Testing for Speculative Bubbles: Revisiting the Rolling Window

J. Chong and A.S. Hurn

School of Economics and Finance, Queensland University of Technology

Abstract

Recent research on detecting asset pricing bubbles in real-time has focussed on recursive and rolling-recursive regressions in combination with the supremum norm of a sequence of right-tailed unit root tests. The rolling-recursive algorithm, in particular, has proved relatively successful in identifying the timeline of bubbles but it does suffer from the disadvantages of being computationally quite intensive and also requiring the use of non-standard limit theory. This paper re-evaluates a more simple and perhaps somewhat neglected approach to the date-stamping of bubbles, namely the rolling-window unit root testing approach, and provides a comprehensive comparison of its performance against the recursive and rolling-recursive methods. The results of a suite of simulation experiments indicate that rolling-window tests may in fact be superior to the other two methods. In addition, the rolling-window approach yields better inference than its competitors when applied to a sample of the Nasdaq stock index and a sample of U.S. housing price-to-rent ratios, both of which are known to contain bubbles.

Keywords

Financial bubble, date-stamping strategy, multiple bubbles, mildly explosive bubbles, Monte Carlo simulations, rolling windows

JEL Classification C12, C14, C22

Corresponding author:

Jieyang Chong <yang.chong@qut.edu.au>

School of Economics and Finance

Queensland University of Technology

2 George Street, Brisbane, 4001 Australia

1 Introduction

The periodic appearance and collapse of speculative bubbles in asset prices has been a source of fascination ever since the infamous bubble in the Dutch tulip market in 1637. Then, just as now, bubbles in asset prices seem to be a precursor to periods of economic instability or even crisis. The Dot-Com bubble in the late 1990s, the United States housing bubble in the mid 2000s and the Chinese stock price bubble in 2015 are all recent examples of bubbles which preceded crises. At the time of writing, many asset markets are conjectured to be in the grips of a speculative bubble: if media claims are to be believed then the United States stock market as well as the housing markets in China, Hong Kong, and Australia are all currently experiencing bubble conditions. Until relatively recently these claims could only be substantiated retrospectively, but given their importance to central banks and regulators, it is no surprise that the detection and dating of bubbles in real time is receiving more attention in the econometric literature.

Various approaches to testing for the presence of bubbles have been proposed. Among these are variance bounds tests (LeRoy and Porter, 1981 and Shiller, 1981), West's two-step test (West, 1987), fractionally integrated models (Cuñado, Gil-Alana and De Gracia, 2005 and Frömmel and Kruse, 2012) and recursive unit root tests (Phillips, Wu and Yu, 2011). This paper focusses on the use of unit root tests to detect the presence of and identify the timeline of bubbles in asset prices. Early attempts at bubble detection employed traditional *left-tailed* unit root tests to test the null hypothesis of nonstationarity of the price process against the alternative of stationarity (Diba and Grossman, 1988). These tests had low power against the presence of bubbles leading Campbell, Lo and MacKinlay (1997) to conclude that these traditional unit root tests provided little or no statistical evidence of explosive behaviour. In a fairly recent development, Phillips, Wu and Yu (2011), suggest using *right-tailed* unit root tests to detect bubbles. In this version of the test, the null hypothesis of non-stationarity is tested against the alternative of mildly explosive behaviour in the price process. When the test is conducted recursively, it is able to detect when the series switches from being generated under the null hypothesis to when it is explosive and vice versa, thus estimating the origination and collapse of bubbles. Early implementations of this testing procedure yielded promising results and subsequent refinements by Phillips, Shi and Yu (2015a, 2015b) have developed rolling-recursive versions of the test that allow for bubble detection and dating the origin and collapse of multiple bubbles. This area of research has now been taken up by, *inter alia*, Phillips and Yu (2011), Homm and Breitung (2012), Gutierrez (2013), Harvey, Leybourne and Sollis (2015a, 2015b) and Harvey, Leybourne, Sollis and Taylor (2015c).

A testing method which featured in Phillips et al. (2011) only as a robustness check is a rolling-window approach, as opposed to the recursive approach, in which a window of fixed length is run

along the entire length of the sample. Of course the rolling-window procedure is a subset of the rolling-recursive algorithm of Phillips et al. (2015b) but the simplicity of the approach, together with some evidence that the more intensive procedures produce results that vary according to the location of bubbles within the sample (Homm and Breitung, 2012 and Phillips et al., 2015b), provide solid motivation for a re-evaluation of the method. In addition, relatively recent work by Gutierrez (2013) shows that the rolling-window approach may have higher power than the recursive method when the degree of explosiveness is low, especially when the bubble is located near the end of the sample.

The central contribution of this paper is provided by a thorough re-evaluation of the rolling-window testing framework for bubble dating by means of a comprehensive suite of simulation experiments and application to time series data known to contain bubbles. One of the central tenets of the research is that estimating the origin and termination of bubbles is perhaps of greater importance and practicality than merely detecting the presence of bubbles within a sample and emphasis will therefore be placed on this aspect of the problem. The evaluation in this paper provides insight into the qualities of the rolling-window framework on its own as well as in comparison to the recursive and rolling-recursive methods of Phillips et al. (2011) and Phillips et al. (2015a, 2015b).

A brief synthesis of the main results generated in this paper is as follows. The rolling-window testing procedure is shown to have more desirable size and power properties, better rates of detection and more accurate estimation of the origination of simulated bubbles than its competitors. In addition, when an asymmetric loss function is used to evaluate bubble detection, based on the reasonable proposition that from a policy perspective, failing to detect a bubble is more costly than a false alarm, the rolling-window procedure is (more often than not) less costly than the recursive and rolling-recursive methods and in some instances is even superior on all counts to its competitors. The method is then used to detect and date bubbles in monthly observations on the Nasdaq composite index for the period February 1973 to July 2015 and quarterly United States house prices from January 1975 to April 2015. Based on the general consensus about the approximate dates of major bubbles within the samples, results from using the rolling-window procedure suggest that it detects bubbles earlier than the other methods commonly used. Finally, an additional advantage of the use of rolling windows is found: successful detection of bubbles at each point in the series is completely independent of when the sample begins, whereas recursive procedures are affected by sample choice.

2 Bubble Detection

The fundamental price of the asset is derived from the no arbitrage condition

$$P_t = \frac{E_t[P_{t+1} + D_{t+1}]}{1+R}, \tag{1}$$

where P_t is the price of the stock at period t, D_t denotes the dividend received from ownership of the stock between $t-1$ and t, and R is the discount rate. Using the present value theory of finance and solving (1) by forward iteration yields

$$P_t^f = \sum_{i=1}^{\infty} \frac{1}{(1+R)^i} E_t(D_{t+i}),$$

in which the fundamental price of an asset in any period is equal to the present value of all expected dividend payments from that point onwards. If the transversality condition

$$\lim_{k\to\infty} E_t\left[\frac{1}{(1+R)^k} P_{t+k}\right] = 0 \tag{2}$$

holds, then the current price of the asset, P_t is equal to the fundamental price of the asset, P_t^f. However if (2) does not hold, an explosive rational bubble can exist. Let B_t denote the bubble component and be defined as

$$E_t[B_{t+1}] = (1+R)B_t. \tag{3}$$

Adding B_t to P_t^f will yield infinitely many solutions for the current price of the asset, which takes the form

$$P_t = P_t^f + B_t.$$

Since stock prices must be nonnegative, it is important to only consider cases where $B_t \geq 0$. Even though the bubble series is restricted to strictly positive values, it need not grow exponentially from start to end of the sample. It can take a constant positive value for some time and begin to grow exponentially at some point in the sample. Since rational bubbles must eventually collapse, it is also relevant to include a subsequent structural break which allows B_t to fall back to some constant positive value, reflecting the bursting of the bubble.

Given this model of asset prices and bubbles, it is natural to attempt to detect bubbles by testing for explosiveness against the null hypothesis of a unit root. Existing unit root tests typically have an alternative hypothesis of stationarity, and are *left-tailed* tests. Conducting unit root tests but looking instead in the *right tail* of the distribution of the test statistic represents testing the null

hypothesis of nonstationarity against the alternative hypothesis of explosive behaviour. It turns out, however, that right tailed unit root tests suffer from the same shortcoming that afflicts their left tail counterparts, namely that they suffer from low power. In fact, in right tailed tests this problem may even be exacerbated by the nature of bubbles, which do not last forever but collapse at some point in time. In particular, Evans (1991) demonstrates that in a model containing a periodically-collapsing bubble, full-sample right-tailed unit root tests have low power.

2.1 Supremum tests

In order to overcome this problem, Phillips et al. (2011) propose a recursive algorithm for right-tailed unit root testing and show that it has superior power to the simple full-sample alternative. The algorithm involves a simple forward recursion in which the test statistic is computed at each recursion and inference is based on the supremum norm of the sequence of test statistics. Formally, for a sample $(y_1, \ldots, y_T)$ the test statistic is given by

$$\mathcal{S} := \sup_{r \in [r_0, 1]} \mathcal{T}_r, \tag{4}$$

where $0 < r_0 < 1$, r_0 determines the smallest sub-sample of the data on which the researcher wishes to conduct the test, $\mathcal{T}_r$ is the test statistic for which $y_{\lfloor Tr \rfloor}$ is a defining point (as shown in formulations of test statistics in this section) and $\lfloor \cdot \rfloor$ denotes the integer part of its argument. A visual representation of the procedure is shown in Panel (a) of Figure 1. The sample contains a period of explosiveness (and hence a bubble) if $\mathcal{S}$ is greater than its relevant critical value. Using right-tailed, as opposed to left-tailed tests allows one to detect the presence explosiveness, instead of merely the absence of stationarity. The recursive method is powerful in the event of a collapsed bubble, whereas a single full-sample right-tailed unit root test would have low power under such circumstances.

The forward recursive bubble detection algorithm can be used with any one of a number of unit root tests. Examples of unit root tests which have been adapted specifically for the purpose of recursive bubble detection are now outlined.

Dickey Fuller statistic

Phillips et al. (2011) and Phillips et al. (2014, 2015b) use a right-tailed Dickey-Fuller (DF) statistic (Dickey and Fuller, 1979) for bubble detection. The test regression is

$$\Delta y_t = \alpha + \phi y_{t-1} + \epsilon_t, \tag{5}$$

in which $\epsilon_t \sim N(0, \sigma_\epsilon^2)$ and α is the deterministic drift term. As expounded on by Phillips et al. (2014), the null and alternative hypotheses are

$$\begin{aligned} H_0: &\quad \Delta y_t = \alpha T^{-\eta} + \phi y_{t-1} + \epsilon_t, &&\quad \phi = 0, \\ H_1: &\quad \Delta y_t = \phi y_{t-1} + \epsilon_t, &&\quad \phi > 0, \end{aligned}$$

where the series has a deterministic drift of the form $\alpha t T^{-\eta}$ under the null hypothesis, which depends on sample size, T, and the localising parameter, η. The null hypothesis is tested using a t-test.

There are couple of things to note. The first is that the DF test is preferred to the augmented DF (ADF) test (Said and Dickey, 1984) because Phillips et al., (2015a) demonstrate that the size distortion of the tests increases with lag length. Omitting lags altogether deals with the size distortion issue, and has the added benefit of simplifying the test. The second interesting point about this particular implementation is the presence of a deterministic drift but no time trend in the test regression. The omission of the time trend stems from the fact that the alternative hypothesis is now mildly explosive behaviour rather than the traditional alternative of stationarity. The concurrent appearance of both mildly explosive behaviour and a deterministic time trend under the alternative hypothesis seems unrealistic (Phillips et al., 2014).

The supremum test statistic takes the form

$$\sup DF(r_0) = \sup_{r_0 \leq r \leq 1} DF_r \qquad \text{with} \qquad DF_r = \frac{\hat{\phi}_r - 1}{\hat{\sigma}_{\phi,r}},$$

where DF_r is the Dickey-Fuller statistic for sub-sample $\{y_1, \ldots, y_{\lfloor Tr \rfloor}\}$ and $\hat{\sigma}_{\phi,r}$ is the usual estimator for the standard deviation of $\hat{\phi}$ over the same sub-sample. The limiting distribution is derived by Phillips et al. (2011) as

$$\sup_{r_0 \leq r \leq 1} DF_r \Rightarrow \sup_{r_0 \leq r \leq 1} \frac{\int_0^r W(z) \mathrm{d}W(z)}{\sqrt{\int_0^r W(z)^2 \mathrm{d}z}},$$

where W is a standard Wiener process.

The Bhargava statistic

The locally most powerful invariant test statistic proposed by Bhargava (1986)

$$BH_0^* = \frac{\sum_{t=1}^T (y_t - y_{t-1})^2}{\sum_{t=1}^T (y_t - y_0)^2}$$

is inverted and modified to get a series of statistics

$$BH_r = \frac{1}{T - \lfloor Tr \rfloor} \left(\frac{\sum_{t=\lfloor Tr \rfloor+1}^{T} (y_t - y_{t-1})^2}{\sum_{t=\lfloor Tr \rfloor+1}^{T} (y_t - y_{\lfloor Tr \rfloor})^2} \right)^{-1},$$
$$= \frac{1}{d_r^2 (T - \lfloor Tr \rfloor)^2} \sum_{t=\lfloor Tr \rfloor+1}^{T} (y_t - y_{\lfloor Tr \rfloor})^2,$$

where $d_r^2 = (T - \lfloor Tr \rfloor)^{-1} \sum_{t=l\lfloor Tr \rfloor+1}^{T} (y_t - y_{t-1})^2$ and $r \in (0, 1)$. To test for the presence of explosiveness within the sample, the supremum statistic

$$\sup BH(r_0) = \sup_{r \in [0, 1-r_0]} BH_r$$

is compared against the relevant critical value. If the statistic is larger than the critical value, explosiveness exists within the sample.

The asymptotic distribution of this supremum test statistic is given in Homm and Breitung (2012) as

$$\sup BH(r_0) \Rightarrow \sup_{r \in [0, 1-r_0]} \left\{ (1 - r)^{-2} \int_r^1 (W(z) - W(r))^2 \, \mathrm{d}r \right\},$$

where $\Rightarrow$ denotes weak convergence and W denotes standard Brownian motion on the interval $[0, 1]$.

The Busetti-Taylor statistic

The statistic proposed by Busetti and Taylor (2004) tests the hypothesis that a time series is stationary against the alternative that it switches from a stationary to a random walk process at an unknown breakpoint. Homm and Breitung (2012) modify the standard Busetti-Taylor statistic to obtain

$$BT_r = \frac{1}{d_0^2 (T - \lfloor Tr \rfloor)^2} \sum_{t=\lfloor Tr \rfloor+1}^{T} (y_T - y_{t-1})^2.$$

The test for the presence of bubbles uses the supremum test statistic

$$\sup BT(r_0) = \sup_{r \in [0, 1-r_0]} BT_r,$$

where s_0^2 is the variance estimator based on the entire sample. Since this is now a test for explosiveness instead of stationarity, this test rejects for large values of $\sup BT(r_0)$. The limiting distribution of the supremum statistic is given as

$$\sup_{r \in [0, 1-r_0]} BT_r \Rightarrow \left\{ (1 - r)^{-2} \int_r^1 W(1 - z)^2 \mathrm{d}z \right\}.$$

The Kim statistic

Another statistic for testing the null of stationarity against the alternative of nonstationarity was proposed by Kim (2000). A modification to the Kim statistic for recursive testing gives

$$\sup K(r_0) = \sup_{r\in[r_0,1-r_0]} K_r, \qquad \text{where} \qquad K_r = \frac{(T-\lfloor Tr\rfloor)^{-2}\sum_{t=\lfloor Tr\rfloor+1}^{T}(y_t - y_{\lfloor Tr\rfloor})^2}{\lfloor Tr\rfloor^{-2}\sum_{t=1}^{\lfloor Tr\rfloor}(y_t-y_0)^2}.$$

The right-tailed version of this test detects the presence of explosiveness for large values of $\sup K(r_0)$. The limiting distribution is given as

$$\sup_{r\in[t_0,1-r_0]} K_t \Rightarrow \sup_{r\in[r_0,1-r_0]} \left\{ \left(\frac{r}{1-r}\right)^2 \frac{\int_r^1 (W(z)-W(r))^2 \mathrm{d}z}{\int_0^r W(z)^2 \mathrm{d}z} \right\}.$$

Chow-type unit root statistic for structural break

This test incorporates the assumption that y_t is not explosive for the first $\lfloor Tr^*\rfloor$ observations of the sample under both the null and alternative hypothesis for some unknown r^*. Should the sample contain a bubble which begins at $\lfloor Tr^*\rfloor + 1$,, the parameter ρ will be $\phi = 0$ for $t = 1,\ldots,\lfloor Tr^*\rfloor$ and $\phi > 0$ for $t = \lfloor Tr^*\rfloor,\ldots,T$. Thus, the model can also be written as

$$\Delta y_t = \phi(y_{t-1}\mathbb{1}_{\{t>\lfloor Tr\rfloor\}}) + \epsilon_t, \tag{6}$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function which equals 1 when the statement in the braces is true and 0 otherwise. The presence of explosiveness can then be tested using a Chow test for structural breaks in ϕ. The statistic which is to be computed recursively is

$$DFC_r = \frac{\sum_{t=\lfloor Tr\rfloor+1}^{T}\Delta y_t y_{t-1}}{\tilde{\sigma}_r\sqrt{\sum_{t=\lfloor Tr\rfloor+1}^{T} y_{t-1}^2}},$$

where

$$\tilde{\sigma}_r^2 = \frac{1}{T-2}\sum_{t=2}^{T}\left(\Delta y_t - \hat{\phi}_r y_{t-1}\mathbb{1}_{\{t>\lfloor Tr\rfloor\}}\right)^2,$$

with $\hat{\phi}_r$ denoting the OLS estimator of ϕ in (6). The statistic to test for a change from random walk to explosive in the interval $r\in[0,1-r_0]$ is then written as

$$\sup DFC(r_0) = \sup_{r\in[0,1-r_0]} DFC_r.$$

The Chow-type test favours the alternative hypothesis for large values of sup $DFC(r_0)$. The limiting distribution is

$$\sup DFC(r_0) \Rightarrow \sup_{r\in[0,1-r_0]} \frac{\int_r^1 W(z)\mathrm{d}W(z)}{\sqrt{\int_r^1 W(z)^2\mathrm{d}z}}.$$

Note that this test explicitly assumes the null distribution for early observations, so selecting a sample which begins in a bubble may lead to incorrect conclusions.

2.2 Generalised supremum test

A refinement by Phillips et al. (2015a, 2015b) uses the *generalised* supremum statistic, which is the supremum norm of DF statistics computed on all sub-samples containing at least $\lfloor Tr_0 \rfloor$ observations. This version of the test has higher power than the supremum DF test when bubbles occur later in the sample and when there are multiple bubbles in a sample. The generalised supremum DF statistic is given by

$$GSDF := \sup_{\substack{r\in[r_0,1]\\ r_1\in[0,r-r_0]}} DF_{r_1}^r, \tag{7}$$

where $DF_{r_1}^r$ is the DF statistic which corresponds to sub-sample $(y_{\lfloor Tr_1 \rfloor}, \ldots, y_{\lfloor Tr \rfloor})$. Bubbles are present in a sample if the $GSDF$ statistic exceeds its critical value. The limiting distribution of the $GSDF$ statistic is given by Phillips et al. (2015a) as

$$GSDF \Rightarrow \sup_{\substack{r\in[r_0,1]\\ r_1\in[0,r-r_0]}} \left\{ \frac{\frac{1}{2}(r-r_1)\left[W(r)^2 - W(r_1)^2 - (r-r_1)\right] - \int_{r_1}^r W(z)\mathrm{d}z\left[W(r)-W(r_1)\right]}{(r-r_1)^{1/2}\left\{(r-r_1)\int_{r_1}^r W(z)^2\mathrm{d}z - \left[\int_{r_1}^r W(z)\mathrm{d}z\right]^2\right\}^{1/2}} \right\}.$$

3 Date-stamping methods

In the context of detecting bubbles in Evans' (1991) periodically collapsing bubbles model, Homm and Breitung (2012) find that the DF statistic has higher power than all the other unit root tests outlined in Section 2 for almost all simulated values of ϕ. In addition, Phillips et al. (2015b) show that for date-stamping, their rolling-recursive DF approach is superior to other approaches against which it was compared. Consequently, the date-stamping methods which will be considered in this paper will focus only on the use of DF statistics. Date-stamping rational bubbles requires that the DF tests be conducted recursively as the null hypothesis needs to be tested against the alternative hypothesis of a mildly explosive process at each point in time.

An alternative to recursive unit root testing estimates the start and end dates of bubbles through the use of model-based minimum sum of squared residuals estimators to find a model which best fits a given price series (Harvey, Leybourne and Sollis, 2015a). This approach will not be considered here

Three different algorithms for detecting the origination and termination dates of bubbles using DF tests will now be outlined.

PWY test (Phillips, Wu and Yu, 2011)

In order to date-stamp bubbles, the series of DF_r statistics described in Subsection 2.1 is compared against relevant right-tailed critical values. Estimates of the origination point, $\hat{r}_e$, and collapse point, $\hat{r}_f$, of the bubble are constructed as

$$\hat{r}_e = \inf_{s \geq r_0} \left\{ s : DF_s > cv_{\beta_T}^{\text{adf}} \right\}, \ \hat{r}_f = \inf_{s \geq \hat{r}_e + L_T} \left\{ s : DF_s < cv_{\beta_T}^{\text{adf}} \right\},$$

where $cv_{\beta_T}^{\text{adf}}$ is the $100(1-\beta_T)\%$ critical value of the relevant DF statistic and L_T is a restriction on the minimum duration of a bubble. In other words, the bubble originates at the first instance at which the DF statistic exceeds its critical value, and terminates at the first instance following the origination at which the DF statistic ceases to exceed its critical value, or after the minimum bubble duration has passed, whichever comes later.

PSY test (Phillips, Shi and Yu, 2015b)

The PWY dating procedure was shown to perform well for the first bubble in a sample, but to have low power against any subsequent bubbles. To compensate for this deficiency, Phillips et al. (2015b) propose a rolling-recursive procedure in which every single sub-sample (subject only to a minimum window requirement) is tested in a systematic way. Consequently, every observation will have a series of DF statistics associated with it and inference is based on sequence of supremum norms. Panel (b) of Figure 1 provides an illustration of the sample sequence of this procedure. The procedure is shown in simulations to have very good power against multiple bubbles.

The PSY date-stamping procedure uses a *backward-supremum DF* statistic, $BSDF$. The $BSDF$ statistic at observation $\lfloor Tr \rfloor$ of the sample is defined as

$$BSDF_r(r_0) := \sup_{r_1 \in [0, r-r_0]} DF_{r_1}^r, \tag{8}$$

where $\lfloor Tr_0 \rfloor$ is the minimum window size, and $\lfloor Tr_1 \rfloor$ is the first observation for each DF test. Thus, the $BSDF_r$ statistic is simply the largest of the DF statistics for all sub-samples larger than $\lfloor Tr_0 \rfloor$

which have $y_{\lfloor Tr \rfloor}$ as their final observation. The origination and termination point of bubbles are constructed as

$$\hat{r}_e = \inf_{s \geq r_0} \left\{ s : BSDF_s(r_0) > cv_{\beta_T}^{\text{BSDF}} \right\}, \ \hat{r}_f = \inf_{s \geq \hat{r}_e + L_T} \left\{ s : BSDF_s(r_0) < cv_{\beta_T}^{\text{BSDF}} \right\},$$

where $cv_{\beta_T}^{\text{BSDF}}$ is the $100(1-\beta_T)\%$ critical value of the relevant $BSDF$ statistic.

Rolling-window test

A procedure mentioned in passing in PWY, but which has since received little academic attention, is a simple fixed rolling-window right-tailed DF method.[1] This approach takes sub-samples of the data, and is thus also a potential solution to the Evans (1991) critique. Bubble origination and collapse dates estimated under the rolling-window procedure are constructed as

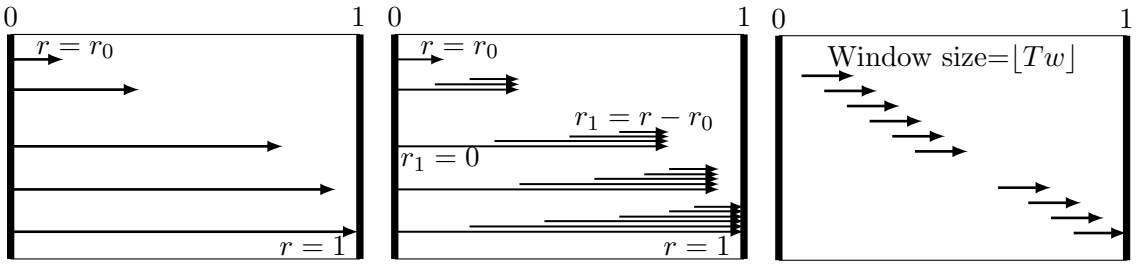
$$\hat{r}_e = \inf_{s \geq w} \left\{ s : DF_s > cv_{\beta_T, w}^{\text{adf}} \right\}, \ \hat{r}_f = \inf_{s \geq \hat{r}_e + L_T} \left\{ s : DF_s < cv_{\beta_T, w}^{\text{adf}} \right\},$$

where $0 < w < 1$ indicates the fixed window size as a proportion of sample size, DF_s is the statistic computed on sub-sample $(y_{s-w+1}, \ldots, y_s)$ for all $s \in [w, T]$, and $cv_{\beta_T, w}^{\text{adf}}$ is the corresponding right-tailed critical value of the DF statistic for sample size w. The sample sequence for this procedure is shown in Panel (c) Figure 1.

If the period between the collapse of a bubble and the origination of a subsequent one is longer than the window size, this method should date multiple bubbles as accurately as the PWY procedure detects the first bubble, subject to optimal selection of w. A schematic comparison of the three recursive algorithms is given in Figure 1.

Note that the rolling-window method is a subset of the rolling-recursive procedure of PSY, replacing $r_1 \in [0, r - r_0]$ in (8) with $r_1 = r - w$. In the PSY test, the set of DF statistics computed for each r are collapsed into a single supremum statistic. The practice of considering only the largest value results in loss of potentially valuable information since all other DF statistics are essentially discarded. Gutierrez (2013) advocates the use of rolling windows over the PWY procedure and provides support for this through Monte Carlo simulations. In the context of forecasting, Clark and McCracken (2009) offer results suggesting that the use of rolling windows produces lower mean squared errors than a recursive method. However, neither Gutierrez (2013) nor Clark and McCracken (2009) draw comparison with rolling-recursive methods akin to PSY.

[1] Phillips et al. (2015a) use the term "rolling window test for bubbles" to refer to their rolling-recursive procedure instead of a fixed-window method.



(a) PWY procedure (b) PSY procedure (c) Rolling window

Figure 1: Sample sequences and window widths for PWY, PSY and rolling-window procedures. Each arrow corresponds to a DF test.

The asymptotic distribution of $BSDF$ statistics used in the PSY test is non-standard and is provided in Phillips et al. (2015b). In contrast, the DF statistics used in the PWY and rolling-window tests rely on the standard DF distribution. However, the PWY test is shown to be inferior to the other two procedures, particularly in the presence of multiple bubbles. Apart from the simplicity of the inference the rolling test also has the advantage of computational parsimony. Of course the test does require the selection of a window size over which to compute the sub-sample tests. For simplicity, this may be set to the minimum window size chosen in the implementation of the PSY test.

A final additional consideration is that the PWY and PSY tests always include regressions from the first observation in a sample. In contrast, the rolling-window test does not. It is likely then that selecting the correct date at which the sample begins is important in order to obtain reliable inference from the PWY and PSY procedures. The fact that the performance of the rolling-window test is independent of sample selection is a potential advantage which is worth exploring.

4 Simulation Experiments

4.1 Simulating Bubbles

The DGP adopted here follows, *inter alia*, Phillips and Yu (2009), Phillips et al. (2015b), and Harvey et al. (2015c). The procedure is to generate a series which follows a random walk under the null hypothesis of no explosive bubbles with a sub-sample in which the process is explosive. For a series $(y_1, \ldots, y_T)$, let τ_{je} and τ_{jf} be the origination and collapse points of the j^{th} bubble

respectively. A general form of the DGP for a series with two bubbles is

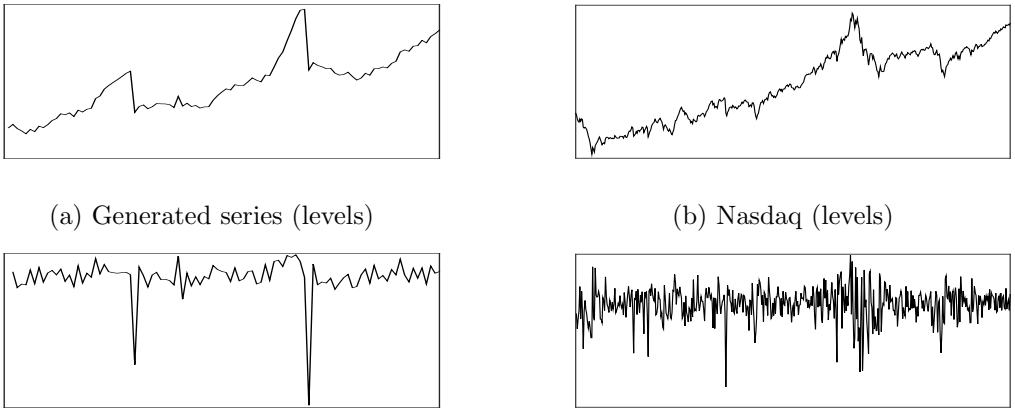
$$y_t = \begin{cases} y_{t-1} + \varepsilon_t, & t = 1, \ldots, \tau_{1e} - 1 \\ (1+\delta)y_{t-1} + \varepsilon_t, & t = \tau_{1e}, \ldots, \tau_{1f} \\ y_1^*, & t = \tau_{1f} + 1 \\ y_{t-1} + \varepsilon_t, & t = \tau_{1f} + 2, \ldots, \tau_{2e} - 1 \\ (1+\delta)y_{t-1} + \varepsilon_t, & t = \tau_{2e}, \ldots, \tau_{2f} \\ y_2^*, & t = \tau_{2f} + 1 \\ y_{t-1} + \varepsilon_t, & t = \tau_{2f} + 2, \ldots, T, \end{cases} \tag{9}$$

where $\delta > 0$ is a parameter to impose explosiveness, y_j^*, $j = 1, 2$ are the values the series takes upon termination of the bubble, and $\varepsilon_t \sim N(0, \sigma^2)$. In Phillips and Yu (2009) and PSY, y_j^* equals $y_{\tau_{je}}$ plus an $O_p(1)$ perturbation. This choice of y_j^* aims to model a series which returns to fundamentals upon the collapse of a bubble. Harvey et al. (2015c) let the series resume a random walk immediately upon the termination of a bubble, i.e. $y_j^* = y_{\tau_{jf}} + \varepsilon_t$, which models a non-collapsing bubble.

Experiments in this paper are conducted by generating models with two bubbles using (9) and the same specifications as Phillips et al. (2015b). Parameter settings are $\sigma = 6.79$, $y_0 = 100$, $T = 100$, and $\delta = 0.06$. Multiple combinations of start points for each of the two bubbles are considered. The first bubble can start at $\tau_{1e} = \{\lfloor 0.20T \rfloor, \lfloor 0.30T \rfloor\}$, and the second bubble starts at $\tau_{2e} = \{\lfloor 0.50T \rfloor, \lfloor 0.60T \rfloor, \lfloor 0.70T \rfloor\}$. The duration of the first bubble, $\tau_{1f} - \tau_{1e}$, can take the values $\lfloor 0.10T \rfloor$ and $\lfloor 0.15T \rfloor$. The duration of the second bubble, $\tau_{2f} - \tau_{2e}$, takes values $\{\lfloor 0.10T \rfloor, \lfloor 0.15T \rfloor, \lfloor 0.20T \rfloor\}$. For each experiment, $5,000$ replications were used. The minimum window, $\lfloor Tr_0 \rfloor$, the for PWY and PSY methods has 12 observations. The rolling window is arbitrarily set at 12 observations, which is equal to r_0. Bubbles were identified using respective finite sample 95% quantiles, obtained from simulations with $5,000$ replications for the $BSDF$ statistic and $20,000$ replications for the DF statistic. For each of the analyses conducted in this section, a DF test regression with constant term included as in equation (5) and with the constant term omitted will be used and their results compared to each other. This comparison will provide some insight into the appropriate choice of DF test regression to use when testing for bubbles.

This DGP, which was developed by Phillips et al. (2011), has been used, with or without slight variations, in almost all subsequent studies in bubble-detection and dating literature, as is the case in this paper. If the model simulated under the DGP provides a reasonable representation of true financial bubbles, any similarities or differences between the testing methods should then be reflected when the tests are conducted on true data. If however this is not the case and the DGP is nothing like the time series for asset prices encountered in practice, then inappropriate

conclusions can be reached. A representative series is illustrated in Figure 2 and compared with actual observed data for the Nasdaq stock index.



(a) Generated series (levels)

(b) Nasdaq (levels)

(c) Generated series (first differences)

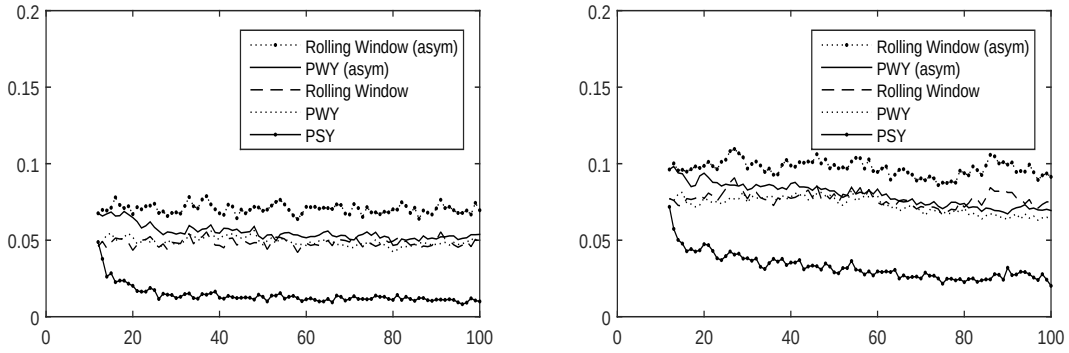
(d) Nasdaq (first differences)

Figure 2: Comparison between first differences of simulated and actual data in the presence of collapsing bubbles

A key characteristic of the simulated data is that the DGP collapses, post bubble, to the fundamentals in a single period. As a result, the first differences of generated data exhibits a single abnormally large downward spike upon the collapse of bubbles, which is not observed in real data. In other words, it is worth remembering that that simulation results should always be approached with a healthy degree of scrutiny, especially when the true DGP of a series is not well-established.

4.2 Size

A common first step in Monte Carlo analysis of any statistical test is to examine its empirical size. The empirical size of each of the three procedures is shown in Figure 3. For the rolling-window and PWY tests, results were obtained for experiments conducted using asymptotic critical values as well as those generated by simulation for relevant window lengths. For the PSY procedure, all critical values were generated for each window size.



(a) $\Delta y_t = \alpha + \phi y_{t-1} + \epsilon_t$ (b) $\Delta y_t = \phi y_{t-1} + \epsilon_t$

Figure 3: Empirical size of rolling-window, PWY and PSY date-stamping procedures under DGP(9) wih $\sigma = 6.79$

Panel (a) illustrates the series of empirical sizes for tests conducted when a constant term is included in the DF test regression. Under this specification, the size of the PSY method is approximately 5% for the smallest sub-sample (12 observations), but decreases as more observations are included. The size of the rolling-window test using asymptotic values is higher than the size under simulated critical values, which is approximately the nominal level of 5%. On the other hand, since the window size of the PWY method increases with each recursion, the empirical size using asymptotic critical values tends towards the nominal value of 5% as the window size increases. Using window-specific critical values, the rolling-window and PWY procedures have empirical sizes which are consistently approximately 5%.

Consider briefly the inclusion of a constant term in the test regression, equation (5). This term affects the DF statistic in the sense that the estimate of ρ is computed after having removed the effect of a possible drift even when the presence of such a drift is infeasible. Under the null hypothesis, this phenomenon does not manifest itself. However it is likely that a sub-sample containing an explosive period would be estimated as having a drift in it, which is considered empirically infeasible (Phillips et al., 2014), therefore altering the value of the DF statistic. From an empirical perspective, Homm and Breitung (2012) have observed that for data with high enough frequency, the effect of a drift, if present, is negligible. In addition, for shorter windows, the drift effect under the null is greatly reduced. Since the rolling window procedure generally considers relatively short windows, ignoring the presence of a drift term should not affect statistical tests in a significant manner.

While on the subject of the test regression, Homm and Breitung (2012) suggest that the constant

term in the test equation may be accounted for by first detrending the data by means of a least squares regression of the series on a constant and linear time trend and using the residuals for the purposes of unit root testing. While detrending a series is common practice for left-tailed DF testing, the situation is slightly more complicated in right-tailed testing for bubbles. Although this method of detrending is valid under the null hypothesis, Phillips et al. (2014) point out that under the alternative, the presence of a deterministic drift component is empirically unrealistic. Therefore, indiscriminate detrending of a series without prior knowledge of whether or not a bubble exists may affect statistical inference. The reason for the inclusion of the constant term in the test regression is that it allows the test to identify possible drift in prices instead of immediately marking them as explosive. In other words, the inclusion of the constant may reduce the chances of rejecting a true null hypothesis.

In the case of using the DF test where a constant is excluded from the test regression, it is important to first note that critical values are generated for a series with 0 as its initial value and $\varepsilon_t \sim N(0,1)$, whereas the DGP has an initial value of $y_0 = 100 \neq 0$. This means that under the null hypothesis, the simulated sample can be re-written as

$$
\begin{aligned}
y_t &= \mu + u_t \\
u_t &= \begin{cases}
u_{t-1} + \varepsilon_t, & t = 1, \ldots, \tau_{1e} - 1 \\
(1+\delta)u_{t-1} + \varepsilon_t, & t = \tau_{1e}, \ldots, \tau_{1f} \\
y_1^*, & t = \tau_{1f} + 1 \\
u_{t-1} + \varepsilon_t, & t = \tau_{1f} + 2, \ldots, \tau_{2e} - 1 \\
(1+\delta)u_{t-1} + \varepsilon_t, & t = \tau_{2e}, \ldots, \tau_{2f} \\
y_2^*, & t = \tau_{2f} + 1 \\
u_{t-1} + \varepsilon_t, & t = \tau_{2f} + 2, \ldots, T,
\end{cases}
\end{aligned}
$$

with $\mu = 100$ and $u_0 = 0$. In this case, it is possible to conduct tests on u_t instead of y_t. The results are very similar to those reported in Panel (a) of Figure 3, in which the size of the rolling-window and PWY tests are approximately 5%, and size of the PSY procedure begins at approximately 5% and decreases as the sample size increases[2]. However, for the same reason that a sample cannot contain a trend under the alternative, it is empirically infeasible to include a mean for a series under the alternative. Since it is not known a priori whether a series contains an explosive period, it is unrealistic in practice to attempt to demean the series before testing. As such, simulations will be conducted on y_t instead of on demeaned u_t.

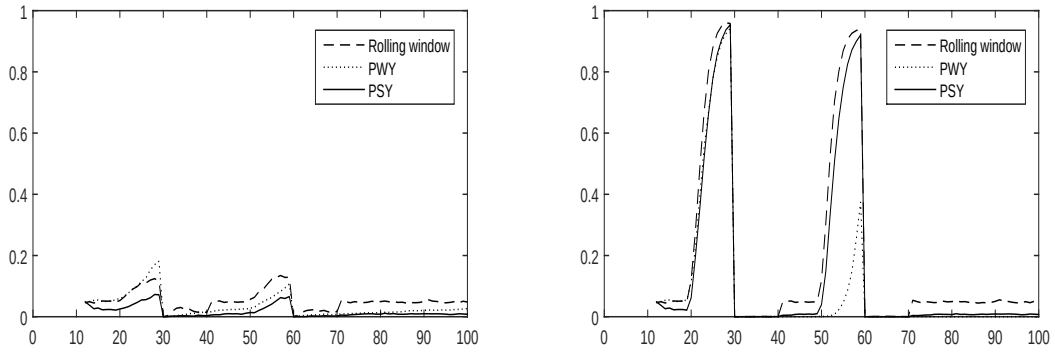
[2]Results for empirical size of tests performed on demeaned series are, for all practical purposes, identical to Panel (a) of Figure (3), and are available upon request.

In Panel (b) of Figure 3, results are obtained using a DF test regression without a constant term. All three tests have empirical sizes which are larger than their respective counterparts using (5), regardless of whether simulated or asymptotic critical values are used. The rolling-window procedure has empirical size of approximately 8% when the critical value for window lengths of 12 is used and 10% when the asymptotic critical value is used. The PWY test with window-specific critical values has empirical size of approximately 8% across the board, whereas the use of the asymptotic critical value leads to empirical size beginning at approximately 10% for the smallest sample and decreasing to approximately 8%. The PSY test has empirical size which begins at around 8% and decreases in a similar way as its counterpart in Panel (a).

Clearly the size properties of the DF tests are slightly adversely affected by the absence of a constant term in the test regression. However, a potentially more important point is that the inclusion of a constant may increase the chances of not rejecting a false null hypothesis. From an economic point of view, the latter could potentially be more harmful. In other words, it is the power of the tests which are more important from the standpoint of the policymaker. The results of this section do serve to emphasise that in computing the power of the tests in the absence of a constant in the test regression, any comparisons can only be drawn on the basis of size adjusted measures of power.

4.3 Power

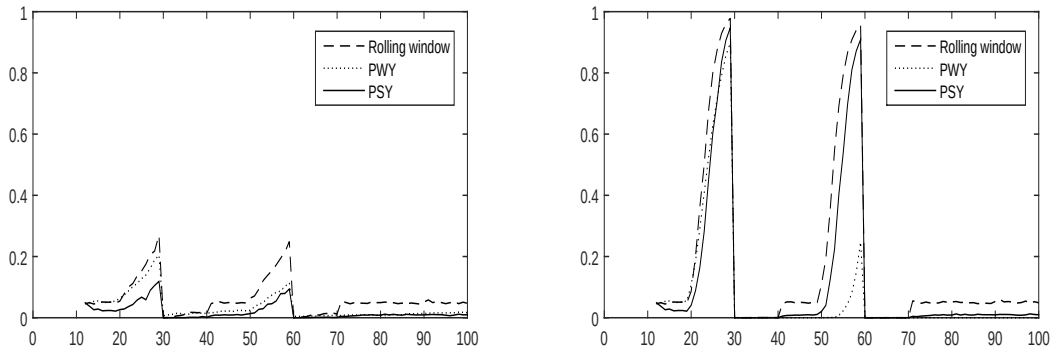
In addition to having low enough empirical size, a good test must have reasonably high power against the alternative. Since this paper seeks to address real-time detection of bubbles, power is displayed for each point in the series at which tests are conducted. Figures 4 and 5 present the empirical power of rolling-window, PWY and PSY date-stamping procedures under DGP (9) with $\sigma = 6.79$, $y_0 = 100$, $T = 100$, $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{2e} = \lfloor 0.50T \rfloor$, $\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$, $\tau_{2f} - \tau_{2e} = \lfloor 0.10T \rfloor$, and for $\delta = \{0.02, 0.10\}$.



(a) $\delta = 0.02$ (b) $\delta = 0.10$

Figure 4: Empirical power of rolling-window, PWY and PSY bubble-dating procedures when a constant term is included in the DF test regression.

It is immediately clear from Figure 4 that for tests which include a constant, the PWY procedure has very low power during the second explosive period for both values of δ. Both the rolling-window and PSY methods appear to perform well, but the rolling-window test has higher power. It is worth noting that the presence of bubbles in a sample appears to impose a lasting reduction in the subsequent size of the recursive and rolling-recursive procedures, whereas the size of the rolling-window test returns immediately to approximately 5% once the window no longer includes observations from within the bubble.



(a) $\delta = 0.02$ (b) $\delta = 0.10$

Figure 5: Size-adjusted empirical power of rolling-window, PWY and PSY bubble-dating procedures when the constant term is excluded from the DF test regression.

Size-adjusted power for the three date-stamping procedures using the DF model without a constant is shown in Panel (a) of Figure (5). From this plot, it can be seen that for a low value of δ not only

does the rolling-window test under the no-constant specification have higher power than the other two procedures, it also has higher power than its counterpart with a constant. On the other hand, for a larger value of δ, it is seen in Panel (b) that although the power of the rolling-window test under the no-constant specification is initially slightly lower than its counterpart with a constant, it eventually catches up and has higher power for later observations within the explosive period. In addition, the rolling-window procedure still outperforms the PWY and PSY methods under this specification.

4.4 Detection Rates

From a practical perspective, it is interesting to consider not only the power of these tests against the alternative hypothesis, but also how long the delay in detection is. In considering this, it is useful to record *detection rates* for the various testing algorithms. Following Phillips et al. (2015b) a successful detection is recorded if the test correctly identifies the origination of a bubble sometime between its actual start and end date. Table 1 reports the detection rate, empirical mean, and standard deviation (in parentheses) of estimated origination dates for PWY, PSY and rolling-window tests with a constant for $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$, and for all combinations of second bubble parameters $\tau_{2e} = \{\lfloor 0.50T \rfloor, \lfloor 0.60T \rfloor, \lfloor 0.70T \rfloor\}$ and $\tau_{2f} - \tau_{2e} = \{\lfloor 0.10T \rfloor, \lfloor 0.15T \rfloor, \lfloor 0.20T \rfloor\}$. Under these parameter combinations, detection rates, means, and standard deviations of the estimated origination of the *first* bubble are virtually identical for the PWY and PSY tests. The difference between detection rates is only 1%, and the mean estimated origination dates are identical. On the other hand, the rolling-window procedure yields a detection rate which is 11% higher than the PWY method, with mean estimated origination which is closer to the true date by $\lfloor 0.01T \rfloor$.

For the second bubble, the PSY procedure provides a clear improvement over the PWY test. The PSY procedure has detection rates that are between 11% and 41% higher than the PWY method. Mean origination points estimated using the PSY procedure are closer to true origination points by between 0.01 and 0.04 of the sample size. This result clearly demonstrates the superiority of the rolling recursive method over the recursive procedure, a result established by Phillips et al. (2015b). However, the detection rate for the second bubble by the rolling-window regressions are consistently higher than PSY detection rates by between 4% and 15%. The origination of the explosive period is also estimated slightly more accurately by up to 0.02 of the sample size using rolling windows. Similar results are obtained for $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$ and for $\tau_{1e} = \lfloor 0.30T \rfloor$, $\tau_{1f} - \tau_{1e} = \{\lfloor 0.10T \rfloor, \lfloor 0.20T \rfloor\}$, and are available upon request.

As before, each of the three procedures for the DF test regression without a constant are also examined. Detection rates, empirical means and standard deviations of origination dates for each

Table 1: Detection rate and estimates of the origination dates under DGP with two bubbles. Parameters are set to $y_0 = 100$, $\sigma = 6.79$, $\delta = 0.06$, $T = 100$, $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{2e} = \{\lfloor 0.50T \rfloor, \lfloor 0.60T \rfloor, \lfloor 0.70T \rfloor\}$, $\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$. Figures in parentheses are standard deviations. DF regressions are conducted for model $\Delta y_t = \alpha + \phi y_{t-1} + \epsilon_t$.

$\tau_{2f} - \tau_{2e} =$	$\lfloor 0.10T \rfloor$			$\lfloor 0.15T \rfloor$			$\lfloor 0.20T \rfloor$		
	PWY	PSY	Roll	PWY	PSY	Roll	PWY	PSY	Roll
Detection rate (1)	0.76	0.75	0.87	0.76	0.75	0.87	0.76	0.75	0.87
$r_{1e} = 0.20$	0.25	0.25	0.24	0.25	0.25	0.24	0.25	0.25	0.24
	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)
Detection rate (2)	0.30	0.71	0.86	0.64	0.87	0.93	0.82	0.93	0.97
$r_{2e} = 0.50$	0.57	0.55	0.54	0.59	0.56	0.54	0.61	0.57	0.55
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Detection rate (2)	0.31	0.69	0.84	0.62	0.84	0.93	0.80	0.91	0.96
$r_{2e} = 0.60$	0.66	0.65	0.64	0.69	0.66	0.65	0.71	0.67	0.65
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Detection rate (2)	0.31	0.68	0.83	0.61	0.83	0.91	0.78	0.90	0.95
$r_{2e} = 0.70$	0.76	0.75	0.74	0.79	0.76	0.74	0.81	0.77	0.75
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)

Note: Calculations are based on $5,000$ replications. $\lfloor Tr_0 \rfloor$ and w have 12 observations.

of the three procedures without a constant are reported in Table 2. It is seen, once again, that the rolling-window test is superior to the others. There are only two instances in which the rolling-window test does not have the highest detection rate, and in any event the difference in performance in these instances is negligible. For all parameter settings, mean origination dates estimated by the rolling-window procedure are closer to actual origination dates than the others.

Overall it is observed that the rolling-window procedure yields higher detection rates than the others and the increased efficacy of the rolling-window test over others is more pronounced when bubbles have shorter durations. This conclusion is further reinforced by the fact that mean estimates of origination dates under the rolling-recursive procedure deviate more from true dates than rolling-window estimations when bubbles are longer.

4.5 Asymmetric Loss

Thus far, analysis of the performance of these three methods has assumed that incorrect rejection of a true null hypothesis and failure to reject a false null hypothesis are penalised equally. In reality, failure to identify the presence of a bubble in a timely fashion may lead to stock market crashes and financial crises akin to the sub-prime crisis and the aftermath of the Dot-Com bubble, and is potentially far more costly than false rejection and a brief period of trepidation. In order to account for this, an asymmetric loss function is used to evaluate these three procedures under different DGP parameter settings. This function attaches a higher cost to the case where a bubble occurs but is not detected than it does to scenarios where the null hypothesis is rejected even though there is no bubble. The asymmetric loss function takes the form

$$Loss = \frac{1}{T - \lfloor Tr_0 \rfloor + 1} \sum_{t=\lfloor Tr_0 \rfloor}^{T} \left(x_t(1+\kappa) + (1-x_t)(1-\kappa)\right) |\hat{x}_t - x_t|, \tag{10}$$

where $x_t = 1$ if a bubble occurs at t or 0 otherwise and $\hat{x}_t = 1$ if a bubble is detected at time t or 0 otherwise. A higher weighting, $(1+\kappa)$, is assigned to the penalty suffered if the tests fail to detect a bubble that exists at time t, and a lower weighting, $(1-\kappa)$ is assigned if false detection occurs. The higher the value of the function, the greater the cost is over the sample.

Table 3 reports the loss associated with each date-stamping method under a range of different parameter settings using regression equation (5). These results reveal that in the presence of two bubbles, the PWY test always has higher loss associated with it than do the other two methods. When the first bubble is shorter ($\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$), the rolling-window test always incurs a lower penalty than the PSY test. When $\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$, the PSY test *almost* always suffers a lower penalty than the rolling-window test. On the other hand, when loss is computed for date-stamping

Table 2: Detection rate and estimates of the origination dates under DGP with two bubbles. Parameters are set to $y_0 = 100$, $\sigma = 6.79$, $\delta = 0.06$, $T = 100$, $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{2e} = \{\lfloor 0.50T \rfloor, \lfloor 0.60T \rfloor, \lfloor 0.70T \rfloor\}$, $\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$. Figures in parentheses are standard deviations. DF regressions are conducted for model $\Delta y_t = \phi y_{t-1} + \epsilon_t$.

$\tau_{2f} - \tau_{2e} =$	$\lfloor 0.10T \rfloor$			$\lfloor 0.15T \rfloor$			$\lfloor 0.20T \rfloor$		
	PWY	PSY	Roll	PWY	PSY	Roll	PWY	PSY	Roll
Detection rate (1)	0.63	0.75	0.83	0.63	0.75	0.83	0.63	0.75	0.83
$r_{1e} = 0.20$	0.25	0.25	0.24	0.25	0.25	0.24	0.25	0.25	0.24
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
Detection rate (2)	0.28	0.74	0.82	0.59	0.88	0.89	0.79	0.93	0.91
$r_{2e} = 0.50$	0.57	0.55	0.54	0.59	0.56	0.55	0.61	0.57	0.55
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Detection rate (2)	0.28	0.72	0.81	0.57	0.86	0.87	0.76	0.91	0.90
$r_{2e} = 0.60$	0.66	0.65	0.64	0.69	0.66	0.65	0.71	0.67	0.65
	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)
Detection rate (2)	0.28	0.70	0.78	0.55	0.84	0.86	0.73	0.89	0.89
$r_{2e} = 0.70$	0.76	0.75	0.74	0.79	0.76	0.75	0.81	0.77	0.75
	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.04)

Note: Calculations are based on $5,000$ replications. $\lfloor Tr_0 \rfloor$ and w have 12 observations.

Table 3: Mean loss from incorrect detection computed using (10) under DGP with two bubbles. Parameters are set to $y_0 = 100$, $\sigma = 6.79$, $\delta = 0.06$, $T = 100$, $\kappa = 0.5$. Figures in parentheses are standard deviations.
All DF regressions are conducted for model $\Delta y_t = \alpha + \phi y_{t-1} + \epsilon_t$.

$\tau_{2f} - \tau_{2e} =$	$\lfloor 0.10T \rfloor$			$\lfloor 0.15T \rfloor$			$\lfloor 0.20T \rfloor$		
	PWY	PSY	Roll	PWY	PSY	Roll	PWY	PSY	Roll
	Panel A ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.50T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2509	0.2236	0.1974	0.2921	0.2428	0.2279	0.3123	0.2526	0.2490
	(0.0605)	(0.0787)	(0.0764)	(0.0802)	(0.0998)	(0.0922)	(0.1018)	(0.1160)	(0.1078)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2835	0.2543	0.2502	0.3646	0.2758	0.2801	0.4349	0.2855	0.3005
	(0.0822)	(0.1058)	(0.1001)	(0.0794)	(0.1255)	(0.1143)	(0.0785)	(0.1401)	(0.1284)
	Panel B ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.60T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2492	0.2239	0.1988	0.2911	0.2448	0.2291	0.3132	0.2560	0.2511
	(0.0612)	(0.0792)	(0.0772)	(0.0822)	(0.1014)	(0.0941)	(0.1051)	(0.1191)	(0.1114)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2832	0.2485	0.2497	0.3628	0.2687	0.2797	0.4298	0.2789	0.3008
	(0.0816)	(0.1080)	(0.1012)	(0.0788)	(0.1277)	(0.1163)	(0.0804)	(0.1429)	(0.1316)
	Panel C ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.70T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2484	0.2242	0.1995	0.2904	0.2460	0.2305	0.3138	0.2590	0.2529
	(0.0617)	(0.0793)	(0.0778)	(0.0835)	(0.1022)	(0.0954)	(0.1073)	(0.1215)	(0.1142)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2830	0.2476	0.2505	0.3615	0.2687	0.2812	0.4251	0.2805	0.3030
	(0.0816)	(0.1083)	(0.1022)	(0.0791)	(0.1286)	(0.1181)	(0.0830)	(0.1453)	(0.1350)

Note: Calculations are based on $5,000$ replications. $\lfloor Tr_0 \rfloor$ and w have 12 observations.

procedures conducted using a test regression without a constant term, as reported in Table 4, the rolling-window test results in lower loss than the other two procedures for all parameter settings. In addition to that, comparisons between Table 3 and Table 4 reveal that for the two methods with high power against the second bubble (PSY and rolling-window test), the use of a DF test regression without a constant term results in lower loss than the use of test equation (5).

The choice of $\kappa = 0.5$ is arbitrary, and penalises incorrect non-detection three times as much as incorrect detection. The asymmetric loss simulation experiments were also conducted for $\kappa = 0.6$, with outcomes which mirrored these.

4.6 Window Length

The results of all the simulation exercises thus far suggest that the rolling-window procedure without a constant in the regression equation is the preferred method for bubble detection. An important point to note is that the number of observations used in each rolling window have been arbitrarily chosen. The question of optimal window-length selection is of course of great importance. If the rolling window contains too many observations the procedure will face a delay in identifying the origination of the bubble. If the window contains too few observations, the overall trend might be ignored by the procedure, leading to meaningless inference resulting from noise instead of from changes in the DGP. Pesaran and Timmerman (2007) and Inoue et al. (2014), among others, propose methods to select an optimal window length for rolling-window regressions in context of forecasting problems. However, there is a clear distinction between the context in which their procedures are conducted, namely forecasting performance, and the objectives in date stamping bubbles, namely, identifying the exact point of the break.

It is tempting to relate the choice of window length to sample size in a similar way to how the minimum sub-samples of the PWY and PSY procedures are dictated by r_0, and are thus related to sample size. On the other hand, in this particular problem, the length of an existing bubble will not change as more observations are included. The data are only informative if in fact a new bubble is present.

As an initial foray into the problem of the optimal choice of window-length for rolling-window

Table 4: Mean loss from incorrect detection computed using (10) under DGP with two bubbles. Parameters are set to $y_0 = 100$, $\sigma = 6.79$, $\delta = 0.06$, $T = 100$, $\kappa = 0.5$. Figures in parentheses are standard deviations.
All DF regressions are conducted for model $\Delta y_t = \phi y_{t-1} + \epsilon_t$.

$\tau_{2f} - \tau_{2e} =$	$\lfloor 0.10T \rfloor$			$\lfloor 0.15T \rfloor$			$\lfloor 0.20T \rfloor$		
	PWY	PSY	Roll	PWY	PSY	Roll	PWY	PSY	Roll
	Panel A ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.50T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2553	0.2142	0.1798	0.3002	0.2265	0.1852	0.3240	0.2330	0.1876
	(0.0678)	(0.0789)	(0.0805)	(0.0908)	(0.0957)	(0.0926)	(0.1142)	(0.1093)	(0.1022)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2946	0.2314	0.1852	0.3767	0.2425	0.1892	0.4493	0.2475	0.1960
	(0.0930)	(0.0981)	(0.0932)	(0.0912)	(0.1130)	(0.1036)	(0.0909)	(0.1239)	(0.1113)
	Panel B ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.60T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2535	0.2152	0.1813	0.2988	0.2296	0.1881	0.3246	0.2378	0.1917
	(0.0686)	(0.0787)	(0.0796)	(0.0919)	(0.0976)	(0.0935)	(0.1165)	(0.1133)	(0.1053)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2944	0.2284	0.1838	0.3753	0.2413	0.1892	0.4453	0.2480	0.1917
	(0.0926)	(0.0991)	(0.0930)	(0.0909)	(0.1158)	(0.1048)	(0.0925)	(0.1288)	(0.1147)
	Panel C ($\boldsymbol{\tau_{1e} = \lfloor 0.20T \rfloor}$, $\boldsymbol{\tau_{2e} = \lfloor 0.70T \rfloor}$):								
$\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$	0.2526	0.2164	0.1826	0.2984	0.2322	0.1904	0.3260	0.2419	0.1950
	(0.0692)	(0.0779)	(0.0801)	(0.0930)	(0.0978)	(0.0945)	(0.1186)	(0.1151)	(0.1081)
$\tau_{1f} - \tau_{1e} = \lfloor 0.20T \rfloor$	0.2943	0.2291	0.1853	0.3743	0.2438	0.1919	0.4421	0.2522	0.1955
	(0.0926)	(0.0992)	(0.0936)	(0.0912)	(0.1168)	(0.1063)	(0.0943)	(0.1314)	(0.1178)

Note: Calculations are based on 5,000 replications. $\lfloor Tr_0 \rfloor$ and w have 12 observations.

bubble-detection tests, the following DGP is simulated

$$y_t = \begin{cases} y_{t-1} + \varepsilon_t, & t = 1, \ldots, \tau_{1e} - 1 \\ (1+\delta)y_{t-1} + \varepsilon_t, & t = \tau_{1e}, \ldots, \tau_{1f} \\ y_1^*, & t = \tau_{1f} + 1 \\ y_{t-1} + \varepsilon_t, & t = \tau_{1f} + 2, \ldots, \tau_{2e} - 1 \\ (1+\delta)y_{t-1} + \varepsilon_t, & t = \tau_{2e}, \ldots, \tau_{2f} \\ y_2^*, & t = \tau_{2f} + 1 \\ y_{t-1} + \varepsilon_t, & t = \tau_{2f} + 2, \ldots, T + T_1, \end{cases}$$

in which τ_{je} and τ_{jf} defined as before for $j = 1, 2$ and for $T = 100$. The change in sample size is achieved by varying T_1. In this way, the bubbles are not affected by increasing the number of observations. The mean loss for $\tau_{1e} = \lfloor 0.20T \rfloor$, $\tau_{2e} = \{\lfloor 0.5T \rfloor, \lfloor 0.6T \rfloor, \lfloor 0.7T \rfloor\}$, $\tau_{1f} - \tau_{1e} = \{\lfloor 0.10T \rfloor, \lfloor 0.20T \rfloor\}$, $\tau_{2f} - \tau_{2e} = \{\lfloor 0.10T \rfloor, \lfloor 0.15T \rfloor \lfloor 0.20T \rfloor\}$, $\delta = 0.06$ and $\kappa = 0.5$ is computed using (10) for window lengths of $12, 13, \ldots, 22$. Setting $T_1 = \{0, 50, 100, 150, 200, 250\}$ lets the length of the sample vary. The window length which leads to the lowest loss across all values of T_1 is consistent for every combination of parameters. For example, the mean loss for $\tau_{2e} = \lfloor 0.5T \rfloor$, $\tau_{1f} - \tau_{1e} = \lfloor 0.10T \rfloor$, $\tau_{2f} - \tau_{2e} = \lfloor 0.10T \rfloor$ is lowest for rolling windows with 12 observations regardless of T_1. This set of results suggest that for a given financial time series, there should be an optimal window length which is fixed and does not grow with sample size.

Of course support for the rolling window procedure in this section thus far is based on the very specific case of $r_0 = w = 12$. Following this line of thought, it is interesting to consider cases where minimum window lengths are the same across all three methods, but for different values of $r_0 = w$. In other words, is the rolling window method still superior to the other two methods if $r_0 = w \neq 12$? All of the simulation experiments in this section were repeated for $r_0 \in \{15, 17, 20\}$ and $w = r_0$. Results of the experiments in terms of size, power, detection rates and loss function still lead to the rolling window approach being selected as the best for every value of $r_0 = w$. Thus, while the optimal window length for a given sample may be unknown, the rolling window test is still preferred to recursive methods as long as the rolling window is of the same length as the minimum window of the recursive approaches. The choice of minimum window length is one that must be made regardless of which of these three methods is used. A corollary of this fact is that if a minimum window length must be selected, it may as well be used with the method which is most likely to provide good results. The results from this section show that the rolling window approach is the best choice.

Overall, the simulation results suggest that in the event of a single bubble, the origination dates

estimated by all three methods should be similar. Should a second bubble occur within a sample, the rolling-window procedure should detect its origination earlier than the PSY procedure, which in turn should detect it sooner than the PWY method, if the PWY method detects it at all. This relationship between estimated bubble origination dates is what one should expect to find in real data, assuming (9) models financial prices to a satisfactory degree of accuracy.

5 Empirical application

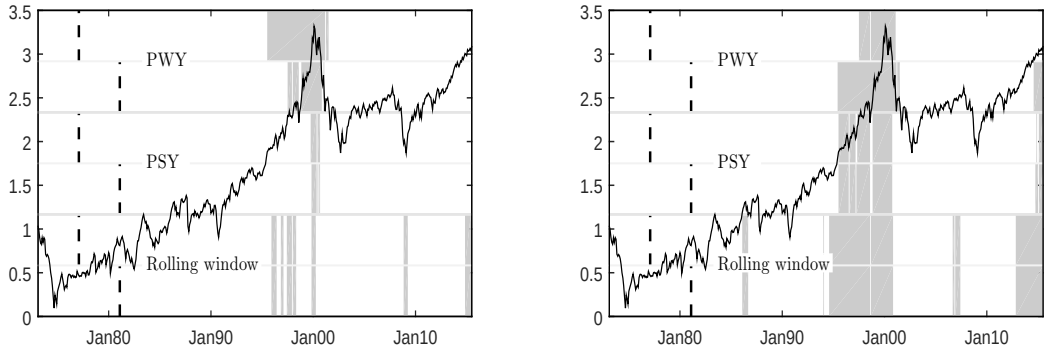
This section applies the alternative bubble detection algorithms to two reasonably well known time series often used in the bubble detection literature. The first series is monthly nominal Nasdaq composite price index data from February 1973 to July 2015 (510 observations) which are obtained from `finance.yahoo.com`. This nominal price index is normalised using the Consumer Price Index (CPI) obtained from the Federal Reserve Economic Data (FRED), which is maintained by the Federal Reserve bank of St. Louis. The second series is the All-Transactions House Price Index for the United States which is used to analyse house prices. Quarterly observations for the period January 1975 to April 2014 (162 observations) are obtained from FRED.

The usefulness of these two series from the point of view of detection algorithms is that both are know to contain at least one bubble. The Nasdaq sample contains what has come to be known as the Dot-com bubble. This bubble occurred in the mid- to late-1990s, and collapsed in the early 2000s. The sample of U.S. house prices contains multiple bubbles, the largest and most significant of which occurring in the mid- to late- 2000s. It is generally believed that the series contains two other bubbles, which peaked in 1979 and 1989, respectively (Gjerstad and Smith, 2009). In order to address the consideration raised in Section 3 regarding the influence of the first observation in the sample on each of the three date-stamping procedures, all three methods will be conducted on full samples as well as on samples which omit some initial observations.

5.1 Nasdaq Composite Index

Based on the detection rates reported in Section 4 all three methods should identify the origination of the Dot-Com bubble at approximately the same date, assuming the Dot-Com bubble is the first time the series exhibits explosivenes. Any subsequent bubbles should be detected first by the rolling-window procedure followed by the PSY test and finally (if at all) by the PWY method. The minimum window size for PWY and PSY has 49 observations, which is the same minimum window used in PWY. The window size for the rolling window algorithm is therefore arbitrarily set at 49

observations to facilitate comparison. All three methods are conducted on the full sample as well as on a sub-sample with the first 48 observations omitted.



(a) DF model with constant

(b) DF model without constant

Figure 6: Monthly logged real Nasdaq prices from February 1973 to July 2015. $\lfloor Tr_0 \rfloor$ and w have 49 observations.

The series of logged real Nasdaq prices is plotted in Figure 6. Panel (a) reports results for tests conducted with a constant in the regression equation and Panel (b) contains tests conducted without the constant. Each panel is divided vertically into three pairs of bands, or six bands in total. The top two bands correspond to the PWY procedure, the middle two to the PSY procedure and the bottom two to the rolling-window method. The higher band of each pair corresponds to full-sample analysis and the lower corresponds to the sub-sample. The shaded regions in the figure represent periods when explosiveness is detected by the respective date-stamping methods. Test statistics are evaluated for all observations to the right of vertical dashed lines in each band.

Upon inspection of Panel (a) it is found that the order of detection does not match expectations. Although all three methods detect the Dot-Com bubble at approximately the same time (as the simulation results suggest they should), the start dates are all very different. In addition, the PSY procedure is very late to pick up explosiveness and the rolling-window method detects bursts of explosiveness instead of one long bubble. By way of contrast, Panel (b) reveals that all the tests conducted without a constant yield what look like more believable results. For full-sample analysis, all three tests detect the Dot-Com bubble, with the rolling-window method leading the way, followed by the PSY approach and finally the PWY procedure.

One apparent conclusion to be drawn from these results is that the rolling-window test equations, with or without a constant term, appear to be unaffected by changes in the starting date of the sample. This follows from the fact that in both panels of Figure 6 the results in the upper and lower

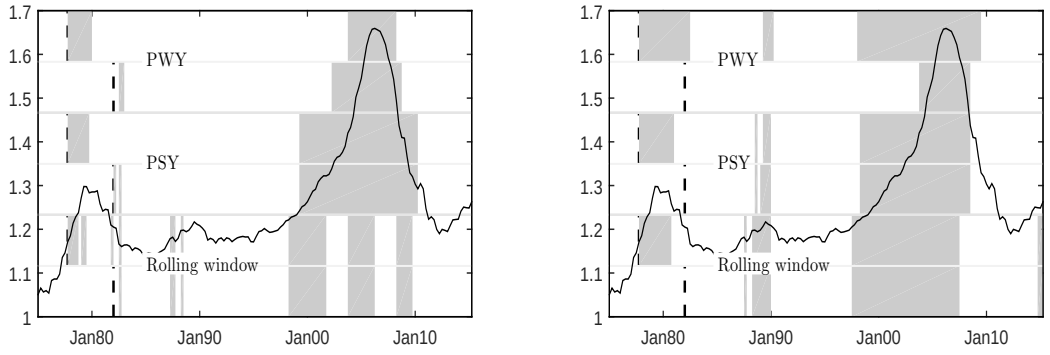
tiers of the rolling window algorithm are identical. Changing the sample has some effect on the results of the PSY procedure, but these differences are negligible. The PWY procedure is affected the most when the first 48 observations are omitted.

Based on inference formed around the Dot-Com bubble, it would appear that the rolling-window test conducted using a DF test equation without the constant term is the best choice. Using this procedure, the Dot-Com bubble is detected earliest and also in a meaningful way as a continuous bubble instead of short, sporadic bursts. Based on these results, the Dot-Com bubble originated in August 1994 and ended in November 2000.

Two final comments are in order. The rolling window test identifies explosive behaviour in the Nasdaq which originates in September 2006 and ends in July 2007, coinciding with the period immediately preceding the global financial crisis. The second observation is that there are now clear signs of a bubble emerging in the Nasdaq in 2015. For the test equation containing the constant term, it is only the rolling window test that picks up an emerging bubble, but when the test equation is specified without a constant there is clear evidence of a bubble in the test results from all the procedures. In other words at the time of writing, these results appear to substantiate claims currently being made by the media that the United States stock market is in a bubble.

5.2 United States House Prices

United States house prices in this sample period are often considered to contain bubbles with peaks in 1979, 1989 and 2006. This time series therefore presents an ideal setting in which to compare the accuracy of the three tests in the presence of multiple (known) bubbles. Minimum windows for PWY and PSY procedures and the rolling window have 12 observations (instead of 49), since this sample has a lower frequency than the Nasdaq sample. All tests are conducted on the full sample as well as on a sub-sample with the first 17 observations omitted. In this example, the omission of these early observations takes on extra significance because in so doing the testing sample begins at the peak of the first bubble instead of before it. The origination of the first bubble in the series occurs too early in the sample to be tested, so



(a) DF model with constant

(b) DF model without constant

Figure 7: Quarterly U.S. house price-to-rent ratio from January 1975 to April 2015. $\lfloor Tr_0 \rfloor$ and w have 12 observations.

Figure 7 presents the results of all three tests conducted on U.S. house price-to-rent ratios. Both panels are set out in the same way as those in Figure 6. Once again, changing the starting point of the sample changes the results of the PWY test quite significantly, the results of the PSY test very slightly, and the results of the rolling-window test not at all.

In Panel (a), the rolling-window test is the only one to detect all three purported bubbles. However the largest and most recent one seems to be split into three explosive periods under the rolling-window test, whereas the PWY and PSY procedures correctly identify it as a single bubble. Even so, the PWY procedure only begins to identify the bubble when it is close to its peak regardless of sample choice. An interesting observation is that the period around the middle of 1982 is shaded for four out of the six bands. This period corresponds to the collapse instead of growth of the first bubble. This observation emphasises the point that in the presence of a bubble, the inclusion of a constant is empirically unrealistic (Phillips et al., 2014).

The results of the tests in Panel (b) are closer to what one might expect, from the simulation results for detection rates. For the full sample, all three procedures detect all three bubbles, with the rolling-window approach being the first to identify the second and third bubbles. In addition, there are no instances of collapses being identified as explosiveness. PSY results for the sub-sample are almost identical to full-sample results, but PWY results vary depending on the start of the sample. For larger initial values, the PWY test faces a larger delay in bubble-detection.

Two conclusions regarding testing methods can be drawn from tests conducted on United States house price-to-rent ratios, both of which concur with conclusions from analysis of the Nasdaq. First, conducting tests using DF equations without a constant is preferable to tests which include a

constant drift term. Secondly, the rolling-window procedure has certain advantages over the other procedures in that it detects bubbles earlier than the other methods and is unaffected by choice of sample. In addition, the rolling-window approach is much simpler and less computationally costly than the PSY method, since it conducts much fewer DF tests. Using the rolling-window approach for the test equation without a constant, the second bubble is detected to have started in the third quarter of 1987, the bubble preceding the sub-prime crisis started in the third quarter of 1997, and house prices are currently in a bubble, which began in the final quarter of 2014.

6 Conclusion

This paper has scrutinised a number of bubble detection and date-stamping methods that have been proposed in the literature both under simulation and in empirical applications. These procedures all involve repeated testing of the null hypothesis of non-stationarity against the alternative hypothesis of mildly explosive behaviour using right-tailed Dickey-Fuller unit root tests. The specification of the Dickey-Fuller test regression for each of these three models was also considered.

The balance of the evidence presented in this paper suggests that the rolling-window test for a test equation without a constant or trend component performs best. The test is the most responsive to explosiveness, results in the lowest loss as computed using an asymmetric loss function, is independent of the sample starts, and is the quickest and easiest to implement. The test also provides reasonable and economically viable estimates of the periods of explosive growth in the samples used in the paper. Specifically, the growth of the Dot-Com bubble in the Nasdaq is estimated to have spanned from August 1994 to November 2000, and the housing bubble which preceded the sub-prime crisis is estimated to have grown from the third quarter of 1997 to the third quarter of 2007.

References

Bhargava, A. (1986). On the theory of testing for unit roots in observed time series. *The Review of Economic Studies*, **53**, 369–384.

Busetti, F., and Taylor, A. M. R. (2004). Tests of stationarity against a change in persistence. *Journal of Econometrics*, **123**, 33–66.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. C. (1997). *The Econometrics of Financial Markets*. Princeton, N.J.: Princeton University Press.

Clark, T. E., and McCracken, M. W. (2009). Improving forecast accuracy by combining rateecursive and rolling forecasts. *International Economic Review*, **50**, 363–395.

Cuñado, J., Gil-Alana, L. A., and De Gracia, F. P. (2005). A test for rational bubbles in the NASDAQ stock index: a fractionally integrated approach. *Journal of Banking & Finance*, **29**, 2633–2654.

Diba, B. T., and Grossman, H. I. (1988). Explosive rational bubbles in stock prices? *The American Economic Review*, **78**, 520–530.

Dickey, D. A., and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, **74**, 427–431.

Evans, G.W. (1991). Pitfalls in testing for explosive bubbles in asset prices. *The American Economic Review*, **81**, 922–930.

Frömmel, M., and Kruse, R. (2012). Testing for a rational bubble under long memory. *Quantitative Finance*, **12**, 1723–1732.

Gjerstad, S., and Smith, V. L. (2009). Monetary policy, credit extension, and housing bubbles: 2008 and 1929. *Critical Review*, **21**, 269–300.

Gutierrez, L. (2013). Speculative bubbles in agricultural commodity markets. *European Review of Agricultural Economics*, **40**, 217–238.

Harvey, D. I., Leybourne, S. J., and Sollis, R. (2015a). Improving the accuracy of asset price bubble start and end date estimators. *Unpublished Manuscript*.

Harvey, D. I., Leybourne, S. J., and Sollis, R. (2015b). Recursive right-tailed unit root tests for an explosive asset price bubble. *Journal of Financial Econometrics*, **13**, 166–187.

Harvey, D. I., Leybourne, S. J., Sollis, R., and Taylor, A. M. R. (2015c). Tests for explosive financial bubbles in the presence of non-stationary volatility. *Journal of Empirical Finance*. forthcoming.

Homm, U., and Breitung, J. (2012). Testing for speculative bubbles in stock markets: A comparison of alternative methods. *Journal of Financial Econometrics*, **10**, 198–231.

Inoue, A., Jin, L., and Rossi, B. (2014). Window selection for out-of-sample forecasting with time-varying parameters. *Unpublished Manuscript*.

Kim, J. (2000). Detection of change in persistence of a linear time series. *Journal of Econometrics*, **95**, 97–116.

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

Pesaran, M. H., and Timmermann, A. (2007). Selection of estimation window in the presence of breaks. *Journal of Econometrics*, **137**, 134–161.

Phillips, P. C. B., and Yu, J. (2009). Limit theory for dating the origination and collapse of mildly explosive periods in time series data. *Unpublished Manuscript*.

Phillips, P. C. B., and Yu, J. (2011). Dating the timeline of financial bubbles during the subprime crisis. *Quantitative Economics*, **2**, 455–491.

Phillips, P. C. B., Wu, Y., and Yu, J. (2011). Explosive behaviour in the 1990s NASDAQ: When did exuberance escalate asset values? *International Economic Review*, **52**, 201–226.

Phillips, P. C. B., Shi, S., and Yu, J. (2014). Specification sensitivity in right-tailed unit root testing for explosive behaviour. *Oxford Bulletin of Economics and Statistics*, **76**, 315–333.

Phillips, P. C. B., Shi, S., and Yu, J. (2015a). Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. *International Economic Review*, forthcoming.

Phillips, P. C. B., Shi, S., and Yu, J. (2015b). Testing for multiple bubbles: Limit theory of real time detectors. *International Economic Review*, forthcoming.

Said, S. E., and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, **71**, 599–607.

Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review*, **71**, 421–436.

West, K. D. (1987). A specification test for speculative bubbles. *The Quarterly Journal of Economics*, **102**, 553–580.