered the following version

$$\alpha_j = \gamma_j w_0 + \left(\sum_{l \in T(j)} \gamma_{j-l} \right) b_0, \tag{46}$$

known as LORD 2, where $T(j)$ is the set of discoveries up to time j and $\boldsymbol{\gamma}$ is a sequence of non-negative and monotone non-increasing numbers such that $\sum_{j=1}^{\infty} \gamma_j = 1$.

The key attribute of LORD to be a monotonic algorithm means that its property of rejecting more p-values early on can only lead to higher testing thresholds later on.

It is proved that LORD ensures online control, namely

$$\sup_n \text{FDR}(n) \le \alpha^*, \tag{47}$$

where $\text{FDR}(n)$ is the FDR at time n, in case of independent p-values and also under a general form of dependence if the test levels α_j are modified according to the rule $\alpha_j = \phi_j = \zeta_j w(j)$, where ζ_j is a sequence such that $\sum_{i=1}^{\infty} \zeta_i(1 + \log(j)) \le \alpha/b_0$.‡

The property (47) holds under the assumption that p-values under the null follow an uniform distribution, hence it is reasonable to implement the LORD procedure by using calibrated p-values.

In the next sections online tests are applied to detect bubbles in both simulated and real data and their results compared to those of Phillips's tests (Phillips and Yu 2011), implemented with either calibrated or non-calibrated p-values, as well as with the outcomes of the Phillips and Shi (2020) test.

3. Simulation studies

In this section, the two approaches, discussed in the previous section have been implemented in a simulation study to investigate their performance in bubble detection. More precisely,

† Should π_0 be replaced by a suitable estimate $\widehat{\pi}_0$ then, upon noting that $\widehat{\pi}_0 \xrightarrow{a.s.} \widetilde{\pi}_0 > \pi_0$, one can easily draw the conclusion that FDR and FNR would asymptotically converge to a conservative estimate for the former and to a liberal estimate the latter.

‡ Recently, Fisher (2022) studied the LORD under the assumption of positive, local dependence (PLD) for the p-values. He proved that LORD controls FDR if the null p-values satisfy the assumption of conditional super-uniformity, meaning that they either follow a uniform distribution conditional on the information used to define their rejection threshold (see Zrnic *et al.* 2021), or the null p-values follow the conventional assumption of positive regression dependence on a subset (see Benjamini and Yekutieli 2001), or users select significance thresholds that are monotonically non-increasing in the p-values.

Table 3. BED descriptive statistics for (C)PH, PS, CLR and ACFNR methods. Column α represents the significance level for the (C)PH and PS tests and the FNR control threshold level for ACFNR and CLR, by assuming for the latter the ending of the null hypothesis rejection streak as BED. Scenario I.

DGP	Type	α	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$	DGP	Type	α	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$
(1) $z_t \sim N(0,1)$ $\sigma_t = \sigma\ \forall t$	PH	0.01	622.2	4.7	607	622	624	625	626	(3) $z_t \sim N(0,1)$ $\sigma_t = \omega t/T$	PH	0.01	622	5.1	605	621	624	625	626
		0.05	626.8	5.7	609	626	629	630	631			0.05	626.2	12.5	608	626	629	630	632
		0.1	630.4	6.5	609	630	633	634	635			0.1	629	18	608.5	629	633	634	636
	CPH	0.01	638.2	56.1	497.1	630	636	644.3	687.1		CPH	0.01	635.8	24.4	606.6	631	638	645	657
		0.05	639.4	50.6	492.2	631	639	647	699.4			0.05	637.7	28.3	607.5	633	640	647	660.5
		0.1	645.4	55.7	610.4	633	640	650	818.2			0.1	640.9	27.9	614	634	642	649	665.1
	PS	0.01	619.4	4.5	605	619	621	622	623		PS	0.01	619.5	14.1	607	619	621	622	625
		0.05	624.5	13.9	607.2	623	625	626	629			0.05	623.3	19	608	623	626	627	630
		0.1	629.3	18.1	609	627	629	630	633			0.1	628.6	33.1	609.4	626	629	631	634
	ACFNR	0.01	644.6	44.3	614.1	633	639	647	781.1		ACFNR	0.01	639	23.9	613.3	633	640	647.8	661.7
		0.05	639	31.8	614	631	637	645	672.8			0.05	636.8	23.3	607.3	632	638	645	659
		0.1	636.2	19	612	629	635	642	662			0.1	635	25.2	607	631	637	644	656
	CLR	0.01	640.2	60.8	604.5	629	635	643	735.6		CLR	0.01	635.4	30.1	606.6	630	637	644	656.4
		0.05	640.2	60.8	604.5	629	635	643	735.6			0.05	635.4	30.1	606.6	630	637	644	656.4
		0.1	640.2	60.8	604.5	629	635	643	735.6			0.1	635.4	30.1	606.6	630	637	644	656.4
(2) $z_t \sim t_v$ $\sigma_t = \sigma\ \forall t$	PH	0.01	622.2	4.3	609	621	624	625	626	(4) $z_t \sim t_v$ $\sigma_t = \omega t/T$	PH	0.01	622.2	4.8	607	621	624	625	626.7
		0.05	626.8	5.2	610.7	626	629	630	631			0.05	626.3	12	606.4	626	629	630	632
		0.1	630.5	5.7	613.7	630	633	634	635			0.1	628.7	21.6	608	629	633	634	636
	CPH	0.01	639.6	59	521.6	630	636	645	729.2		CPH	0.01	636.6	32.3	608.2	631	639	646	660.8
		0.05	639.5	63.8	432.9	631	638	648	788.5			0.05	639.4	35.7	610.9	633	641	649	670.6
		0.1	638.5	71.8	425.5	632	640	649	802.8			0.1	643.7	38.5	613	635	643	651	683
	PS	0.01	619.5	14.6	608	619	621	622	623.1		PS	0.01	618.3	20.4	605	619	621	623	625
		0.05	623.6	11.5	605.2	623	625	626	628			0.05	622.8	24.9	607	623	626	627	630
		0.1	626.7	24.9	608	626	629	630	632			0.1	628.2	26.9	608.6	626	629	631	634.4
	ACFNR	0.01	640.2	58.9	451.6	631	637	646	782.6		ACFNR	0.01	640.6	32.5	613	633	641	649	666.8
		0.05	634.6	48.6	451.4	630	636	644	693.5			0.05	638.2	26.8	612.3	632	639	647	662
		0.1	636.3	47.3	609.4	628.3	634.5	642.8	674			0.1	636.6	25.5	611.3	631	638	645	659.8
	CLR	0.01	644.4	71.5	603	629	635	645	951.6		CLR	0.01	635.6	26.9	606.6	631	637	645	659.5
		0.05	644.4	71.5	603	629	635	645	951.6			0.05	635.6	26.9	606.6	631	637	645	659.5
		0.1	644.4	71.5	603	629	635	645	951.6			0.1	635.6	26.9	606.6	631	637	645	659.5

Table 4. BED descriptive statistics for (C)PH, PS, CLR and ACFNR methods. Column α represents the significance level for the (C)PH and PS tests and the FNR control threshold level for ACFNR and CLR, by assuming for the latter the ending of the null hypothesis rejection streak as BED. Scenario II.

DGP	Type	α	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$	DGP	Type	α	Mean	SD	$q_{0.025}$	$q_{0.25}$	$q_{0.5}$	$q_{0.75}$	$q_{0.975}$
(1) $z_t \sim N(0,1)$ $\sigma_t = \sigma\ \forall t$	PH	0.01	667.6	6.1	653	664	669	673	675	(3) $z_t \sim N(0,1)$ $\sigma_t = \omega t/T$	PH	0.01	667.1	14	652.1	663	669	673	676
		0.05	671.5	7.6	652.7	667	674	677.5	680			0.05	668.4	34	647.8	666.3	674	678	681
		0.1	674.9	8.4	654	671	677	681	685			0.1	669.7	43.6	557.3	669	678	682	686
	CPH	0.01	673.3	66.5	432.4	670	680	688	776.6		CPH	0.01	680.5	62.6	477	671	682	691	836
		0.05	680.1	75.1	448.6	672	682	691	874.6			0.05	682.6	60.9	463.6	673	684	692	838.6
		0.1	681.3	80.2	447.4	673.8	683	693	894			0.1	689.1	71.3	465.3	675	685	694	909.4
	PS	0.01	619.5	14.6	608	619	621	622	623.1		PS	0.01	664.4	16.9	652	662.3	667	670	672
		0.05	623.6	11.5	605.2	623	625	626	628			0.05	665.3	28.3	651	665	670	673	675
		0.1	626.7	24.9	608	626	629	630	632			0.1	668.1	29.7	652	667	673	675	679
	ACFNR	0.01	683.6	67.9	480.5	673	682	691	885.5		ACFNR	0.01	683.8	61.3	509.3	673	683	692	845.4
		0.05	677.5	59.1	484.5	672	680	688	790.5			0.05	680.9	54.9	525.4	673	682	691	804.1
		0.1	673.1	47.1	482.3	670	679	686	705.2			0.1	679.3	54.6	502.7	672	681	690	800.4
	CLR	0.01	683.4	77.7	465.6	671	679	687	997.1		CLR	0.01	678.7	61	477.5	670	680	690.5	834.6
		0.05	683.4	77.7	465.6	671	679	687	997.1			0.05	678.7	61	477.5	670	680	690.5	834.6
		0.1	683.4	77.7	465.6	671	679	687	997.1			0.1	678.7	61	477.5	670	680	690.5	834.6
(2) $z_t \sim t_v$ $\sigma_t = \sigma\ \forall t$	PH	0.01	667.8	6	652	665	670	673	674	(4) $z_t \sim t_v$ $\sigma_t = \omega t/T$	PH	0.01	666.6	13.3	652	662	669	672	675
		0.05	672.2	7.1	654	669	675	677	680			0.05	667.7	34.5	647	666	674	677.5	681
		0.1	673.7	24.4	653.7	672	679	682	684			0.1	671.5	36.1	648.9	669	677	681	686
	CPH	0.01	671.7	62.9	451.3	673	680	687	725.3		CPH	0.01	675.5	61.3	469.8	670	680	689	783.8
		0.05	675.5	83.3	397.6	674	682	690	885			0.05	682.5	59.9	521	672	683	692	847.6
		0.1	682.2	84.1	406.4	676	683.5	692	905.5			0.1	686	74.5	475	674	684	694	896.4
	PS	0.01	618.3	20.4	605	619	621	623	625		PS	0.01	662.1	31.8	646	661	667	670	674
		0.05	622.8	24.9	607	623	626	627	630			0.05	661.7	45.1	470.2	665	671	673	678
		0.1	628.2	26.9	608.6	626	629	631	634.4			0.1	666.1	47.9	480	667	673	676	695.5
	ACFNR	0.01	681.4	65	453.6	675	682	689	832.7		ACFNR	0.01	683.4	62.2	525	673	683	691	881.8
		0.05	676.9	60.3	443.8	674	681	687.5	732.3			0.05	679.9	56.7	526	672	681	690	831.9
		0.1	675	43	646.5	673	679	686	697.5			0.1	677.7	55.2	528.4	671	680	689	747
	CLR	0.01	680.1	68.6	486	673	680	686	913.3		CLR	0.01	676.4	58.9	517.9	670	680	689	785.3
		0.05	680.1	68.6	486	673	680	686	913.3			0.05	676.4	58.9	517.9	670	680	689	785.3
		0.1	680.1	68.6	486	673	680	686	913.3			0.1	676.4	58.9	517.9	670	680	689	785.3

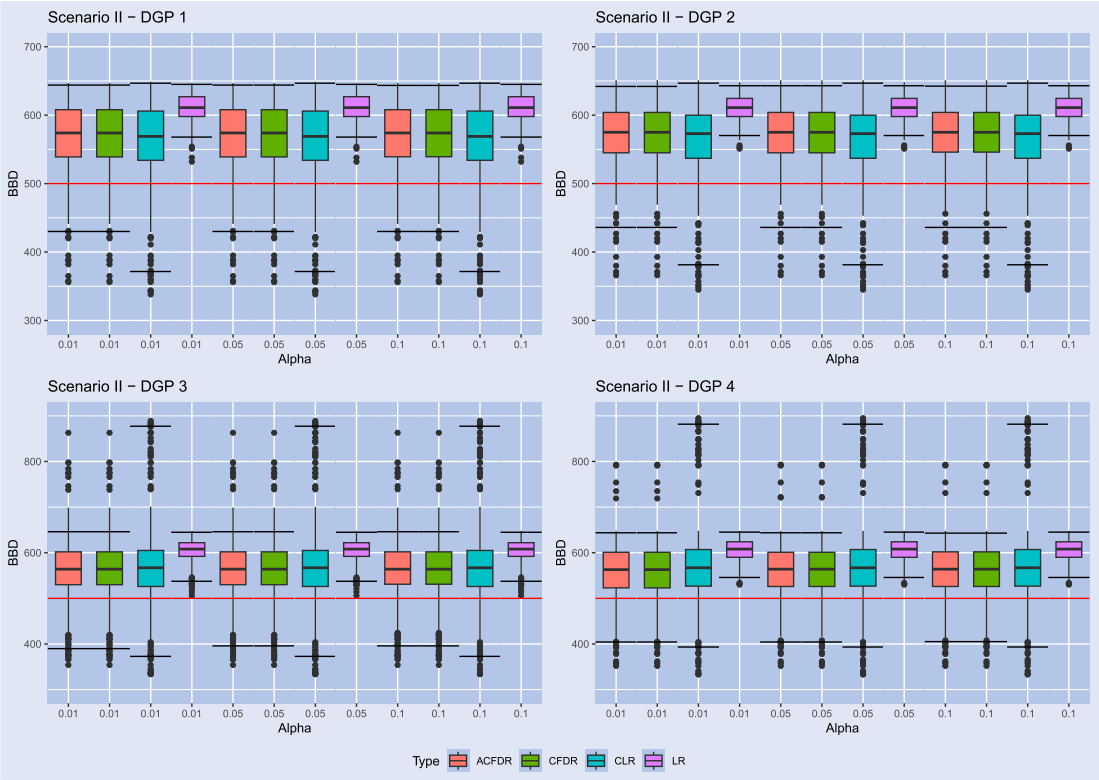

Figure 7. BBD estimates obtained with the (C)LR and (AC-C)FDR methods under Scenario II. Along the x-axis, α represents the significance level for the CPH test, and the threshold level for (C)LR and (AC-C)FDR procedures. Scenario II. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

the tests used in this simulation study are: (1) the original (Phillips and Yu 2011) test [PH]; (2) the PH test, based on calibrated p-values, as defined in (17) [CPH]; (3) the Phillips and Shi (2020) test [PS];† (4) the FDR-based procedures, based on calibrated p-values, employing the Storey method (based on the MSE of the bootstrap p-values) for detecting the tuning parameter [CFDR]; (5) the self-calibrated version of the FDR/FNR-based procedure [ACFDR and ACFNR respectively]; (6) the online LORD procedure [LR]; and (7) the online LORD procedure based on calibrated p-values [CLR].‡

The bubble has been simulated according to the following DGPs, initialized at $y_1 = 1$

$$y_t = \theta_t y_{t-1} + \sigma_t z_t \quad t = 2, \ldots, T$$

and two possible scenarios have been considered for the parameter θ_t:

(I) a 'fast bubble' scenario engendering a sharp bubble

$$\theta_t = \begin{cases} 1 + 1/T^{0.5} & \text{if } t = T/2 + 1, \ldots, T/2 + 100 \\ 1 - 1/T^{0.5} & \text{if } t = T/2 + 101, \ldots, T/2 + 200 \\ 1 & \text{otherwise;} \end{cases}$$

(II) a 'slow bubble' scenario engendering a less marked bubble, which is more difficult to identify

$$\theta_t = \begin{cases} 1 + 1/T^{0.7-0.05(j-1)} & \text{if } t = T/2 + t_j \\ & \quad +1, \ldots, T/2 + t_{j+1} \\ 1 - 1/T^{0.5} & \text{if } t = T/2 + 151, \ldots, \\ & \quad T/2 + 200 \\ 1 & \text{otherwise,} \end{cases}$$

where $j = \{1, 2, 3, 4\}$ and $t_j = \{0, 20, 75, 100, 150\}$.

In both scenarios, four DGPs characterized by the following specifications for the innovation terms are accounted for:

(i) DGP with homoskedastic and normally distributed error: $z_t \sim N(0, 1)$, $\sigma_t = \sigma = 0.5$;
(ii) DGP with homoskedastic and heavy-tailed (Student-t) distributed errors: $z_t \sim t(\nu)$, $\nu = 8$, $\sigma_t = \sigma = 0.5$;
(iii) DGP with heteroskedastic and normally distributed errors: $z_t \sim N(0, 1)$, where $\sigma_t = 0.5\, t/T$;

† The package `psymonitor` was used for this purpose. However, since the procedure only provides the critical values of the bootstrap distribution of the BSADF statistic, p-value calibration is not applicable in this context.
‡ In the paper we have considered the Phillips and Yu (2011) instead of Phillips *et al.* (2015), because our aim is to apply the ADF test sequentially, as the observations arrive, as it happens when the test is applied in real time. Thanks to the way of working of the Phillips and Yu (2011) test, its outcomes are comparable with the online tests considered in the previous section.

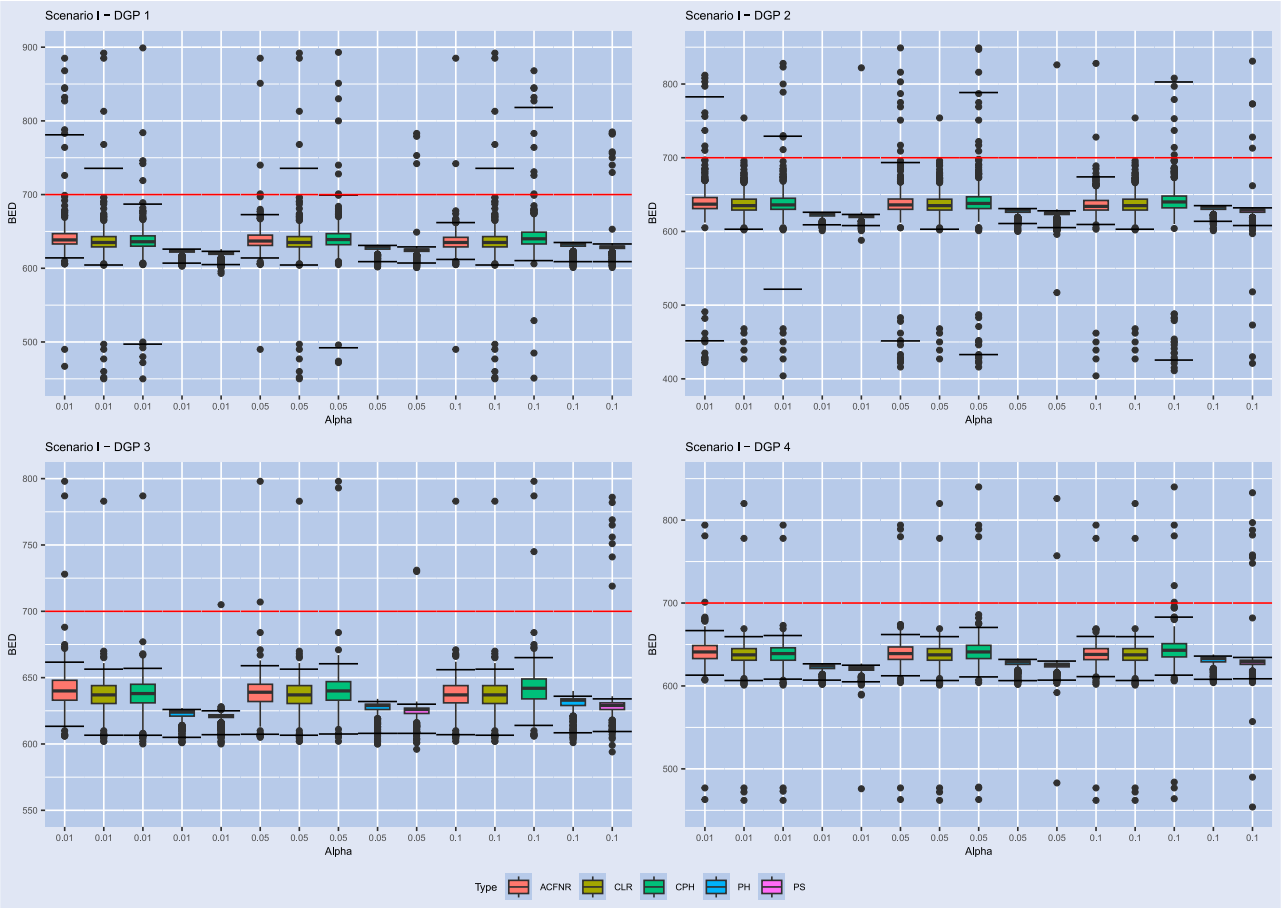

Figure 8. BED comparison between (C)LR and ACFNR methods. Along the x-axis, α represents the FNR control threshold level for ACFNR and CLR, by assuming for the latter the ending of the null hypothesis rejection streak as BED. Scenario I. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

(iv) DGP with heteroskedastic and heavy-tailed (Student-t) distributed errors: $z_t \sim t(\nu)$, $\nu = 8$, variance-scaled to get $\sigma_t = 0.5\, t/T$.

The ability of all the mentioned methods to capture the origination and ending dates of the bubble, set by design at $t = T/2 = 500$ and $t = 700$ for both scenarios, has been measured using both the average and median distances between the beginning of the longest streak of rejections of the null hypothesis and the true start date of the bubble. The number of replications in each scenario was set at $B = 500$ in both simulation studies. Finally, since not all simulated series may show a bubble episode due to randomization, replications for which bubble origination is not detected, or it is signaled at $t < 350$ or $t > 800$ were discarded in the results analysis (as clearly being non-indicative of a bubble episode).

As for the fast bubble scenario, the results pertaining to the detection of the bubble burst obtained from the PH, the CPH, and the PS test have been compared, to appreciate the role played by the p-values calibration. Figure 4 displays the box-plots for the estimated bubble beginning date (BBD hereafter) provided by each test at different significance levels: $\alpha = \{0.01, 0.05, 0.1\}$. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across simulation experiment replications (95% confidence bands).

Looking at the boxplots, it emerges that calibration helps in detecting the bubble burst, as it reduces the discrepancy between the true bubble origination date and its estimated value via the test, by anticipating the date of the first rejection. The PS and PH tests give mixed results, depending on the DGP, since they are both sensitive to the variance structure of the erratic terms. The 95% confidence interval includes the real bubble date of inception across all methods only for $\alpha = 0.1$.

It is worth noting that the crude application of CPH has the side effect of engendering several early rejections or false, premature discoveries.

In order to discard premature discoveries, methods based on either the estimation or the control of the FDR are needed. Therefore, such approaches have been implemented to the replications for which the PH test detects bubbles, with the aim to investigate if these methods confirm the outcomes of the mentioned test.

Figure 5 displays the boxplots for the BBD estimates obtained from the CPH test at $\alpha = \{0.01, 0.05, 0.1\}$, along with the LR, CLR, CFDR and ACFDR at FDR threshold levels $\alpha^* = \{0.01, 0.05, 0.1\}$ (see (47)).

As explained in the previous section, the significance α level of both LR and CLR varies during their sequential application. To guarantee that CFDR or ACFDR satisfy the threshold level α^* (see (47)), the significance level for

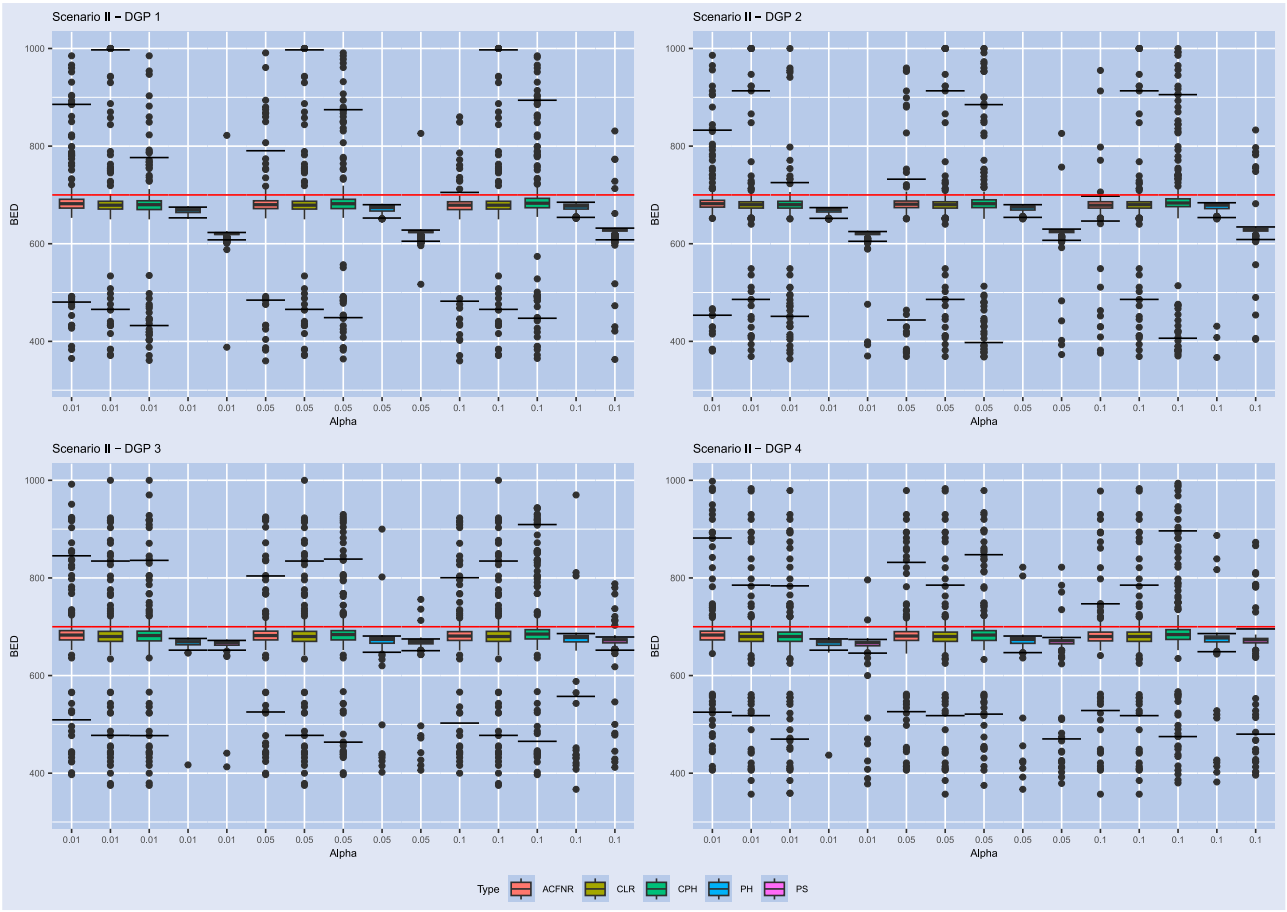

Figure 9. BED comparison between (C)LR and ACFNR. Along the x-axis, α represents the FNR control threshold level for ACFNR and CLR, by assuming for the latter the ending of the null hypothesis rejection streak as BED. Scenario II. The horizontal lines of each box-plots represent the 2.5% and the 97.5% quantiles of the dates detected across all replications in the simulation experiments (95% confidence bands).

the detection of bubble episodes is adjusted to meet this condition.† In this case, a new level α, assuring the respect of the given threshold, is determined and used to detect BBD.

Figure 5 shows that, unlike the CPH test, the outcomes of the mentioned procedures avoid several premature discoveries for the BBD, with the consequence that the lower confidence interval extremes of the BBD estimates increase. For instance, in DGP 1, it is equal to $q^{\text{CPH}}_{0.025} = 383$ for the CPH test at $\alpha = 0.1$, and it shifts to $q^{\text{CFDR}}_{0.025} = 447$ for the CDFR approach. Similarly, the confidence interval extremes of the CLR method shift forward: the 95% bands shift from $q^{\text{CPH}}_{0.025} = 383$ to $q^{\text{CLR}}_{0.025} = 417$ and $q^{\text{CPH}}_{0.975} = 576$ to $q^{\text{CLR}}_{0.975} = 582$, for the threshold level $\alpha^* = 0.1$.

Thus, the conclusion is that calibrating shifts BBD backward, while FDR corrects and moves it forward.

Notably, under this scenario, where the explosive event is considerably sharp and immediate, the CFDR, ACFDR and CLR methods do not provide significantly different results, a result that reasonably stems from the particular characteristics of the data generating process.

Table 1 summarizes the main descriptive statistics for the detection of the BBD together with the percentages of retained detections with all the considered approaches. It shows that in nearly all the replications for which the CPH test detects a bubble, the same occurs for the CFDR, ACFDR, LR, and CLR approaches. Nonetheless, the LR and CLR methods identify a greater number of explosive events than CFDR and ACFDR. The innovation distribution has a less marked effect on the detection of the BBD with respect to the presence of heteroskedasticity. In particular, the latter engenders variability across replications which in turn broaden the confidence intervals.

It is worth noting that the LR procedure is the only FDR-based method which fails to capture the BBD, which means that not calibrating p-values before employing LORD leads to a poor performance of this procedure.

Let us now consider the slow bubble scenario or Scenario II. Figure 6 reports the estimates of the BBD when the PH, CPH and PS tests are employed at different significance levels $\alpha = \{0.01, 0.05, 0.1\}$. Here, the boxplots show the crucial role played by calibration for an effective detection of the

† In practice, a grid of the significance level $0 < \alpha \leq \alpha^*$ is considered, and the new significance threshold $\hat{\alpha}$ is chosen by taking

$$\hat{\alpha} = \min_{\alpha}(\text{FDR}(\alpha) - \alpha^*)^2 = \min_{\alpha}\left(\frac{F_0(\alpha)\pi_0}{F(\alpha)} - \alpha^*\right)^2$$

$$= \min_{\alpha}\left(\frac{\alpha\pi_0}{F(\alpha)} - \alpha^*\right)^2$$

where π_0 is obtained via the Storey method.

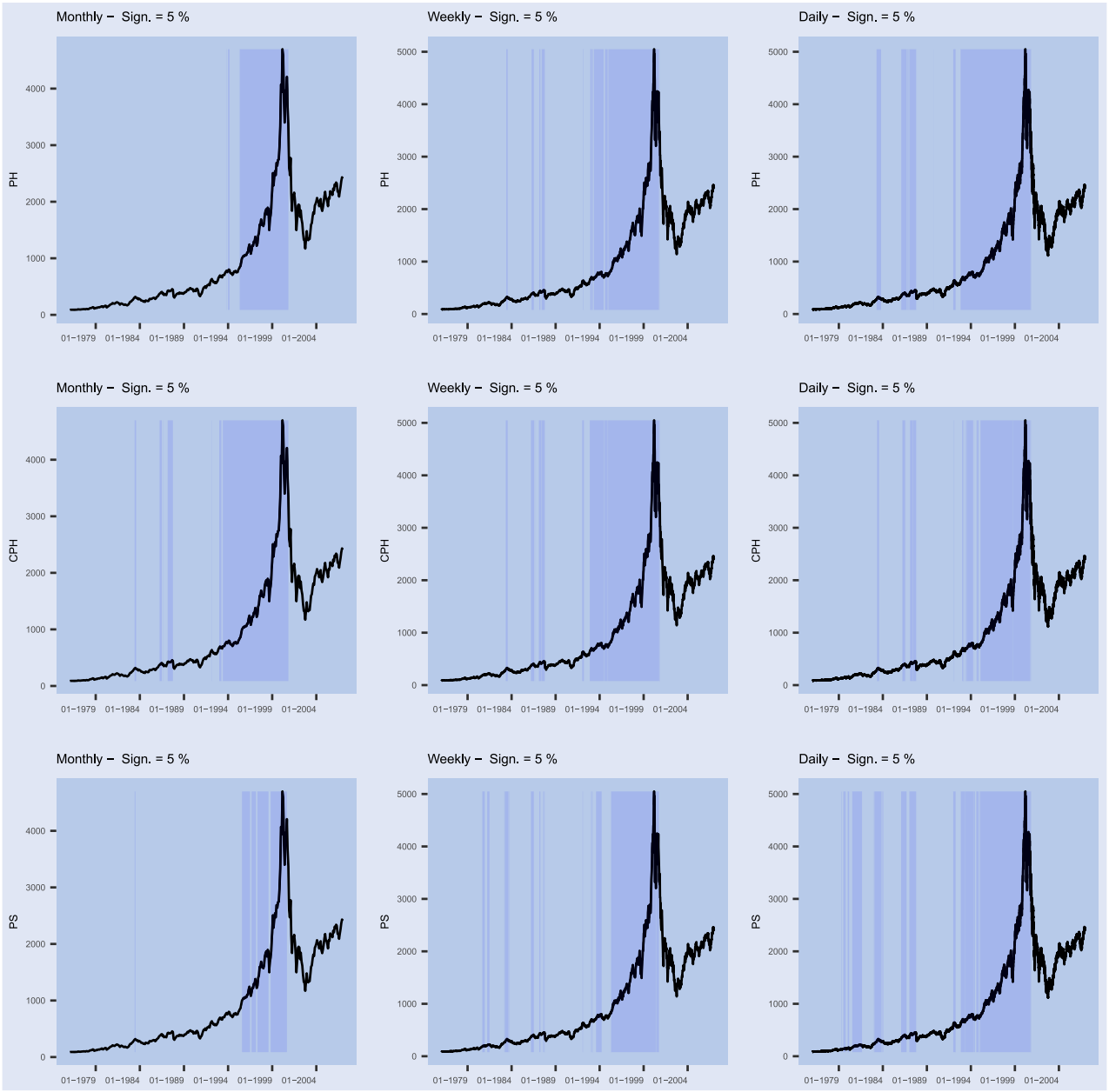

Figure 10. Detection of BBD: PH, CPH and PS at significance $\alpha = 0.05$. Shadowed areas denote sample periods when the null is rejected.

BBD. In fact, approaches based on non-calibrated p-values may not detect the bubble's origination date, especially in DGP 1 and for lower levels of α. Differences among methods are less prominent in the presence of heteroskedasticity or heavy tailed distribution (Student-t) for the innovation term. The BBD outcomes of PS exhibit lower variability across replications in DGPs 2 and 4. Thus, PS better detects the BBD under these DGP's, compared to DGPs 1 and 3. CPH has a stable outcome regardless of the level α and error term structure.

Figure 7 displays the boxplots for the BBD estimates obtained from LR, CLR, CFDR and ACFDR procedures, at FDR threshold levels $\alpha^* = \{0.01, 0.05, 0.1\}$ (see equation (47)) under scenario II.

Similarly to Scenario I, when implementing CFDR and ACFDR, if the threshold is not met, the significance level is adjusted so as to reach it. Then the new level α, assuring the respect of the threshold, is determined and used to detect the BBD.

It can be observed that FDR-based procedures efficiently discard false discoveries. Note that the CLR's boxplot shifted above the others, since the detection of BBD with this method lags behind the others, especially for DGPs 3 and 4. Nevertheless, it still remain within the 95% confidence intervals. The CFDR and ACFDR approaches are more accurate in detecting BBD than CLR, and signal the bubble burst with greater accuracy. As table 2 however, shows that they do not retain as many replications as the CLR test. Introducing heteroskedasticity makes the simulation results more variable, (AC-C)FDR detect more bubble events in DGPs 2 and 4, while the number of detections vary to a lesser extent in DGPs when applying CLR.

Just like in Scenario I, the LR approach is the only method which fails to capture the bubble's origination date.

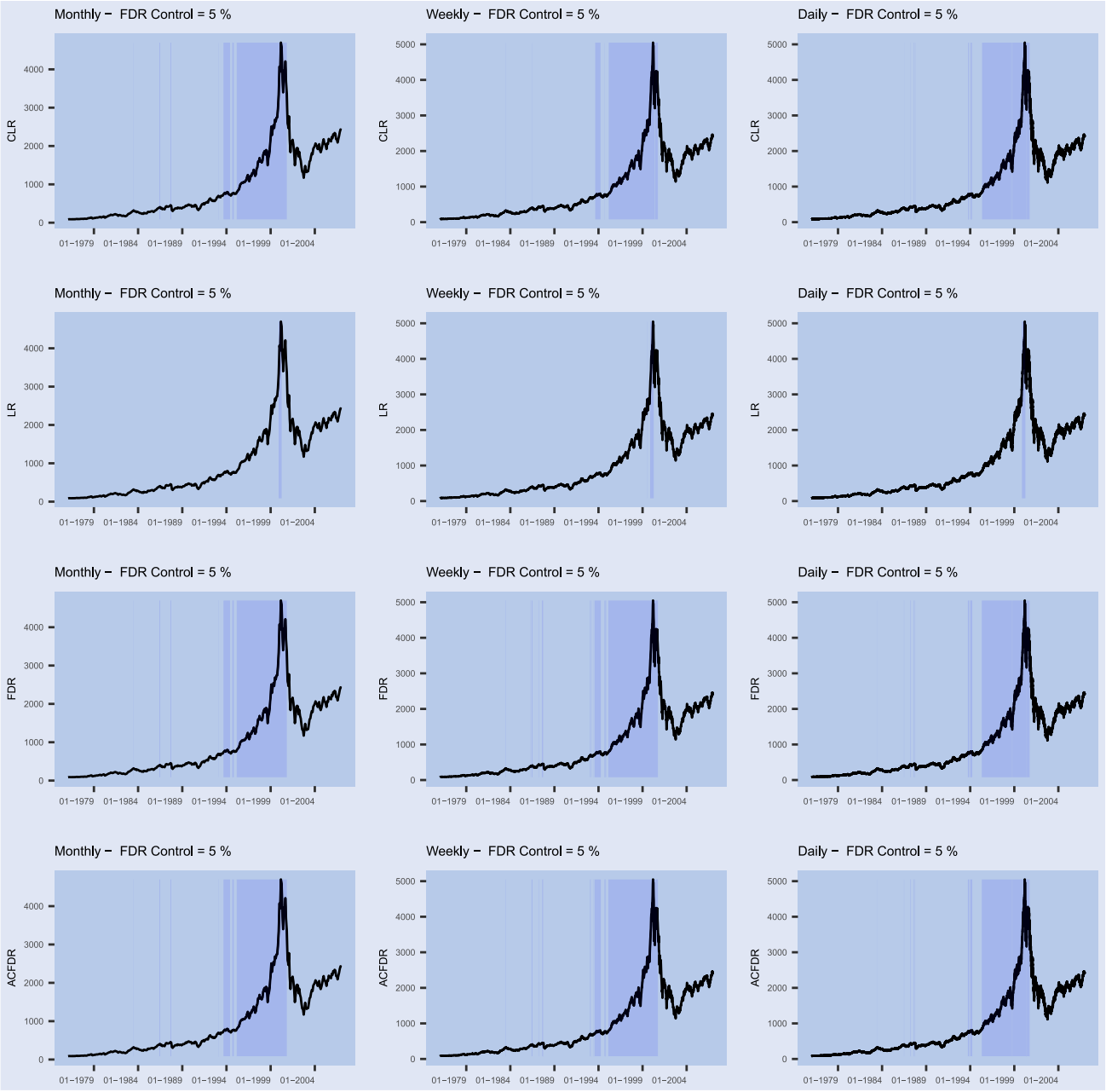

Figure 11. Detection of BBD: (C-AC)FDR and (C)LR at FDR control $\alpha = 0.05$. Shadowed areas denote sample periods when the null is rejected.

Table 2 also shows that the LORD method with calibrated p-values manages to retain more bubble episodes than those detected by the CPH test, especially in the slow bubble scenario and in the homoscedastic error structure. As a last noteworthy observation, in terms of computational burden, the ACFDR approach proves to be less demanding than the CFDR approach. In this regard, since their results are closely comparable in both scenarios, it can be concluded that the ACFDR method can be deemed preferable to the CFDR approach.

The best performing FDR-based methods considered so far have been also employed to detect the ending date of the bubble, BED hereafter, with the exception of LORD without calibrated p-values, and the FNR approach with no self-calibration. The former was excluded because it showed a poor performance in estimating the BBD, and the latter because it was too computationally intensive or underperforming in the previous analysis with respect to the self-calibrated FNR. As no online methods are currently available to control for FNR, the CLR approach, which was specifically devised to control for FDR, was used for this purpose by assuming the end of the null hypothesis rejection streak as an estimation of the BED.

Tables 3 and 4 displays the results, while figures 8 and 9 show the BED boxplots for each method across $B = 500$ simulations.

The methods show different performances in the two scenarios considered. Scenario I reveals that ACFNR and CLR are the only effective methods for BED detection, but solely in DGPs 1 and 2 and with high variability across replications. Conversely, PH, CPH and PS fail to detect BED: the mentioned approaches instead perform very well in Scenario

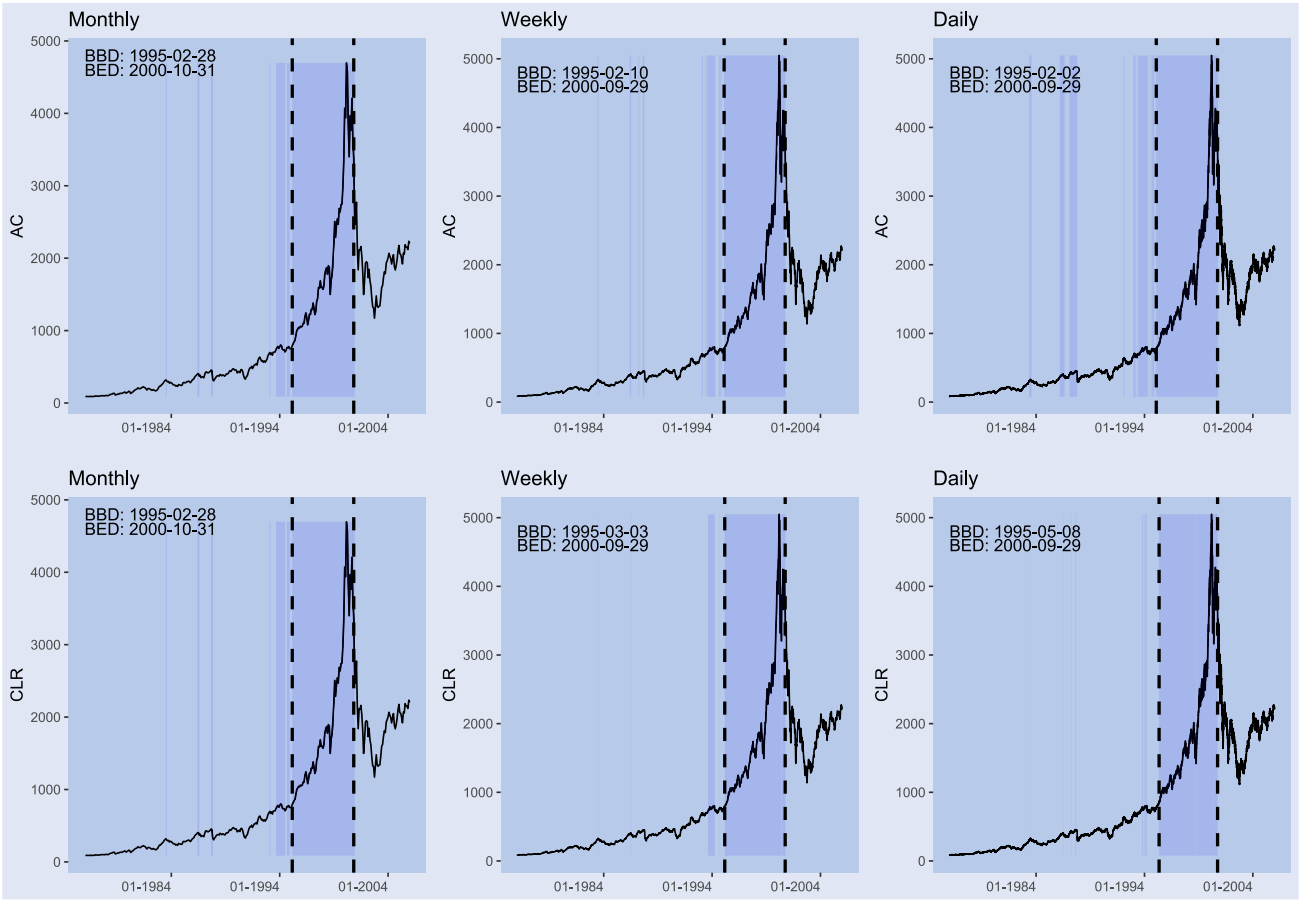

Figure 12. Dating of BBD and BED at FDR/FNR control $\alpha = 0.05$. Black vertical lines denote BBD and BED.

II, since all the methods, except PH and PS, manage to detect the BED. Indeed their estimates fall within the 95% confidence intervals. The ACFNR and CLR methods turn out to be the least variable ones and behave similarly. The outcomes of Scenario II allow us to conclude that when the pace of the bubble bursts is slow enough, the detection of its ending can be pinpointed more easily. In contrast, scenarios with sharp bubble bursts, as observed in Scenario I, exhibit a delay of several days in detecting the bubble's end date.

Furthermore, ACFNR is the only method among the ones here considered that actually makes use of the false non-discovery rate, while the CLR, despite behaving similarly, is inaccurately based on the FDR premise. Lastly, it is also worth mentioning that there may be a trade-off between the FDR and FNR that is similar to that between type I and type II errors in single hypothesis testing. This topic, however, has not been assessed since it does not play a relevant role in our analysis, insofar as different threshold levels can be employed for obtaining these statistics.

4. An empirical application

In this section the FDR-based approaches considered thus far are now applied to real data. To this end, as in Phillips and Yu (2011), we have considered the Nasdaq Composite (IXIC), from Jan, 1976 to Dec, 2006 at different frequencies of observation: end-of-month, end-of-week, and daily. In the authors' original work, only end-of-month prices were considered. Therefore, to add a further layer of interpretation and better gauge the robustness of the methods, we have considered other index frequencies. For each of the above sampling frequencies, we have implemented the PH, CPH, PS, CFDR, ACFDR, ACFNR, LR and the CLR approaches for detecting the bubble burst date (BBD) and the end date (BED). The results are displayed in figures 10 and 11 for the level $\alpha = 0.05$ (the outcomes for significance levels $\alpha = \{0.01, 0.1\}$ are shown in appendix). The shadowed areas in these figures denote the sample periods when the null hypothesis is rejected.

The conclusions that can be drawn on BBD, based on this analysis, are the following: the PH test at all significance levels, except for $\alpha = 0.1$, behaves similarly to the LORD approach when monthly data are considered. In addition, the PH test is not robust to the sampling frequency. In fact, in the case of monthly data, the tests' outcomes are coherent with those of Phillips's original work and do not show many early discoveries which instead occur once the sampling frequency increases. Finally, PS works best with daily data, since the monitoring window for the BSADF statistic is based on a wider sample, or with a significance level of $\alpha = 0.1$. However, in these cases some potential false discoveries are observed. On the contrary, PS performs poorly at $\alpha = 0.01$ when using monthly data, as displayed in figure A1.

Calibration of the p-values allows the PH test to discard several discoveries, especially for daily sampling frequency data. However, calibration does not permit an appropriate control of false discoveries.

The CFDR method provides results similar to those engendered by the CPH test, especially for daily sampling frequencies, where both methods provide poor results. An explanation could be that measuring daily returns produces more frequent price fluctuations, making the 'offline' procedure too sensitive to the rejection of the null hypothesis.

Employing the LR procedure without p-value calibration raises several concerns, since the windows detected by this method for measuring bubble length are the narrowest among the FDR based methods here considered. It is widely known that, one of the assumptions which guarantees the FDR's control of the LR procedure is the super-uniformity of the p-value distribution under the null, given in (16). This assumption is not guaranteed using non-calibrated p-values as the test does.

The CLR method provides results similar to those achievable with the other FDR based methods, namely ACFDR and CFDR. In particular, the latter engenders rejections earlier than both CLR and ACFDR, especially as the significance level increases. Both CLR and ACFDR are the best operating methods, since they immediately detect the sharp price increases observed in the graph. Moreover, they are quite stable regardless of the sampling frequency. Lowering either the significance level α (for significance based methods) or the threshold level α^* (for FDR-based methods) make both the CLR and CFDR capable of engendering more accurate estimates for BBD, while this does not occur for the PH and CPH tests.

As for BED, we have restricted the analysis among the best-performing methods, namely ACFNR and CLR. As mentioned above, ACFNR is the only method that exploits the false non-discovery rate, while CLR is essentially an FDR-based procedure. Figure 12 shows the results for $\alpha = 0.05$; our final choice for BBD and BED is highlighted by dashed lines (The outcomes of the different tests for levels $\alpha = \{0.01, 0.1\}$ are shown in appendix).

Overall, the conclusion is that there is little to no difference between the proposed methods for detecting BED. The small discrepancies in the outcomes of the mentioned methods are difficult to analyze. In fact, unlike in the simulation study, where the origination and ending dates of the bubble are set by design, in the empirical application here considered, it is debatable where to pinpoint the exact origination and ending dates.

5. Conclusion

In the context of targeting and dating financial bubbles, this work has explored two new avenues to revisit and improve well-established results in the literature.

First, the calibration of the p-values of the ADF test, which is implemented sequentially as proposed by Phillips and Yu (2011) for detecting bubbles, guarantees good theoretical properties of the testing procedure, thus allowing to perform more accurate inferences. Indeed, simulation experiments prove that unit root tests, when applied sequentially, do not enjoy the super-uniformity property needed to carry out valid inferences. p-value calibration recovers this property making them usable for detecting and dating bubbles.

Secondly, to accommodate the multiple testing nature of the bubble dating problem, false and non-false discovery (FDR/FNR) detection methods are employed to further refine our proposal. An online procedure finalized to control for false discovery, known as LORD, is also considered. All the mentioned approaches have been implemented with either original p-values or calibrated p-values. In particular, original self-calibrated versions of both FDR and FNR are provided in order to make our proposal easier to use.

A simulation study highlights the strength of the calibration and the necessity to accurately control for false discoveries. LORD and FDR with self-calibration emerge as the two best-performing methods, each with its own specificities: when the bubble explosion is not sharp and immediate, LORD is more sensitive and detects bubbles more frequently, yet at a slightly later date, while FDR with self-calibration is a little less sensitive but it detects bubbles a little earlier than LORD.

Finally, an empirical application reinforces our objective. Using p-value calibration and the appropriate strategy for false discoveries detection allows to correctly identify the bubble, regardless of the observation frequency. In this context, LORD is more effective at detecting possible false discoveries than the self-calibrated version of the false discovery rate, especially at daily frequency. Nevertheless, FDR with self-calibration provides more accurate estimates of the bubble burst.

Further research should be devoted to detecting the ending date of financial bubbles. As remarked by Demos and Sornette (2017), it is indeed the ends of bubbles that are devilishly difficult to detect and forecast with any useful accuracy. Due to the particular nature of the false-non-discovery rate, also the methods presented in this work have difficulties in detecting the end date, and none of the employed FNR strategies emerge as outperforming the others.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Gianmarco Vacca http://orcid.org/0000-0002-8996-5524

References

Astill, S., Harvey, D.I., Leybourne, S.J., Sollis, R. and Taylor, A.M.R., Real-time monitoring for explosive financial bubbles. *J. Time Ser. Anal.*, 2018, **39**, 863–891.

Astill, S., Harvey, D.I., Leybourne, S.J. and Taylor, A.R., Tests for an end-of-sample bubble in financial time series. *Econom. Rev.*, 2017, **36**, 651–666.

Bajgrowicz, P. and Scaillet, O., Technical trading revisited: False discoveries, persistence tests, and transaction costs. *J. Financ. Econ.*, 2012, **106**, 473–491.

Bajgrowicz, P., Scaillet, O. and Treccani, A., Jumps in high-frequency data: Spurious detections, dynamics, and news. *Manage. Sci.*, 2016, **62**, 2198–2217.

Barras, L., Scaillet, O. and Wermers, R., False discoveries in mutual fund performance: Measuring luck in estimated alphas. *J. Finance*, 2010, **65**, 179–216.

Benjamini, Y. and Hochberg, Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, 1995, **57**, 289–300.

Benjamini, Y. and Yekutieli, D., The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, 2001, **29**, 1165–1188.

Brunnermeier, M.K., Bubbles. In *Banking Crises: Perspectives from The New Palgrave Dictionary*, pp. 28–36, 2016 (Palgrave Macmillan: London).

Campbell, J.Y. and Shiller, R.J., Cointegration and tests of present value models. *J. Political Econ.*, 1987, **95**, 1062–1088.

Demos, G. and Sornette, D., Birth or burst of financial bubbles: Which one is easier to diagnose? *Quant. Finance*, 2017, **17**, 657–675.

Diba, B. and Grossman, H., Rational bubbles in the price of gold. NBER Working Paper: 1300. Cambridge, MA: National Bureau of Economic Research, 1984.

Efron, B., Correlated z-values and the accuracy of large-scale statistical estimates. *J. Am. Stat. Assoc.*, 2010, **105**, 1042–1055.

Evans, G.W., Pitfalls in testing for explosive bubbles in asset prices. *Am. Econ. Rev.*, 1991, **81**, 922–930.

Fisher, A.J., Online control of the false discovery rate under "Decision Deadlines". In *International Conference on Artificial Intelligence and Statistics*, pp. 8340–8359, 2022.

Foster, D.P. and Stine, R.A., α-investing: A procedure for sequential control of expected false discoveries. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 2008, **70**, 429–444.

Gürkaynak, R.S., Econometric tests of asset price bubbles: Taking stock. *J. Econ. Surv.*, 2008, **22**, 166–186.

Hafner, C.M., Testing for bubbles in cryptocurrencies with time-varying volatility. *J. Financ. Econom.*, 2020, **18**, 233–249.

Hamilton, J.D. and Whiteman, C.H., The observable implications of self-fulfilling expectations. *J. Monet. Econ.*, 1985, **16**, 353–373.

Harvey, D.I., Leybourne, S.J., Sollis, R. and Taylor, A.R., Tests for explosive financial bubbles in the presence of non-stationary volatility. *J. Empir. Finance*, 2016, **38**, 548–574.

Harvey, D.I., Leybourne, S.J. and Zu, Y., Testing explosive bubbles with time-varying volatility. *Econom. Rev.*, 2019, **38**, 1131–1151.

Harvey, D.I., Leybourne, S.J. and Zu, Y., Sign-based unit root tests for explosive financial bubbles in the presence of deterministically time-varying volatility. *Econ. Theory*, 2020, **36**, 122–169.

Hu, Y., A review of Phillips-type right-tailed unit root bubble detection tests. *J. Econ. Surv.*, 2023, **37**, 141–158.

Jarrow, R.A. and Protter, P., The martingale theory of bubbles: Implications for the valuation of derivatives and detecting bubbles. In *Financial Crisis: Debating the Origins, Outcomes, and Lessons of the Greatest Economic Event of Our Lifetime*, edited by Arthur Berd, 2010 (Risk Publications).

Jarrow, R.A., Protter, P. and Shimbo, K., Asset price bubbles in incomplete markets. *Math. Finance*, 2010, **20**, 145–185.

Javanmard, A. and Montanari, A., Online rules for control of false discovery rate and false discovery exceedance. *Ann. Stat.*, 2018, **46**, 526–554.

LeRoy, S.F. and Porter, R.D., The present-value relation: Tests based on implied variance bounds. *Econometrica*, 1981, **49**, 555–574.

Lui, Y.L., Phillips, P.C.B. and Yu, J., Robust testing for explosive behavior with strongly dependent errors. *J. Econom.*, 2024, **238**(2), 105626.

Magdalinos, T., Mildly explosive autoregression under weak and strong dependence. *J. Econom.*, 2012, **169**, 179–187.

Pedersen, T.Q. and Schütte, E.C.M., Testing for explosive bubbles in the presence of autocorrelated innovations. *J. Empir. Finance*, 2020, **58**, 207–225.

Phillips, P.C. and Magdalinos, T., Limit theory for moderate deviations from a unit root. *J. Econom.*, 2007, **136**, 115–130.

Phillips, P.C. and Shi, S., Detecting financial collapse and ballooning sovereign risk. *Oxf. Bull. Econ. Stat.*, 2019, **81**, 1336–1361.

Phillips, P.C.B. and Shi, S., Chapter 2–real time monitoring of asset markets: Bubbles and crises. In *Financial, Macro and Micro Econometrics using R*. Vol. 42 of Handbook of Statistics, edited by H. D. Vinod and C. Rao, pp. 61–80, 2020 (Elsevier).

Phillips, P.C. and Shi, S.P., Financial bubble implosion and reverse regression. *Econ. Theory*, 2018, **34**, 705–753.

Phillips, P.C., Shi, S.P. and Yu, J., Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *Int. Econ. Rev.*, 2015, **56**, 1043–1078.

Phillips, P.C., Wu, Y. and Yu, J., Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *Int. Econ. Rev.*, 2011, **52**, 201–226.

Phillips, P.C. and Yu, J., Dating the timeline of financial bubbles during the subprime crisis. *Quant. Econom.*, 2011, **2**, 455–491.

Sala, S., Quatto, P., Valsasina, P., Agosta, F. and Filippi, M., pFDR and pFNR estimation for brain networks construction. *Stat. Med.*, 2014, **33**, 158–169.

Scherbina, A. and Schlusche, B., Asset price bubbles: A survey. *Quant. Finance*, 2014, **14**, 589–604.

Shao, J. and Tu, D., *The Jackknife and Bootstrap*, 2012 (Springer Science & Business Media: New York).

Shiller, R.J., *et al.*, Alternative tests of rational expectations models: The case of the term structure. *J. Econom.*, 1981, **16**, 71–87.

Stiglitz, J.E., Symposium on bubbles. *J. Econ. Perspect.*, 1990, **4**, 13–18.

Storey, J.D., A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 2002, **64**, 479–498.

Su, C.W., Wang, K.H., Chang, H.L. and Dumitrescu-Peculea, A., Do iron ore price bubbles occur? *Resour. Policy*, 2017, **53**, 340–346.

Van der Vaart, A.W., *Asymptotic Statistics*, 2000 (Cambridge University Press: Cambridge).

Whitehouse, E.J., Explosive asset price bubble detection with unknown bubble length and initial condition. *Oxf. Bull. Econ. Stat.*, 2019, **81**, 20–41.

Whitehouse, E.J., Harvey, D.I. and Leybourne, S.J., Real-time monitoring of bubbles and crashes. *Oxf. Bull. Econ. Stat.*, 2023, **85**, 482–513.

Yu, H., A Glivenko-Cantelli lemma and weak convergence for empirical processes of associated sequences. *Probab. Theory Relat. Fields*, 1993, **95**, 357–370.

Zrnic, T., Ramdas, A. and Jordan, M.I., Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.*, 2021, **22**, 33-1.

Appendix. Additional figures in the empirical application

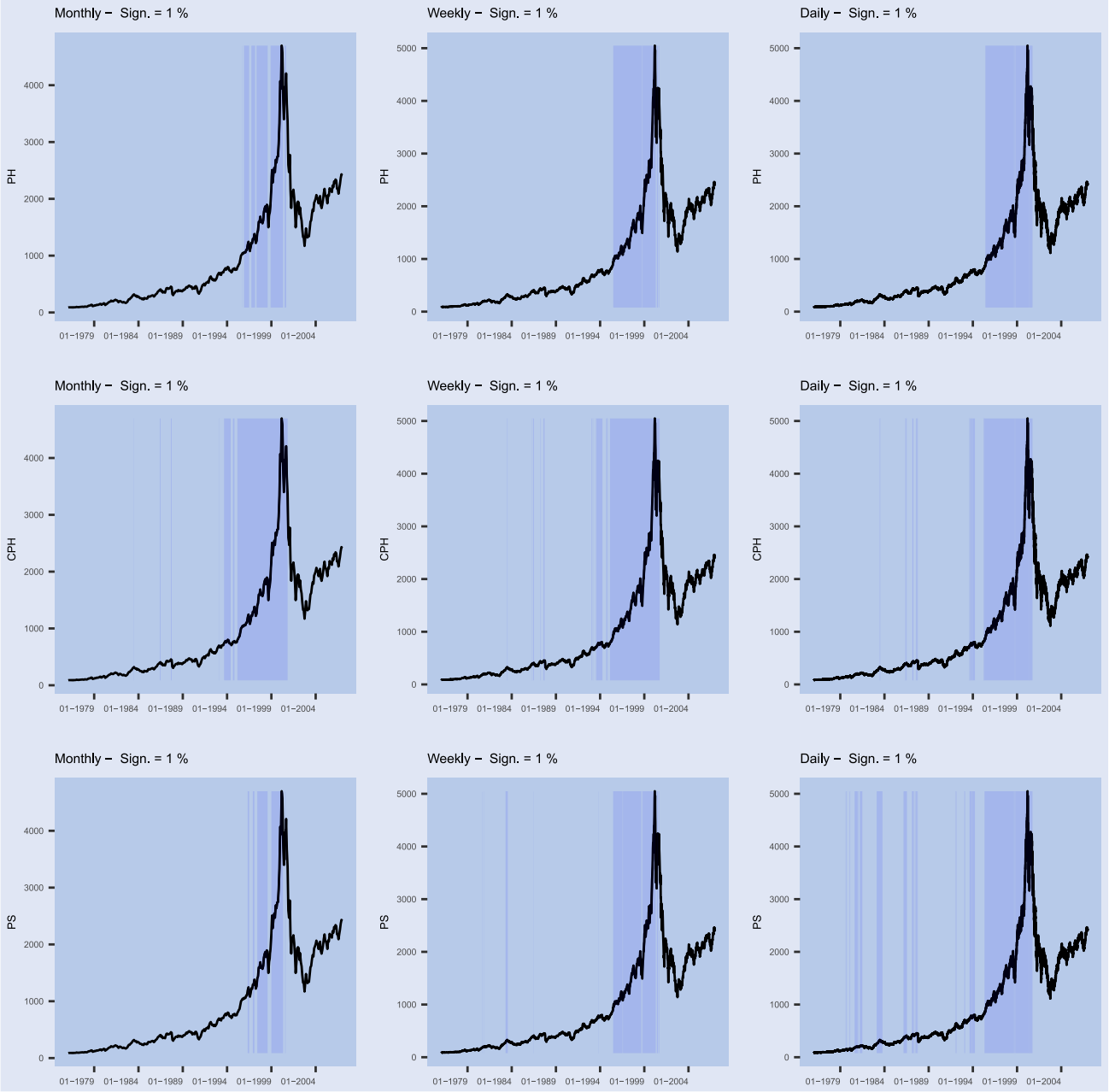

Figure A1. Detection of BBD: PH, CPH and PS at significance $\alpha = 0.01$. Shadowed areas denote sample periods when the null is rejected.

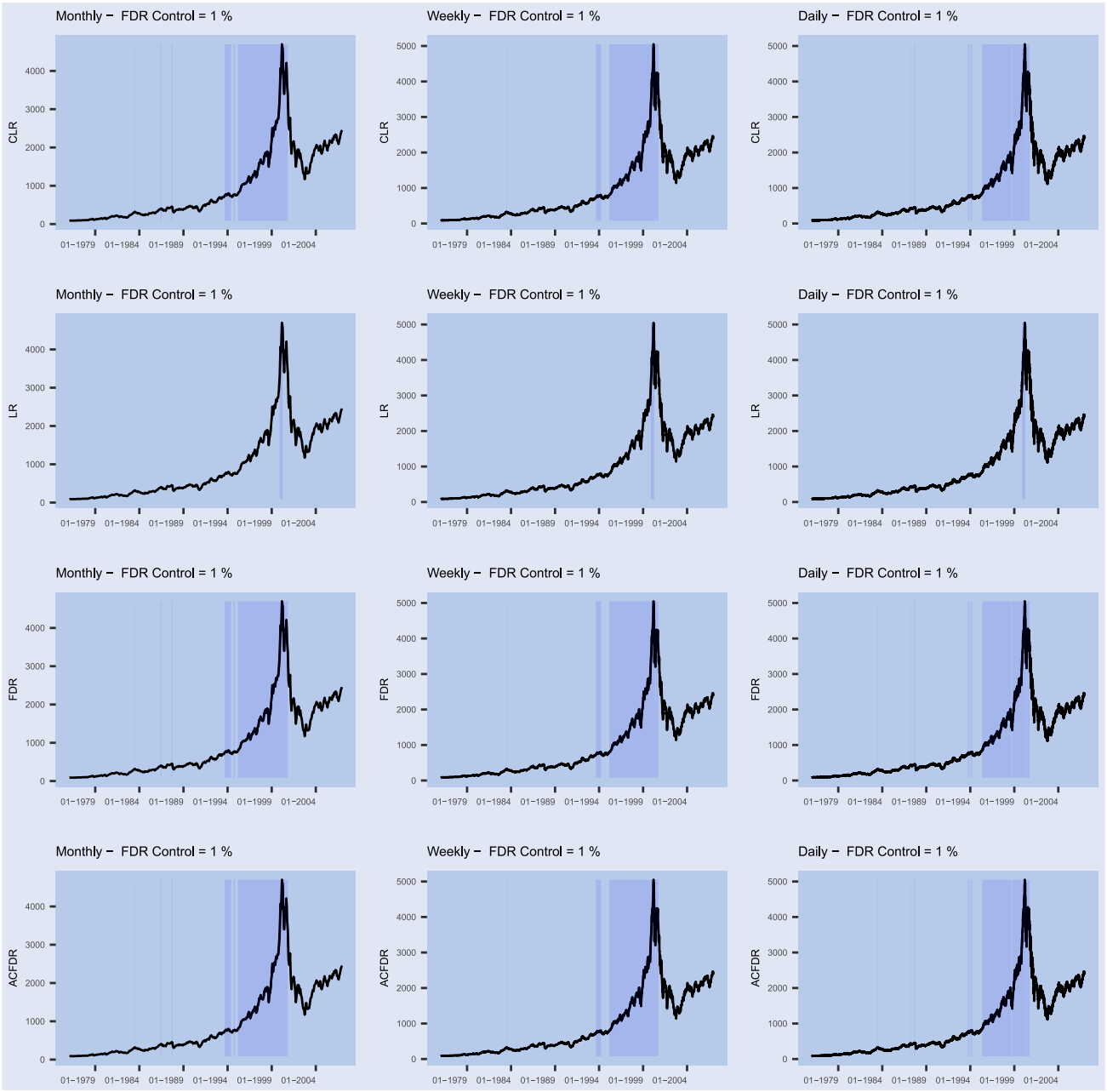

Figure A2. Detection of BBD: (C-AC)FDR and (C)LR at FDR control $\alpha = 0.01$. Shadowed areas denote sample periods when the null is rejected.

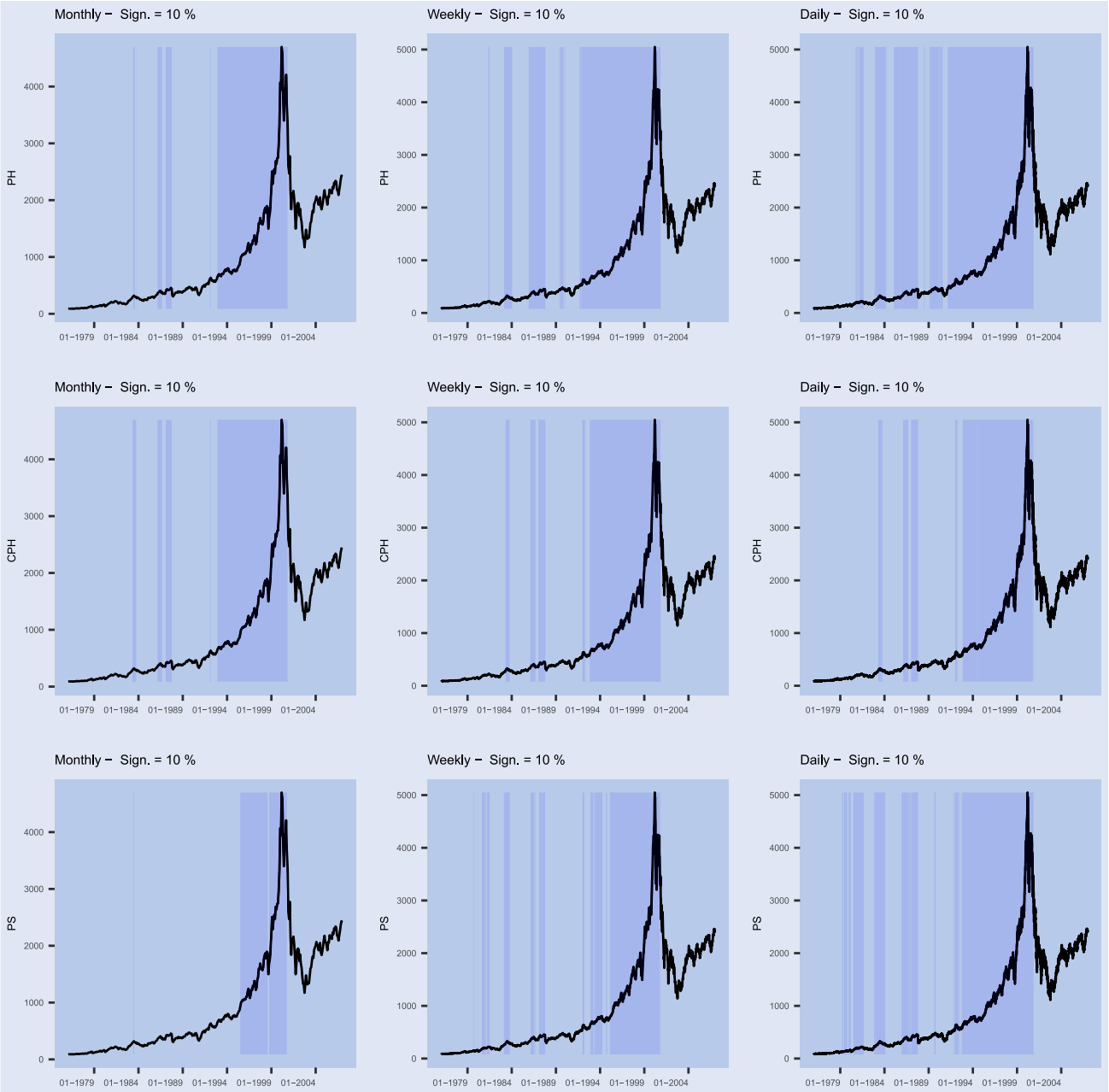

Figure A3. Detection of BBD: PH, CPH and PS at significance $\alpha = 0.1$. Shadowed areas denote sample periods when the null is rejected.

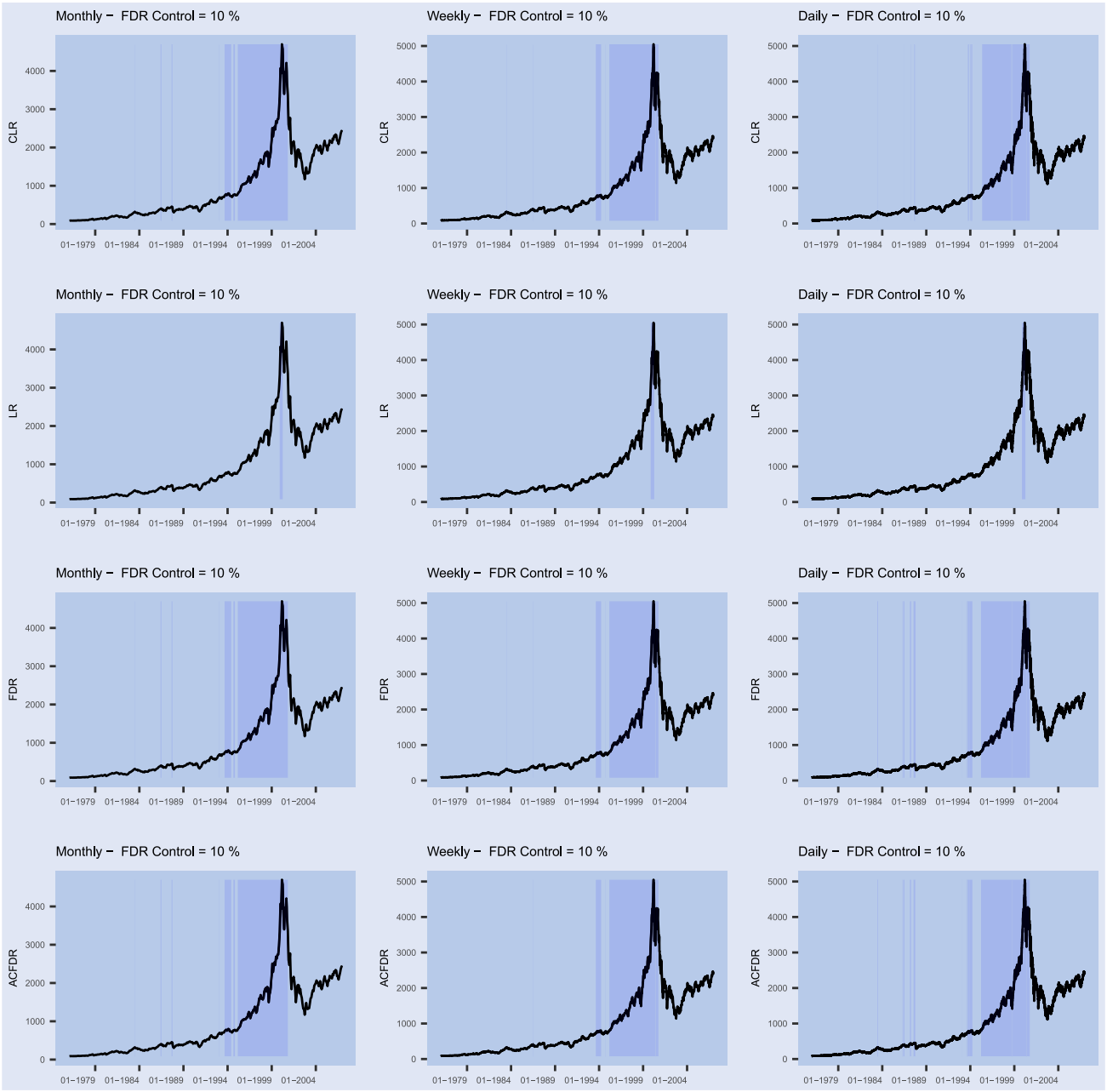

Figure A4. Detection of BBD: (C-AC)FDR and (C)LR at FDR control $\alpha = 0.1$. Shadowed areas denote sample periods when the null is rejected.

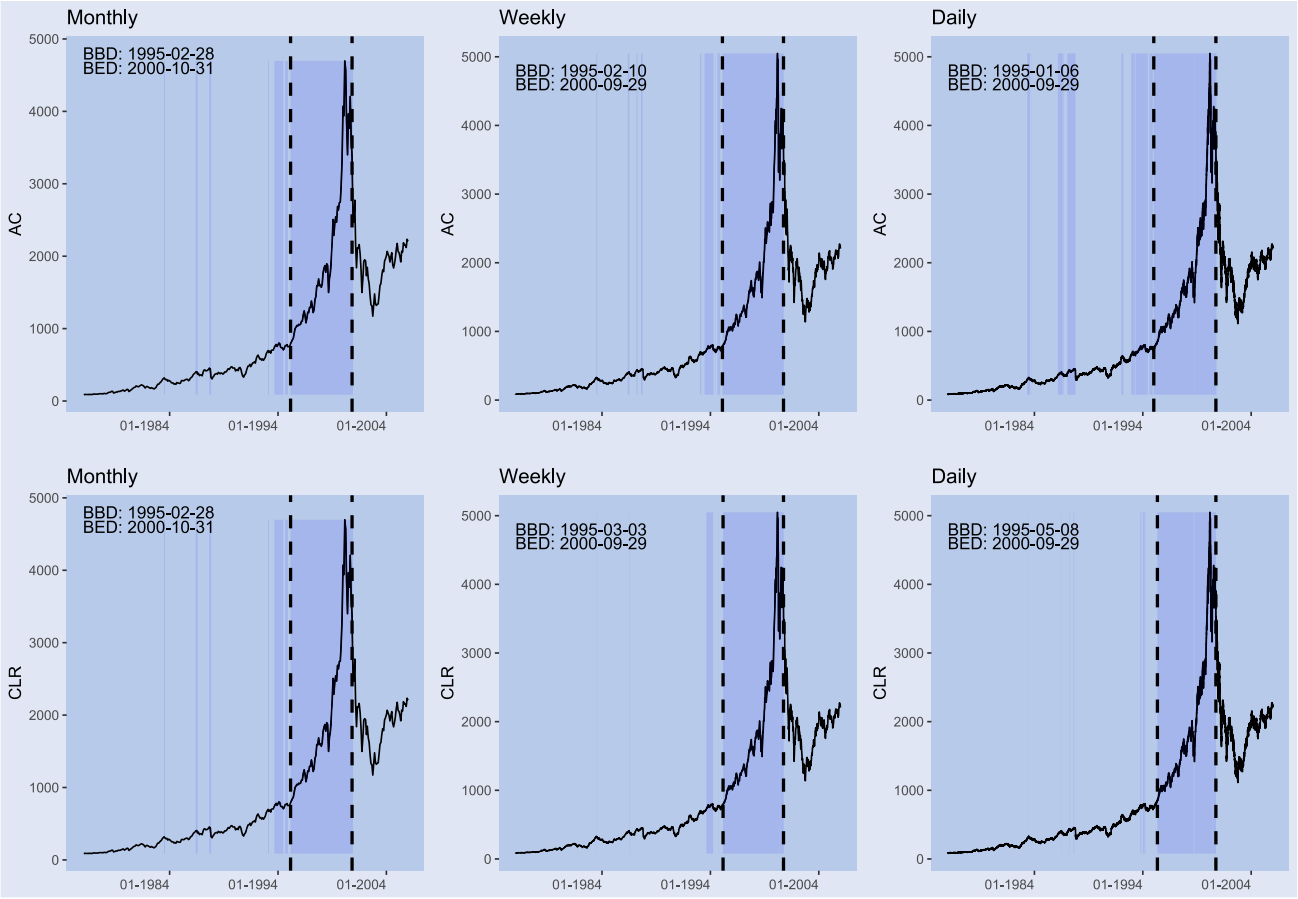

Figure A5. Dating of BBD and BED at FDR/FNR control $\alpha = 0.01$. Black vertical lines denote BBD and BED.

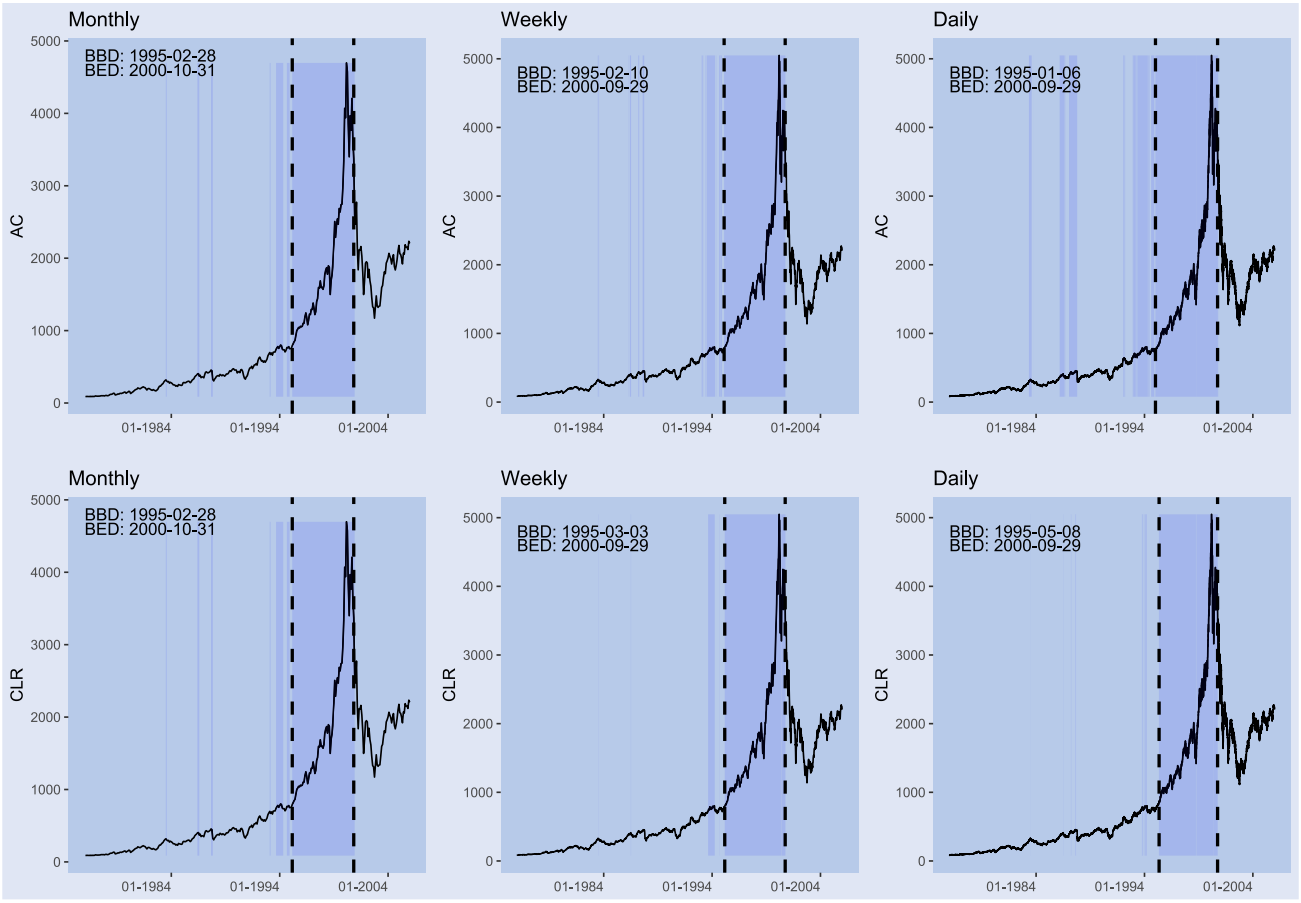

Figure A6. Dating of BBD and BED at FDR/FNR control $\alpha = 0.1$. Black vertical lines denote BBD and BED.