Journal of Empirical Finance 58 (2020) 207–225

Contents lists available at ScienceDirect

Journal of Empirical Finance

journal homepage: www.elsevier.com/locate/jempfin

Testing for explosive bubbles in the presence of autocorrelated innovations☆

Thomas Quistgaard Pedersen, Erik Christian Montes Schütte *

Department of Economics and Business Economics and CREATES, Aarhus University, Denmark

ARTICLE INFO

JEL classification:
C58
G12

Keywords:
Right-tailed unit root tests
SADF
GSADF
Size and power properties
Sieve bootstrap
International housing market

ABSTRACT

We analyze an empirically important issue with recursive right-tailed unit root tests for bubbles in asset prices. First, we show that serially correlated innovations, which is a feature that is present in most financial series used to test for bubbles, can lead to severe size distortions when using either fixed or automatic (based on information criteria) lag-length selection in the auxiliary regressions underlying the test. Second, we propose a sieve-bootstrap version of these tests and show that this results in tests which control size well across a number of simulation designs both with and without highly autocorrelated innovations. We also find that these improvements in size come at a relatively low cost for the power of the tests. Finally, we apply the bootstrap tests on the housing market of OECD countries, and generally find much weaker evidence of housing bubbles compared to existing evidence.

1. Introduction

Since the surge and subsequent collapse in stock prices around the turn of the century and in house prices a few years later, research on speculative bubbles in financial and housing markets has received renewed interest. In recent years we have seen many papers suggesting new methods to detect the presence of speculative bubbles; for example, Phillips et al. (2011, 2015), Homm and Breitung (2012), Engsted and Nielsen (2012), and Harvey et al. (2016). These test procedures have been widely used in empirical research on bubbles in many different markets, including stock markets, housing markets, commodity markets, and even the art market.[1] Especially, the recursive right-tailed unit root test procedures developed by Phillips et al. (2011, 2015), also called the SADF and GSADF test, respectively, play an important role in the literature on bubble detection. These tests are motivated by the seminal paper by Diba and Grossman (1988), who first suggested testing the null hypothesis that a given time series follows a random walk process against the explosive and not the stationary alternative. In response to the critique by Evans (1991) that such a test procedure has very low power in detecting partially collapsing bubbles, Phillips et al. (2011) suggest a test procedure (SADF) that entails performing a series of unit root tests using various subsamples of the data. Phillips et al. (2015) show that this test and a generalized version of the test (GSADF) greatly improve power in detecting partially collapsing bubbles.

Since the bubble tests by Phillips et al. (2011, 2015) in essence are sequences of unit root tests, they are naturally also subject to the pitfalls associated with this type of test. Thus, a critical assumption behind the recursive right-tailed unit root tests is that

☆ We thank Rossen Valkanov (the editor), an anonymous associate editor, two anonymous reviewers, Jörg Breitung, Peter Boswijk, Bent Jesper Christensen, Tom Engsted, Niels Haldrup, Robert Taylor, Anders Rahbek and participants at a CREATES seminar and the 10th International Conference in Computational and Financial Econometrics for valuable comments. This research is supported by CREATES, Center for Research in Econometric Analysis of Time Series (DNRF78), founded by the Danish National Research Foundations, the Danish Council of Independent Research (DFF 4003-00022) and CONACYT, National Council for Science and Technology of Mexico.

* Corresponding author.
E-mail addresses: tqpedersen@econ.au.dk (T.Q. Pedersen), christianms@econ.au.dk (E.C.M. Schütte).

[1] Examples include Phillips et al. (2011, 2015), Homm and Breitung (2012), Engsted and Nielsen (2012), Kivedal (2013), Pavlidis et al. (2015), Harvey et al. (2015), Engsted et al. (2016), Shi et al. (2016), Figuerola-Ferretti and McCrorie (2016), and Kraussl et al. (2016).

https://doi.org/10.1016/j.jempfin.2020.06.002
Received 19 March 2019; Received in revised form 25 May 2020; Accepted 15 June 2020
Available online 17 June 2020

innovations to the relevant time series are homoskedastic and serially uncorrelated under the null hypothesis. In a simulation study, Phillips et al. (2015) explore the properties of the SADF and GSADF tests in the presence of time-varying but stationary volatility and generally find that this does not lead to noticeable size distortions. In contrast, in a recent paper Harvey et al. (2016) consider the case with non-stationary volatility and show that the SADF test is severely oversized.

In this paper, we analyze the impact of serially correlated innovations and lag-length selection on the properties of the recursive right-tailed unit root tests by Phillips et al. (2011, 2015), which to our knowledge has not yet been addressed in the literature. As pointed out by Phillips et al. (2015), the limit distributions of SADF and GSADF also hold with autocorrelated innovations under suitable moment conditions and provided the lag-length goes to infinity at a suitable chosen rate as $T \to \infty$. However, in finite samples the tests may suffer from size distortion caused by autocorrelated innovations similar to what we know from standard unit root tests; for example, Schwert (1989) and Ng and Perron (1995, 2001). In practice, researchers typically deal with the issue of serially correlated innovations by including lags of the dependent variable in the auxiliary regression used to compute the test statistic. There are two important reasons for why we cannot just use our knowledge from the existing literature on unit root testing to guide us in our choice of lag-length in the auxiliary regressions in the SADF and GSADF tests. First, we test against the explosive and not the stationary alternative. Second, the test statistics are computed as the supremum of a sequence of unit root tests. Consequently, given the multiple testing feature of the recursive tests, there is a non-negligible risk that the supremum test statistic will be based on a sample window with very few observations and hence more likely to be size distorted. The importance of analyzing the effect of serially correlated innovations and lag-length selection is emphasized by the fact that we often find evidence of autoregressive and moving average components in the first difference of time series used in bubble tests. This includes, among others, the price–dividend ratio in case of stock markets and the price–rent ratio in case of housing markets.

Through a simulation study with parameter values motivated by empirical findings, we show that the presence of serially correlated innovations can lead to large size distortions for the recursive right-tailed unit root tests by Phillips et al. (2011, 2015). Size distortions do decrease with the sample size as expected, but even for very large samples, the tests can be critically oversized, especially the GSADF test. These results imply that we reject the null hypothesis of a random walk against the explosive alternative too often, i.e. we risk concluding the presence of a bubble when it does not exist. This result holds irrespective of how the lag-length is determined, either fixed or variable using information criteria such as the Bayesian Information Criterion (BIC).

In response to the size distortions caused by autocorrelated innovations and motivated by Park (2003), Chang and Park (2003) and Palm et al. (2008), we propose sieve bootstrap versions of the recursive right-tailed unit root tests. We show that the bootstrap tests are almost perfectly sized both with and without autocorrelated innovations and even in very small samples. Furthermore, we show that this strong improvement in size comes at a reasonably low cost in terms of power. We also pay special attention to the tests' ability to detect multiple bubbles and the date stamping feature of the tests. Consistent with a slightly lower power, the bootstrap tests are a bit more conservative in the number of bubbles they detect and provide a slightly longer delay for detection of bubble eruption. However, in contrast to the standard tests, the bootstrap tests provide a more or less perfect signal about the bubble collapse.

We apply the bootstrap GSADF test to the housing markets of a panel of OECD countries, and compare the results to those obtained using the standard GSADF test with both fixed and variable lag-length selection using BIC. Using the bootstrap test, we find less strong evidence of bubbles in the international housing market compared to existing evidence. For example, using the GSADF with automatic lag-length selection, we find evidence of bubbles in 17 out of 18 countries in our sample using a 1% significance level. In contrast, with the bootstrap test only 11 out of 18 countries are subject to bubbles using a 5% significance level. With a 1% significance level the number drops to five countries.

The rest of the paper is organized as follows. Section 2 provides an overview of the recursive right-tailed test procedures, we consider in this paper. Section 3 contains a simulation study of the size properties of recursive right-tailed unit root tests when innovations are serially correlated as well as empirical results emphasizing the importance of this issue. Section 4 describes the sieve bootstrap version of the tests and contains a simulation study analyzing their size and power. Section 5 provides an empirical application of the sieve bootstrap tests and Section 6 some concluding remarks. An online appendix provides supplementary results.

2. Right-tailed unit root tests for bubbles

Diba and Grossman (1988) were the first to propose a test that exploits the explosive characteristic of rational bubbles to detect exuberance in the stock market. They utilize unit root tests but instead of testing the null of a unit root against the stationary alternative, they consider the right tail of the distribution and test against the explosive alternative. However, through a simulation study (Evans, 1991) shows that right-tailed unit root tests have low power when trying to detect periodically collapsing bubbles.[2]

Phillips and Yu (2011) build upon the idea developed by Diba and Grossman (1988), but instead of running a single test over the whole sample, they implement right-tailed augmented Dickey–Fuller (ADF) tests using subsets of the data incremented by one observation at each run, where the largest of these test statistics is used to test for explosiveness. They name this method the Supremum Augmented Dickey–Fuller (SADF) test and show that it not only improves power – even in the presence of periodically collapsing bubbles – but also allows us to pinpoint the start and ending date of a bubble. Phillips et al. (2015) develop a generalized version of the SADF test (GSADF) by allowing both the start and end date of the sample window to vary. They find that the GSADF test has even higher power in detecting periodically collapsing bubbles. We describe the SADF and GSADF tests in more detail in the following sections.

[2] An alternative strand of the literature shows that the explosive characteristics of asset price bubbles can also be modeled with heavy-tailed non-causal linear autoregressive processes, see for example Gouriéroux and Zakoïan (2017) and Cavaliere et al. (2018).

2.1. The SADF test

The null hypothesis of the SADF test is that the series in question follows a random walk with asymptotically negligible drift,

$$y_t = dT^{-\eta} + \theta y_{t-1} + \varepsilon_{t,} \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2), \tag{2.1}$$

where d is a constant and $\eta > 1/2$ is a coefficient that determines the size of the drift as the sample size T goes to infinity and $\theta = 1$ is the autoregressive parameter. Before continuing with the description of the test it will be helpful to introduce some important notation. Let r_1 and r_2 be fractions of the total sample with $r_2 = r_1 + r_w$, where $r_w > 0$ is the fractional window size used in the auxiliary regressions embedded in the test. The methodology proposed by Phillips and Yu (2011) is to set the starting point of the regression window equal to the first observation (i.e. $r_1 = 0$) and using a minimum fractional window size of r_0, expand this window from r_0 to 1. This recursive methodology is based on a standard ADF regression given by

$$\Delta y_t = \alpha_{r_1,r_2} + \beta_{r_1,r_2} y_{t-1} + \sum_{i=1}^{k} \psi^i_{r_1,r_2} \Delta y_{t-i} + \varepsilon_t, \tag{2.2}$$

where k is the lag order and the subscripts r_1, r_2 indicate that the fractional regression window starts at r_1 and ends at r_2. More specifically, since the SADF test fixes the starting point at 0 the first regression will have a sample size of $\lfloor Tr_0 \rfloor$, where $\lfloor . \rfloor$ denotes the floor function, and increase by one observation at a time until the sequence reaches the end of the sample (i.e. $\lfloor Tr_w \rfloor = T$). Each of the ADF test statistics obtained from this recursive sequence is denoted by $ADF_0^{r_2}$ and the SADF test statistic is defined as the supremum of this sequence:

$$SADF(r_0) = \sup_{r_2 \in [r_0, 1]} \left\{ ADF_0^{r_2} \right\}. \tag{2.3}$$

The distribution of the SADF test statistic under the null is nonstandard, but asymptotic and finite sample critical values can be obtained by simulation.

2.2. The GSADF test

The GSADF test builds upon the idea of the SADF test but allows both the start, r_1, and end, r_2, points of the sample window to vary. Thus, for a given r_0, the GSADF test entails a double recursion scheme by allowing the end point of the regression window r_2 to vary from r_0 to 1 and the starting point r_1 to range from 0 to $r_2 - r_0$. Based on the data-generating process for the null given in (2.1) and the empirical regression model given in (2.2), the GSADF test statistic is defined as the largest ADF statistic that we obtain from this double recursion over all feasible ranges from r_1 to r_2 given a minimal window size r_0:

$$GSADF(r_0) = \sup_{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}} \left\{ ADF_{r_1}^{r_2} \right\}. \tag{2.4}$$

As in the case with the SADF test, the GSADF test statistic follows a nonstandard distribution and critical values are obtained by means of simulation. Our simulation study and the empirical application are based on these two tests.

2.3. The date-stamping of bubbles

One of the advantages of these recursive tests is that they allow us to pin-point the origin and collapse of the bubble. The date-stamping algorithm works by performing SADF tests in a backward expanding sample sequence, where the end point, r_2, at time τ is fixed such that $Tr_2 = \tau$, and the starting point varies from 0 to $r_2 - r_0$. For a given r_2, the supremum of this sequence defines the Backwards SADF (BSADF) test:

$$BSADF_{r_2}(r_0) = \sup_{r_1 \in [0, r_2 - r_0]} \left\{ ADF_{r_1}^{r_2} \right\} \tag{2.5}$$

We can infer whether or not observation τ is part of a potential bubble-period by comparing the BSADF test statistic for that observation to its corresponding critical value (based on a sample size of Tr_2). In our main analysis, we consider the size and power properties of the SADF test and its bootstrap counterpart. Since the BSADF test can be seen a sequence of SADF tests, it is intuitive to conclude that the results below also apply to the BSADF test. We explicitly consider the BSADF test in an analysis of multiple bubbles, cf. Phillips et al. (2015), as well as delay in the date-stamping of bubbles. In our empirical application we also compare the performance of the BSADF test to its bootstrapped counterpart.

2.4. Autocorrelation

In obtaining the distributions of the SADF and GSADF tests, we simulate under the null given by (2.1), where innovations are assumed $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$. In other words, innovations are among other things assumed to be serially uncorrelated. However, Phillips et al. (2015) conjecture that the asymptotic distributions of the SADF and GSADF test remain valid under the following assumption for innovations:

Assumption (A1): Let $u_t = \omega(L)\varepsilon_t = \sum_{j=0}^{\infty} \omega_j \varepsilon_{t-j}$, where $\sum_{j=0}^{\infty} j \left|\omega_j\right| < \infty$. Moreover, $\mathbb{E}\left(\varepsilon_t^2\right) = \sigma^2$ and $\mathbb{E}\left|\varepsilon_t\right|^{4+\delta} < \infty$ for some $\delta > 0$, provided that the autoregressive lag length k in (2.2) satisfies the following deterministic rate condition:

Assumption (A2): As $T \to \infty$, $k \to \infty$ with $k = O\left(T^{1/3}\right)$.

These assumptions are standard in the unit root testing literature; see, for example, Said and Dickey (1984) and Zivot and Andrews (2002).

Consequently, in large samples the SADF and GSADF test statistics should be reasonably sized also in the presence of serially correlated innovations provided a sufficiently long lag augmentation. It is, however, unclear how large the size distortions for the two tests are using empirically relevant sample sizes. There is strong evidence that standard ADF tests suffer from significant size distortions in small samples when innovations are autocorrelated; see, for example Schwert (1989) and Ng and Perron (1995, 2001). Given that we compute SADF and GSADF test statistics as suprema of a sequence of ADF test statistics we should expect even larger size distortions for these recursive tests. Phillips et al. (2015) suggest setting the minimum window length to $r_0 = 0.01 + 1.8/\sqrt{T}$. With an empirically relevant sample size of $T = 200$, this implies samples as small as 25 observations. Furthermore, given the multiple testing feature of the recursive tests, there is a non-negligible risk that the supremum ADF test statistic is in fact based on a sample window with very few observations and hence more likely to be size distorted.

3. Testing for bubbles with autocorrelated innovations

Given the potential size distortion caused by autocorrelated innovations, this section examines the effect that autoregressive and moving average components have on the SADF and GSADF tests at different lag-lengths in the auxiliary regressions. Section 3.1 motivates why we should be concerned about autocorrelated innovations in empirical applications of bubble test and furthermore presents the basic framework for the size analysis in Section 3.2.

3.1. Motivation and basic framework

The original Dickey–Fuller test statistic and corresponding critical values are based upon a regression model with a white noise error process. To deal with potential autocorrelation, Said and Dickey (1984) extended the original Dickey–Fuller regression model with lagged differences of the series, resulting in the augmented Dickey–Fuller (ADF) regression model presented in (2.2). The theoretical foundation for this regression is that we can approximate an ARMA(p,q) model of unknown order by an AR(k) process, where $k = O\left(T^{1/3}\right)$. Chang and Park (2002) show that this approximation holds for a general class of linear models and that, assuming (A2), the ADF test statistic has the same limiting distribution as the simple Dickey–Fuller t-statistic. The challenge is then to select the right autoregressive lag-length, k, in (2.2) such that the test is correctly sized and there is no loss of power. In the classic ADF test against stationarity the test statistic is usually oversized if k is too small, while if k is too large power is low. Further, the test against stationarity suffers from severe size distortions if the moving average component is negative and has a root close to unity; see, for example, Phillips and Perron (1988), Schwert (1989), Agiakloglou and Newbold (1992) and Ng and Perron (1995). While Phillips et al. (2015) and Harvey et al. (2016) analyze the effects of heteroskedastic innovations on the SADF and GSADF tests, the effects of serially correlated innovations and the impact of varying truncation lags under these circumstances have not previously been examined, at least to our knowledge.[3] The relevance of such an analysis is emphasized by the fact that we commonly find non-negligible autoregressive and moving average components in the series usually used to test for bubbles, namely price-dividend and price–rent ratios, where the latter pertains to housing markets.

Using the Bayesian Information Criterion (BIC) to determine the presence and order of ARMA components on the first difference of the annual price–dividend ratio of the market cap index of all American stocks in the period 1926–2011 obtained from the CRSP database, we find that the series contains a single MA component with a coefficient of 0.26 (t-statistic of 2.46). Using the same procedure on the differenced monthly S&P 500 price–dividend ratio in the period January 1871 to December 2010, obtained from Robert Shiller's website, we also find that the series contains a statistically significant MA component with a coefficient of 0.28 (t-statistic of 11.84).

The relevance of this issue is even greater on the housing market since many of the series used to test for bubbles in this market (i.e. house price indices and price–rent ratios) have autocorrelated innovations by construction (Ghysels et al., 2013). Again, using the same procedure as in the case of stocks to select the best ARMA(p,q) fit for the first difference of price–rent ratios based on housing data collected for the OECD, we find that the most common models are either MA(3) or AR(1) models. Table 1 shows the autoregressive and moving average coefficients and t-statistics for selected countries.

In our simulation study of the impact of autocorrelated innovations on right-tailed unit root tests, we start from the null hypothesis given in (2.1) and calculate finite sample critical values for the SADF and GSADF tests using 10,000 simulations. When simulating these critical values we follow Phillips et al. (2015) and set d, η, and θ equal to unity and the minimum window length to $r_0 = 0.01 + 1.8/\sqrt{T}$. The lag length used in the regressions to calculate the critical values is set to zero. To isolate the effects of AR and MA components in innovations when analyzing the size of the tests, we utilize the same data-generating process and

[3] Phillips et al. (2015) use a simulation study to show that stationary conditional heteroskedastic volatility does not seriously affect the size of the tests. However, Harvey et al. (2016) show that volatility of the non-stationary type results in heavily oversized tests and propose a wild bootstrap version of the SADF test that is robust to this issue. While we do not formally test for the presence of non-stationary volatility in the empirical time series we use to test for bubbles, visual inspection of the first differenced series does not seem to suggest any permanent shifts in volatility.

Table 1
ARMA(p,q) model fit and coefficients for first differenced price–rent ratios.

Country	Period	T	ARMA(p,q) coefficient / *(t-statistic)*					
			Constant	AR(1)	AR(2)	MA(1)	MA(2)	MA(3)
Australia	1972Q3 - 2019Q3	189	0.39	–	–	0.77	0.84	0.62
			(1.86)	–	–	*(14.35)*	*(12.62)*	*(9.42)*
Canada	1970Q1 - 2019Q4	199	0.51	0.51	–	–	–	–
			(3.20)	*(10.72)*	–	–	–	–
Denmark	1970Q1 - 2019Q2	198	0.23	–	–	0.62	0.59	0.34
			(0.81)	–	–	*(10.84)*	*(8.99)*	*(4.94)*
Finland	1970Q1 - 2019Q2	198	0.23	–	–	0.62	0.57	0.45
			(0.81)	–	–	*(10.91)*	*(8.37)*	*(7.98)*
France	1970Q1 - 2019Q2	198	0.25	0.75	–	0.09	0.47	–
			(0.94)	*(15.11)*	–	*(1.11)*	*(6.53)*	–
Germany	1970Q1 - 2019Q2	198	-0.03	0.25	0.37	–	–	–
			(−0.13)	*(5.50)*	*(7.16)*	–	–	–
Ireland	1970Q1 - 2019Q2	198	0.43	0.63	–	–	–	–
			(0.58)	*(14.52)*	–	–	–	–
Italy	1970Q1 - 2019Q2	198	0.22	–	–	0.76	0.59	0.32
			(0.38)	–	–	*(21.50)*	*(10.33)*	*(5.41)*
Japan	1970Q1 - 2019Q2	198	0.05	0.76	–	0.10	0.31	–
			(0.11)	*(13.92)*	–	*(1.31)*	*(4.24)*	–
Netherlands	1970Q1 - 2019Q2	198	0.20	0.81	–	-0.58	0.41	–
			(0.34)	*(18.22)*	–	*(−9.24)*	*(7.90)*	–
New Zealand	1970Q1 - 2019Q2	198	0.40	0.66	–	–	–	–
			(1.93)	*(17.80)*	–	–	–	–
Norway	1979Q1 - 2019Q3	163	0.44	–	–	0.51	0.61	0.33
			(2.90)	–	–	*(10.18)*	*(8.54)*	*(3.97)*
Portugal	1988Q1 - 2019Q2	126	−0.00	0.90	–	−0.61	–	–
			(0.00)	*(13.29)*	–	*(−5.15)*	–	–
Spain	1971Q1 - 2019Q2	194	0.44	0.83	–	–	–	–
			(0.81)	*(23.11)*	–	–	–	–
Sweden	1980Q1 - 2019Q2	158	0.14	0.77	–	–	–	–
			(0.41)	*(19.89)*	--	--	--	–
Switzerland	1970Q1 - 2019Q3	199	0.12	0.76	–	-0.56	0.35	–
			(0.28)	*(13.70)*	--	*(−6.72)*	*(5.93)*	–
United Kingdom	1970Q1 - 2019Q3	199	0.33	–	–	0.61	0.64	0.39
			(1.24)	–	–	*(12.49)*	*(10.88)*	*(6.92)*
United States	1970Q3 - 2019Q3	199	0.09	0.39	0.35	–	–	–
			(0.30)	*(8.73)*	*(7.96)*	–	–	–

ARMA(p,q) fit for the first difference in price–rent indices. The optimal p and q are selected by the BIC using a maximum AR order of 2 and a maximum MA order of 5. All the estimated models are invertible.

parameters as the ones used in the calculation of critical values, with the only difference that now innovations are autocorrelated. Thus, the data-generating process for these series is given by

$$\begin{aligned} y_t &= dT^{-\eta} + \theta y_{t-1} + v_t, \\ v_t &= \phi_1 v_{t-1} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + \vartheta_3 \varepsilon_{t-3}, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_v). \end{aligned} \tag{3.1}$$

The parameter combinations for v_t and the size of the coefficients are motivated by the empirical findings for the price-dividend and price–rent ratios. We consider the case where v_t follows an MA(1) process, which is mainly relevant for stock market data and the cases where innovations follow an MA(3) or AR(1) process, which are relevant for housing market indices. The simulated processes for v_t are invertible and have standard normal innovations, i.e. $\sigma_v = 1$. We also consider the case where v_t is a white noise process. This case is interesting since choosing the wrong autoregressive lag-length in (2.2) can also distort the size of the test, even without the presence of autocorrelation.

We generate S replications using (3.1) and count the proportion of test statistics exceeding the critical values. For the SADF test we use $S = 4000$ and for the GSADF test we use $S = 2000$. If the tests are unaffected by serially correlated innovations or an incorrectly specified lag-length, the proportion of test statistics surpassing the critical values should be equal to the nominal size. All simulations are conducted using a nominal size of 5%.

Fig. 1 shows a simulated process using (3.1) together with the price–rent ratio of Norway from 1979Q1 to 2019Q3. The parameter values are set to match the coefficients of the estimated MA(3) model for Norway presented in Table 1. Although not shown in Table 1, the R^2 of the regression for Norway is 0.57, which implies a fairly good fit. We set $d = \eta = \theta = 1$, $y_0 = 37.4$ and $\sigma_v = 1.06$ where the latter two are set to match the initial value of the price–rent ratio in Norway and the standard deviation of the first difference of the same series. The figure shows that the simulated unit root process with MA innovations is realistically capturing the dynamics that drive the price–rent ratio since both series display a fairly similar and rather persistent behavior. More importantly, we can intuitively see why MA components can result in over-rejections of the null when using the SADF and GSADF tests since this type of persistence in innovations can make the series look temporarily explosive even though this is not the case.

Table 2
Empirical size for SADF and GSADF tests with fixed lag-length.

SADF											
T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
100	**0**	**0**	**0**	**0**	0.05	0.06	0.09	0.10	0.13	0.16	0.20
	0.5	**0**	**0**	**0**	0.24	0.04	0.14	0.11	0.16	0.17	0.23
	0.5	**0.5**	**0.5**	**0**	0.52	0.13	0.09	0.09	0.26	0.26	0.27
	0	**0**	**0**	**0.5**	0.42	0.10	0.12	0.14	0.17	0.20	0.26
200	**0**	**0**	**0**	**0**	0.04	0.06	0.07	0.08	0.10	0.11	0.13
	0.5	**0**	**0**	**0**	0.25	0.03	0.12	0.08	0.12	0.12	0.15
	0.5	**0.5**	**0.5**	**0**	0.58	0.09	0.05	0.05	0.21	0.18	0.15
	0	**0**	**0**	**0.5**	0.48	0.08	0.09	0.10	0.13	0.14	0.16
400	**0**	**0**	**0**	**0**	0.05	0.06	0.07	0.07	0.08	0.09	0.09
	0.5	**0**	**0**	**0**	0.27	0.02	0.10	0.06	0.09	0.09	0.10
	0.5	**0.5**	**0.5**	**0**	0.65	0.08	0.03	0.04	0.20	0.15	0.11
	0	**0**	**0**	**0.5**	0.56	0.08	0.09	0.09	0.11	0.11	0.12
1600	**0**	**0**	**0**	**0**	0.05	0.05	0.05	0.06	0.06	0.06	0.06
	0.5	**0**	**0**	**0**	0.31	0.02	0.09	0.04	0.07	0.06	0.07
	0.5	**0.5**	**0.5**	**0**	0.75	0.06	0.02	0.02	0.17	0.11	0.06
	0	**0**	**0**	**0.5**	0.65	0.06	0.07	0.07	0.08	0.08	0.08
GSADF											
T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
100	**0**	**0**	**0**	**0**	0.05	0.09	0.18	0.26	0.39	0.53	0.74
	0.5	**0**	**0**	**0**	0.29	0.08	0.28	0.31	0.48	0.59	0.82
	0.5	**0.5**	**0.5**	**0**	0.74	0.31	0.30	0.32	0.66	0.78	0.90
	0	**0**	**0**	**0.5**	0.60	0.20	0.32	0.40	0.54	0.69	0.86
200	**0**	**0**	**0**	**0**	0.05	0.08	0.15	0.20	0.30	0.38	0.51
	0.5	**0**	**0**	**0**	0.37	0.06	0.25	0.23	0.39	0.43	0.57
	0.5	**0.5**	**0.5**	**0**	0.86	0.29	0.22	0.23	0.60	0.63	0.70
	0	**0**	**0**	**0.5**	0.73	0.20	0.29	0.35	0.45	0.53	0.65
400	**0**	**0**	**0**	**0**	0.05	0.08	0.13	0.17	0.24	0.29	0.37
	0.5	**0**	**0**	**0**	0.43	0.04	0.24	0.19	0.30	0.34	0.42
	0.5	**0.5**	**0.5**	**0**	0.93	0.25	0.14	0.14	0.55	0.53	0.52
	0	**0**	**0**	**0.5**	0.82	0.15	0.22	0.27	0.34	0.40	0.47
1600	**0**	**0**	**0**	**0**	0.05	0.07	0.10	0.12	0.14	0.17	0.20
	0.5	**0**	**0**	**0**	0.55	0.02	0.18	0.10	0.19	0.19	0.25
	0.5	**0.5**	**0.5**	**0**	0.99	0.16	0.04	0.04	0.48	0.38	0.26
	0	**0**	**0**	**0.5**	0.96	0.12	0.15	0.17	0.21	0.23	0.27

The tests are applied to series generated by (3.1) with parameters given in the table. Initial windows are set to $r_0 = \{0.190, 0.130, 0.100, 0.055\}$ for $T = \{100, 200, 400, 1600\}$, respectively. We use 4000 replications for the SADF and 2000 for the GSADF. Nominal size is 5%.

This problem is compounded by the fact that both tests work by running ADF regressions on different subsets of the data and, as a consequence, some of these regressions will have trouble differentiating the explosive-like behavior of persistent innovations from that of a true explosive autoregressive root.

3.2. Size analysis

In this section we perform a size analysis for the SADF and GSADF tests using both a fixed and variable lag length in the auxiliary regressions, i.e. we analyze how changes in the lag length and the presence of AR or MA components lead to incorrect rejections of the null hypothesis of "no bubbles". Table 2 shows the results for fixed lag length across the sample sizes $T = \{100, 200, 400, 1600\}$ and with k taking integer values between 0 and 6. Motivated by the empirical evidence from the stock and housing markets, we consider an MA(1) model with $\vartheta_1 = 0.50$, an MA(3) model with $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, an AR(1) model with $\phi_1 = 0.5$ and the white noise case for comparison. The online appendix contains results for additional empirically relevant data-generating parameters, but the overall conclusion remains robust also in these cases.

With serially correlated innovations, Table 2 shows that fixing the lag length k at zero results in severe size distortions for both SADF and GSADF. For example, for GSADF with $T = 200$ and $k = 0$ we incorrectly reject the null hypothesis in 37%, 86% and 73% of the cases for the MA(1), MA(3) and AR(1) model, respectively. With $k > 0$ we are generally able to reduce the size distortions, but only in a few cases and mainly for SADF do we obtain an empirical size matching the nominal one. We generally find the smallest size distortions for positive but low values of k, while a large number of lags in the auxiliary regression leads to heavily oversized tests. This is especially the case for GSADF, where the size distortions in some cases even surpass those for $k = 0$. These results imply

Table 3
Empirical size for SADF and GSADF tests with variable lag-length.

SADF										
T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$k_{max}=1$	$k_{max}=2$	$k_{max}=3$	$k_{max}=4$	$k_{max}=5$	$k_{max}=6$
100	**0**	**0**	**0**	**0**	0.06	0.07	0.07	0.08	0.09	0.11
	0.5	**0**	**0**	**0**	0.10	0.14	0.14	0.16	0.17	0.19
	0.5	**0.5**	**0.5**	**0**	0.17	0.17	0.18	0.25	0.27	0.30
	0	**0**	**0**	**0.5**	0.16	0.18	0.19	0.20	0.21	0.24
200	**0**	**0**	**0**	**0**	0.05	0.06	0.06	0.06	0.06	0.07
	0.5	**0**	**0**	**0**	0.06	0.11	0.11	0.12	0.12	0.12
	0.5	**0.5**	**0.5**	**0**	0.15	0.14	0.14	0.22	0.22	0.22
	0	**0**	**0**	**0.5**	0.12	0.13	0.14	0.15	0.15	0.16
400	**0**	**0**	**0**	**0**	0.06	0.06	0.06	0.06	0.06	0.06
	0.5	**0**	**0**	**0**	0.04	0.10	0.10	0.10	0.10	0.10
	0.5	**0.5**	**0.5**	**0**	0.10	0.08	0.08	0.19	0.18	0.18
	0	**0**	**0**	**0.5**	0.09	0.09	0.10	0.10	0.11	0.11
GSADF										
T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$k_{max}=1$	$k_{max}=2$	$k_{max}=3$	$k_{max}=4$	$k_{max}=5$	$k_{max}=6$
100	**0**	**0**	**0**	**0**	0.10	0.17	0.22	0.31	0.42	0.65
	0.5	**0**	**0**	**0**	0.24	0.34	0.37	0.48	0.58	0.78
	0.5	**0.5**	**0.5**	**0**	0.59	0.61	0.63	0.75	0.84	0.92
	0	**0**	**0**	**0.5**	0.49	0.53	0.57	0.64	0.71	0.85
200	**0**	**0**	**0**	**0**	0.09	0.14	0.17	0.21	0.24	0.28
	0.5	**0**	**0**	**0**	0.25	0.32	0.34	0.40	0.43	0.49
	0.5	**0.5**	**0.5**	**0**	0.60	0.62	0.64	0.74	0.77	0.80
	0	**0**	**0**	**0.5**	0.53	0.55	0.57	0.61	0.63	0.67
400	**0**	**0**	**0**	**0**	0.08	0.10	0.12	0.14	0.15	0.17
	0.5	**0**	**0**	**0**	0.22	0.30	0.31	0.34	0.36	0.38
	0.5	**0.5**	**0.5**	**0**	0.51	0.53	0.54	0.67	0.69	0.70
	0	**0**	**0**	**0.58**	0.42	0.48	0.50	0.53	0.55	0.56

The tests are applied to series generated by (3.1) with parameters given in the table. Initial windows are set to $r_0 = \{0.190, 0.130, 0.100\}$ for $T = \{100, 200, 400\}$, respectively. We use 4000 replications for the SADF and 2000 for the GSADF. Nominal size is 5%.

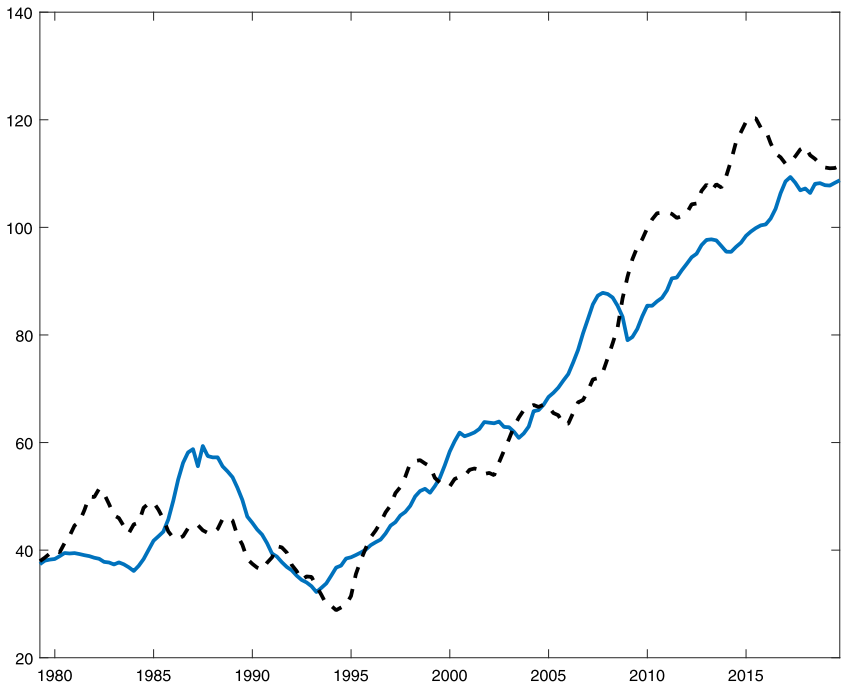

Fig. 1. Simulated unit root process with MA(3) innovations and the price–rent ratio in Norway. The figure shows a simulated unit root process with MA(3) innovations (dashed line) and the price–rent ratio in Norway from 1979Q1 to 2019Q3 (solid line). The data generating process for the simulated series is (3.1) with $d = \eta = \theta = 1$, $\phi_1 = 0$, $\vartheta_1 = 0.51$, $\vartheta_2 = 0.61$, $\vartheta_3 = 0.33$, $y_0 = 37.4$ and $\sigma_v = 1.06$.

that the usual suggestion for ADF tests against stationarity, of selecting a large k in the presence of serially correlated innovations, is not recommended when testing against explosiveness.[4]

Comparing SADF and GSADF it is clear that the latter is subject to the largest size distortions, which follows naturally from the regression windows used to compute the SADF test being a subset of those used in the GSADF case. Table 2 also shows that for $k > 0$, size distortions decrease with the sample size as expected. For SADF and across the three data-generating processes and six non-zero values of k, the average empirical size is 0.16 for $T = 100$, and 0.07 for $T = 1600$. For GSADF the corresponding numbers are 0.49 and 0.19. For $k = 0$ size deteriorates with the sample size and for some data-generating processes it can lead to almost 100% rejection rates when using the GSADF test.

While it seems from Table 2 that $k = 1$ works well when v_t follows an MA(1) process, this relatively good performance does not extend to the AR(1) and MA(3) cases, especially for the GSADF test. From a practical point of view, the autoregressive and moving average orders of the data are not known with certainty. Furthermore, given that both tests work by running regressions on different subsets of the sample, it seems natural to allow for flexibility in the lag order across regressions due to possible structural breaks in the data. For these reasons it appears that an optimal solution is to allow for a data-dependent choice of k (for each of the auxiliary regressions). Table 3 shows the empirical size for SADF and GSADF using the BIC to determine the lag length. We let the maximum number of lags k_{max} take integer values between 1 and 6. For computational tractability we here only consider the empirically most relevant sample sizes $T = \{100, 200, 400\}$. The general conclusions from Table 3 are similar to the ones based on fixed lags: Only in very few cases do we obtain an empirical size matching the nominal one, GSADF suffers from larger size distortions than SADF, size distortions decrease with the sample size and a high number of (potential) lags increases the size distortions. Consequently, given empirically relevant sample sizes a data-dependent choice of k does not eliminate the size distortions that arise due to serially correlated innovations. However, automatic lag-length selection methods appear to be asymptotically valid. For example, in unreported simulation results, we find that the using the BIC with a maximum lag length of $k_{max} = int[8(T/100)^{1/4}]$ with a sample size of $T = 10{,}000$ results in an empirical size of 0.06 for the MA(3) case with $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, 0.05 for the MA(1) case with $\vartheta_1 = 0.5$ and 0.05 for the AR(1) case with $\phi_1 = 0.5$.

As a final note, Tables 2 and 3 show that also in the white noise case, increasing the lag length leads to large size distortions in empirically relevant sample sizes, especially for GSADF. Since the presence of serial correlation is not always clear, an important feature of a bootstrap approach designed to eliminate size distortions due to serial correlation is that it also performs well in the case of white noise innovations.

4. The sieve bootstrap SADF and GSADF tests

As shown in the previous section, using either a fixed lag length or some automatic lag selection method is generally not an effective strategy to control the size of the SADF and GSADF tests in the presence of serially correlated innovations. This problem is particularly serious for the GSADF test, which is severely oversized. Motivated by Park (2003), Chang and Park (2003) and Palm et al. (2008) who consider sieve bootstrap versions of the ADF test, we propose sieve bootstrap SADF and GSADF tests. Our algorithm is also similar to the bootstrap algorithm proposed by Gutierrez (2011) for the SADF test.[5] Chang and Park (2003) and Palm et al. (2008) show that under assumption (A1) and assuming that the lag order of the ADF regression follows (A2), the asymptotic distribution of the ADF bootstrap test is the same under the null as the asymptotic distribution of the original ADF test. By means of simulations they show that the bootstrap ADF test has better empirical size in the presence of serially correlated innovations, and more importantly, these improvements in size come at no cost for the power of the test. In this section, we begin by describing the sieve bootstrap algorithm and then investigate the empirical size and power of the bootstrap versions of the tests in the case where innovations are serially correlated.

4.1. The sieve bootstrap algorithm

The sieve bootstrap algorithm consists of the following steps:

Step 1. Based on the full sample estimate by OLS the ADF regression, to obtain estimates $\hat{\psi}_i$ and residuals:

$$\hat{\varepsilon}_t = y_t - \hat{\alpha} - \hat{\beta} y_{t-1} - \sum_{i=1}^{k^*} \hat{\psi}_i \Delta y_{t-i}, \qquad t = k^* + 1, \ldots, T \tag{4.1}$$

where, for a given k_{max}, we let an information criterion such as the AIC or BIC select the optimal order, k^*, for the approximated autoregression.

Step 2. Generate an iid sample of bootstrap errors, ε_t^*, by drawing randomly with replacement from

$$\hat{\varepsilon}_t - \left(T - k^*\right)^{-1} \sum_{t=1+k^*}^{T} \hat{\varepsilon}_t \tag{4.2}$$

[4] Schwert (1989) suggests that when using ADF tests to test for stationarity, if the series in question has large negative MA components, it is preferable to select a large k since that would result in a test that is close to nominal significance levels.

[5] The main difference between our approach and the one proposed by Gutierrez (2011) is that the latter imposes a unit root restriction in the regression whereas we estimate the autoregressive coefficient β in (4.1). In this sense, our algorithm follows Palm et al. (2008) who suggest estimating β instead of imposing a unit root restriction. We find that this small change results in better power properties in the case of the SADF and GSADF tests.

Step 3. Construct u_t^* recursively from ε_t^* as

$$u_t^* = \sum_{i=1}^{q^*} \hat{\psi}_i u_{t-i}^* + \varepsilon_t^*. \tag{4.3}$$

To have a full bootstrap sample of size T and eliminate any initialization effect, we draw $(T - q^*) + b$ bootstrap errors from step 2 and then discard the first $b - q^*$ values of u_t^*. With u_t^* we can build y_t^* as $y_t^* = y_{t-1}^* + u_t^*$, $t = 1, \dots, T$ with $y_0^* = 0$.

Step 4. Using y_t^* we calculate the bootstrap test statistics:

$$SADF^*(r_0) = \sup_{r_2 \in [r_0, 1]} \left\{ ADF_0^{*r_2} \right\}, \tag{4.4}$$

$$GSADF^*(r_0) = \sup_{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}} \left\{ ADF_{r_1}^{*r_2} \right\}. \tag{4.5}$$

The lag-length in the auxiliary ADF regressions that conform these bootstrap test statistics should be fixed at k^* as determined in step 1.

Step 5. Calculate the bootstrap critical values $cv(q)^{SADF^B}$ or $cv(q)^{GSADF^B}$ for nominal significance level q, by repeating steps 2 to 4 M^* times and obtaining the q-quantile of the ordered bootstrap tests statistics. More specifically, for $m = 1, \dots, M^*$ we obtain $\left\{ SADF_m^*(r_0) \right\}_{m=1}^{M^*}$ or $\left\{ GSADF_m^*(r_0) \right\}_{m=1}^{M^*}$ and calculate $cv(q)^{SADF^B}$ or $cv(q)^{GSADF^B}$ as

$$cv(q)^{SADF^B} := \max \left\{ x : M^{*-1} \sum_{m=1}^{M^*} I\left(SADF_m^*(r_0) < x \right) \leq q \right\}, \tag{4.6}$$

$$cv(q)^{GSADF^B} := \max \left\{ x : M^{*-1} \sum_{m=1}^{M^*} I\left(GSADF_m^*(r_0) < x \right) \leq q \right\}. \tag{4.7}$$

Step 6. Calculate the actual test statistics, $SADF^B(r_0)$ and $GSADF^B(r_0)$, based on y_t and using a lag order equal to k^* as determined in step 1. Reject the null of "no bubbles" if the test statistic is larger than the bootstrap critical value calculated in step 5.[6]

Note that in the implementation of the bootstrap algorithm above, we did not specify the lag order of the sieve, q^*, in step 4. As noted by Cavaliere and Robert Taylor (2009) the choice of q^* is motivated purely for finite sample concerns, and q^* does not have to increase to infinity with sample size. However, the following assumption for k^* and q^* is required:

Assumption (A3): Let $k^* = O\left(T^{1/3}\right)$. There is a T^* such that $q^* \leq k^*$ for all $T > T^*$.

In practice, we find that setting $q^* = k^*$ results in the best (finite sample) size properties. Indeed, setting $k^* > q^*$ results in finite sample oversizing similar to the one shown in Section 3.2 for over-specified lag order models.

4.2. Size of the $SADF^B$ and $GSADF^B$ tests

To analyze the empirical size properties of the sieve bootstrap SADF and GSADF tests ($SADF^B$ and $GSADF^B$, respectively), we use (3.1) as the data-generating process with parameter combinations presented in Tables 2 and 3. Further results using other data-generating parameters that support the conclusion from this section are given in the online appendix. For computational tractability, we consider the sample sizes $T = \{100, 200, 400\}$ and use 4000 simulations in analyzing the $SADF^B$ test and 2000 simulations in case of the $GSADF^B$ test. For each simulated series, we use $M^* = 899$ bootstrap replications to calculate the critical values, $cv(q)^{SADF^B}$ and $cv(q)^{GSADF^B}$. We set the maximum lag-length of the bootstrap tests to $k_{max} = int[8(T/100)^{1/4}]$, and let BIC select the optimal lag-length, k^*.

Table 4 shows that the bootstrap procedure effectively controls the size of the tests when innovations are serially correlated, even for relatively small sample sizes. In the $\vartheta_1 = 0.5$ and $\phi_1 = 0.5$ cases, empirical size is within one percentage point from nominal size, irrespective of the sample size. The results for the MA(3) case are also robust, and the difference between empirical and nominal size is within three percentage points. Considering the size distortions in standard SADF and GSADF tests in the presence of serially correlated innovations (Tables 2 and 3), the performance of the $SADF^B$ and $GSADF^B$ tests seems comparatively impressive. We also note that the sieve bootstrap SADF and GSADF tests are perfectly sized with serially uncorrelated innovations, which in light of uncertainty about the degree of serial correlation further strengthens the applicability of the bootstrap tests.

We note that instead of the sieve bootstrap it is also possible to use the recolored version of the wild bootstrap proposed by Harvey et al. (2016). The recolored version of the wild bootstrap has the advantage of being robust to both autocorrelation and non-stationary volatility. However, unreported simulation results show that in the absence of non-stationary volatility the sieve bootstrap has moderately better finite sample size properties. We also test if GARCH innovations, with reasonable parameters driving the process, has any impact on the results and find no significant differences between the sieve bootstrap and wild bootstrap versions of the tests. In fact, using the same data-generating parameters as Phillips et al. (2015) in their analysis of GARCH errors, we find both $SADF^B$ and $GSADF^B$ to be more or less perfectly sized. Given the focus on autocorrelated innovations, we do not explore time-varying volatility further in this paper.

[6] MATLAB programs implementing $SADF^B$, $GSADF^B$ and $BSADF^B$ tests are available on Pedersen's website.

Table 4
Empirical size for SADF and GSADF bootstrap tests.

T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$SADF^B$	$GSADF^B$
100	**0**	**0**	**0**	**0**	0.05	0.05
	0.5	**0**	**0**	**0**	0.04	0.04
	0.5	**0.5**	**0.5**	**0**	0.07	0.07
	0	**0**	**0**	**0.5**	0.06	0.05
200	**0**	**0**	**0**	**0**	0.05	0.05
	0.5	**0**	**0**	**0**	0.05	0.05
	0.5	**0.5**	**0.5**	**0**	0.08	0.08
	0	**0**	**0**	**0.5**	0.05	0.05
400	**0**	**0**	**0**	**0**	0.05	0.05
	0.5	**0**	**0**	**0**	0.06	0.06
	0.5	**0.5**	**0.5**	**0**	0.08	0.08
	0	**0**	**0**	**0.5**	0.05	0.06

The tests are applied to series generated by (3.1) with parameters given in the table. Initial windows are set to $r_0 = \{0.190, 0.130, 0.100\}$ for $T = \{100, 200, 400\}$, respectively. We use 4000 replications for the $SADF^B$ and 2000 for the $GSADF^B$. The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

Table 5
Empirical power for the SADF bootstrap test.

T	ϑ_1	ϑ_2	ϑ_3	ϕ_1	A: Base case		B: Long bubble		C: Early bubble		D: Late bubble	
					$SADF^A$	$SADF^B$	$SADF^A$	$SADF^B$	$SADF^A$	$SADF^B$	$SADF^A$	$SADF^B$
100	**0**	**0**	**0**	**0**	0.45	0.40	0.65	0.62	0.35	0.32	0.50	0.50
	0.5	**0**	**0**	**0**	0.47	0.33	0.66	0.54	0.34	0.22	0.52	0.39
	0.5	**0.5**	**0.5**	**0**	0.47	0.32	0.66	0.52	0.34	0.24	0.51	0.35
	0	**0**	**0**	**0.5**	0.48	0.37	0.67	0.58	0.37	0.28	0.53	0.43
200	**0**	**0**	**0**	**0**	0.86	0.82	0.95	0.94	0.91	0.87	0.80	0.80
	0.5	**0**	**0**	**0**	0.77	0.68	0.89	0.87	0.79	0.69	0.74	0.69
	0.5	**0.5**	**0.5**	**0**	0.70	0.58	0.87	0.81	0.67	0.51	0.72	0.59
	0	**0**	**0**	**0.5**	0.74	0.69	0.88	0.86	0.71	0.65	0.73	0.67
400	**0**	**0**	**0**	**0**	0.88	0.87	0.96	0.96	0.90	0.90	0.86	0.86
	0.5	**0**	**0**	**0**	0.83	0.81	0.94	0.94	0.84	0.82	0.86	0.83
	0.5	**0.5**	**0.5**	**0**	0.83	0.76	0.93	0.90	0.80	0.72	0.84	0.78
	0	**0**	**0**	**0.5**	0.83	0.80	0.94	0.93	0.81	0.79	0.85	0.82

The tests are applied to series generated by (4.8). Initial windows are set to $r_0 = \{0.190, 0.130, 0.100\}$ for $T = \{100, 200, 400\}$, respectively. We set $y_0 = 100$, $\sigma_v = 6.79$,$\delta_1 = 1 + T^{-\alpha}$ and $\delta_2 = 1 - T^{-\alpha}$ with $\alpha = 0.6$. Other parameters are given in the table. Panel A represents the base case with $r_e = 0.4, r_c = 0.6, r_x = 0.7$, Panel B a long bubble with $r_e = 0.4, r_c = 0.7, r_x = 0.8$, Panel C an early bubble with $r_e = 0.2, r_c = 0.4, r_x = 0.5$ and Panel D a late bubble with no collapse $r_e = 0.8$. We use 4000 replications. The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

4.3. Power of the $SADF^B$ and $GSADF^B$ tests

As shown in the previous section the sieve bootstrap is very successful in restoring the size of both the SADF and GSADF test in the presence of autocorrelated innovations. In this section, we evaluate to what extent this improvement in size comes at a cost in terms of power. To this end, we consider mildly explosive processes with autocorrelated innovations where the bubble collapse is modeled as a stationary process, cf. Harvey et al. (2016) and Phillips and Shi (2018). Compared to alternative bubble processes with either no or an abrupt collapse, the stationary collapse appears most relevant empirically. The bubble process is given as:

$$y_t = \begin{cases} y_{t-1} + \nu_t, & t = 1, \dots, \tau_e - 1 \\ \delta_1 y_{t-1} + \nu_t, & t = \tau_e, \dots, \tau_c - 1 \\ \delta_2 y_{t-1} + \nu_t, & t = \tau_c, \dots, \tau_x - 1 \\ y_{t-1} + \nu_t, & t = \tau_x, \dots, T \end{cases} \tag{4.8}$$

$$\nu_t = \phi_1 \nu_{t-1} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + \vartheta_3 \varepsilon_{t-3} \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_v).$$

Table 5 reports power for the $SADF^B$ test given a single bubble modeled according to (4.8) with innovations generated as white noise, $MA(1)$ with $\vartheta_1 = 0.5$, $MA(3)$ with $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$ and $AR(1)$ with $\phi_1 = 0.5$, respectively. For computational tractability we do not report results for the $GSADF^B$ test, but unreported results show that in terms of power the test is very similar to $SADF^B$ in the single-bubble case. In Section 4.5 we consider multiple bubbles and pay explicit attention to $BSADF$, which forms the basis for $GSADF$. For comparison, Table 5 also shows the (infeasible) size-adjusted power of the SADF test, $SADF^A$, for each case. Size-adjusted power is calculated by using critical values under the null, but allowing for autoregressive or moving average components. In other words, we use (3.1) instead of (2.1) as null hypothesis. Following Phillips et al. (2015) we set $y_0 = 100$, $\sigma_v = 6.79$, and $\delta_1 = 1 + T^{-\alpha}$ with $\alpha = 0.6$. Likewise we define $\delta_2 = 1 - T^{-\alpha}$. Under this bubble process, the series starts as a random

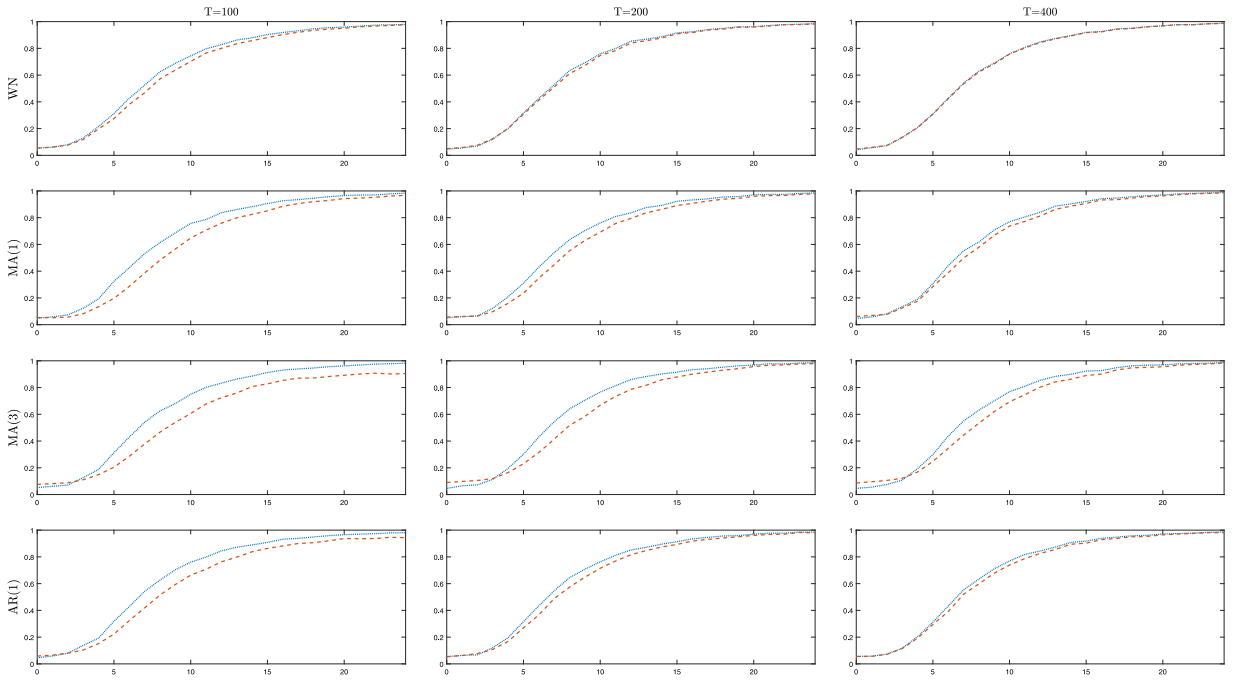

Fig. 2. Local power of the $SADF^B$ test. The figure shows the finite sample power of the $SADF^B$ test (dashed line) and the infeasible size-adjusted power of the $SADF^A$ test (dotted line).

walk and continues to be so until τ_e, where the series becomes explosive with a local to unity autoregressive coefficient, δ_1, and this explosivity continues until observation τ_c, where the stationary collapse begins. After the collapse at τ_x, the series continues as a random walk until the end of the series. We consider the samples sizes $T = \{100, 200, 400\}$, which result in autoregressive coefficients during the bubble period of $\delta_1 = \{1.06, 1.04, 1.03\}$ and during the collapse of $\delta_2 = \{0.94, 0.96, 0.97\}$.[7] To account for bubbles of various length and start date, Table 5 shows results for a base case ($r_e = 0.4, r_c = 0.6, r_x = 0.7$), a longer bubble ($r_e = 0.4, r_c = 0.7, r_x = 0.8$), an early bubble ($r_e = 0.2, r_c = 0.4, r_x = 0.5$), and a late bubble with no collapse ($r_e = 0.8$).[8]

From Table 5 it is clear that for both $SADF^A$ and $SADF^B$, power increases with the sample size and the length of the bubble. Furthermore, we see that with white noise innovations, there is close to no power loss for $SADF^B$ relative to $SADF^A$, even with a very small sample size. Given that $SADF^A$ is infeasible in the presence of autocorrelated innovations due to lack of knowledge of the exact data-generating process, the white noise case represents a fair evaluation of the potential power loss of the bootstrap test. The conclusion from the white noise case also holds with autocorrelated innovations in larger sample sizes like $T = 400$. For smaller sample sizes, the conclusion somewhat depends on the data-generating process for innovations as well as the length and start date of the bubble. Generally, the power loss is smallest for the AR(1) process and largest for the MA(3) process. For example, for $T = 200$ and the base case (Panel A), the power of $SADF^B$ decreases compared to $SADF^A$ by 12 percentage point with MA(3) innovations and 6 percentage points with AR(1) innovations. We also notice that the power loss decreases with the length of the bubble. In the long bubble case (Panel B) the corresponding numbers are 6 and 2 percentage points for the MA(3) and AR(1) process, respectively. The power loss increases slightly for $T = 100$, but overall the loss is of a limited magnitude.

To further support the general conclusion that the bootstrap test suffers from limited loss of power, Fig. 2 plots power curves for $SADF^A$ and $SADF^B$ for the base case ($r_e = 0.4, r_c = 0.6, r_x = 0.7$) and the same sample sizes and data-generating processes for innovations as in Table 5. We follow the literature in constructing asymptotic local power curves and set $y_0 = 0$, $\sigma_v = 1$, $\delta_1 = 1 + c/T$ and $\delta_2 = 1 - c/T$ with $c \in \{1, 2, \ldots, 24\}$. From Fig. 2 it is clear that as expected the larger the autoregressive root (large c), the larger the power. Comparing $SADF^A$ and $SADF^B$ we again see that there is no power loss in large samples. Likewise with white noise innovations, irrespective of the sample size. Fig. 2 also shows that there is virtually no loss of power for highly explosive bubbles (large c) as well as bubbles with an autoregressive root only slightly above one (small c). Only for intermediate values of δ_1 and δ_2 do we observe a noticeable difference between $SADF^A$ and $SADF^B$. The autoregressive roots for $T = \{100, 200\}$ used in Table 5 correspond to c equal to 6 and 8, respectively. Fig. 2 shows that the power loss is largest for autoregressive roots in this vicinity, which suggests that the results given in Table 5 provide a conservative estimate of the loss of power by using the bootstrap test.

[7] In response to some bubbles collapsing very rapidly, we have experimented with a less persistent but still empirically relevant collapse, but this has close to no effect on the results.

[8] r_e refers to the fractional observation where the bubble erupts. Likewise, r_c refers to bubble collapse and r_x to the return to a random walk.

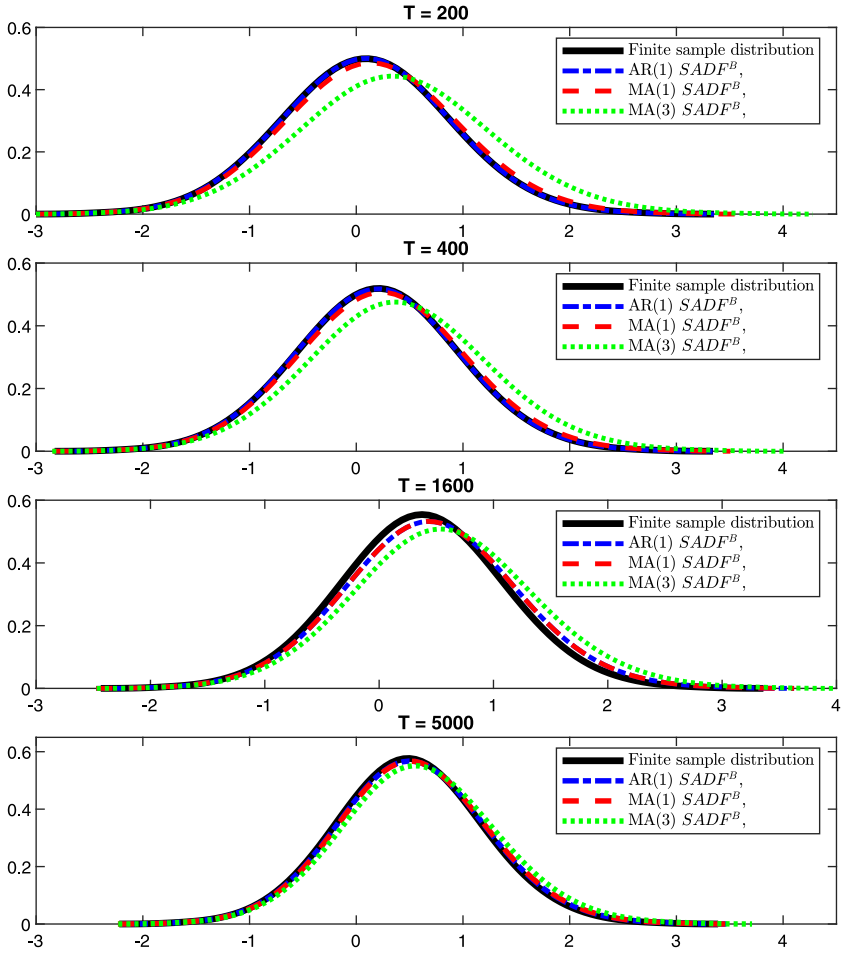

Fig. 3. Bootstrap distribution under the null. The figure shows the distribution of the $SADF$ test under the null and no serial correlation together the bootstrap distribution under the null using (3.1) as data generating process with $d = \eta = \theta = 1$ and $\vartheta_1 = 0.50$, $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, and $\phi_1 = 0.5$, for the MA(1), MA(3) and AR(1) cases, respectively.

4.4. Bootstrap consistency

So far we have shown that the bootstrap algorithm presented in Section 4.1 is able to control size of the SADF and GSADF tests at a reasonable low cost for power. Consistency of the algorithm implies that under A1-A3, the bootstrap distribution should converge to the distribution under the null without any serial correlation as $T \to \infty$. We do not show a formal proof of bootstrap consistency, but to present suggestive evidence of the asymptotic validity of the bootstrap, we generate MA(1), MA(3) and AR(1) series following (3.1) with $\vartheta_1 = 0.50$, $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, and $\phi_1 = 0.5$, respectively, and plot the average bootstrap distribution across 2000 simulations, using 499 bootstrap repetitions. We can then compare the average bootstrap distribution to the finite sample distribution of the $SADF$ test under the null.[9] This is shown in Fig. 3 for sample sizes $T = \{200, 400, 1600, 5000\}$. The figure shows that for the AR(1) and MA(1) cases, the bootstrap distribution is almost exactly the same as the distribution under the null although, and this seems to hold even for $T = 200$. The MA(3) case results in a bootstrap distribution that is both shifted to the right and has fatter tails. However, as the sample size increases the bootstrap distribution converges towards the distribution under the null without serial correlation and at $T = 5000$ the difference between the two distributions is negligible. This reflects the fact that lag augmentation with BIC lag length selection using the whole sample (as the algorithm does) works well for the AR(1) and MA(1) cases, but results in oversizing for the MA(3) case in empirically relevant sample sizes.

It is equally important to show that the bootstrap tests also converge under reasonable choices of the alternative. To show that this is the case we follow the same experiment as above but use (4.8) as data generating process with $y_0 = 0$, $\sigma_v = 1$, $\delta_1 = 1 + c/T$, $\delta_2 = 1 - c/T$, $r_e = 0.4$, $r_c = 0.6$ and $r_x = 0.7$ to match the local asymptotic power curves in the previous section. We set $c = 10$ since

[9] Again, due to computational constraints we only consider the $SADF^B$ test, but unreported results under the null and under the alternative with a smaller number of simulations show that the results also hold for the $GSADF^B$, although convergence happens at a slower rate, particularly in the MA(3) case.

Table 6
Detection frequencies for $BSADF^A$ and $BSADF^B$.

	ϑ_1	ϑ_2	ϑ_3	ϕ_1	$BSADF^A$			$BSADF^B$		
					One	Two	More	One	Two	More
A: Short/long bubble	**0**	**0**	**0**	**0**	0.19	0.70	0.09	0.45	0.48	0.04
	0.5	**0**	**0**	**0**	0.33	0.56	0.05	0.59	0.29	0.03
	0.5	**0.5**	**0.5**	**0**	0.38	0.46	0.06	0.59	0.29	0.03
	0	**0**	**0**	**0.5**	0.37	0.47	0.06	0.55	0.29	0.03
B: Long/short bubble	**0**	**0**	**0**	**0**	0.03	0.81	0.15	0.11	0.66	0.21
	0.5	**0**	**0**	**0**	0.09	0.66	0.20	0.19	0.54	0.21
	0.5	**0.5**	**0.5**	**0**	0.15	0.57	0.21	0.19	0.54	0.21
	0	**0**	**0**	**0.5**	0.13	0.62	0.19	0.24	0.47	0.22
C: Long/long bubble	**0**	**0**	**0**	**0**	0.02	0.83	0.14	0.05	0.77	0.18
	0.5	**0**	**0**	**0**	0.09	0.71	0.19	0.17	0.53	0.28
	0.5	**0.5**	**0.5**	**0**	0.12	0.64	0.20	0.17	0.53	0.28
	0	**0**	**0**	**0.5**	0.11	0.68	0.18	0.18	0.49	0.29

The tests are applied to series generated by (4.8). The sample size is $T = 200$ with initial window set to $r_0 = 0.130$. We set $y_0 = 100$, $\sigma_v = 6.79$, $\delta_1 = 1 + T^{-\alpha}$ and $\delta_2 = 1 - T^{-\alpha}$ with $\alpha = 0.6$. Other parameters are given in the table. Panel A represents the case with a short bubble followed by a long bubble with $r_{e1} = 0.2, r_{c1} = 0.3, r_{x1} = 0.35$ and $r_{e2} = 0.6, r_{c2} = 0.8, r_{x2} = 0.85$. Panel B represents a long bubble followed by a short bubble with $r_{e1} = 0.2, r_{c1} = 0.4, r_{x1} = 0.45$ and $r_{e2} = 0.6, r_{c2} = 0.7, r_{x2} = 0.75$. Panel C represents a long bubble followed by another long bubble with $r_{e1} = 0.2, r_{c1} = 0.4, r_{x1} = 0.45$ and $r_{e2} = 0.6, r_{c2} = 0.8, r_{x2} = 0.85$. We use 4000 replications. The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

this number results in a local asymptotic power of approximately 0.8. The results are shown in Fig. 4. Generally, the picture that emerges under the alternative is similar to the one we have under the null, although convergence towards the distribution under the null is slower. Moreover, in contrast to the case under the null, the bootstrap distribution in the AR(1) and MA(1) cases is also significantly shifted to the right when $T < 1600$. Although we are careful not to claim that our bootstrap algorithm is consistent under all explosive alternatives, the simulations presented here suggest that it is consistent for mild deviations from unity during the explosive and collapse regime.

4.5. Multiple bubbles

A main feature of the GSADF test and its associated date stamping procedure is its ability to detect multiple bubbles. Phillips et al. (2015) show that with a two-bubble data-generating process, BSADF detects two bubbles in 60%–90% of the cases depending on the length of the bubbles. To evaluate the bootstrap test's ability to handle multiple bubbles, we simulate from (4.8), but allow for two bubbles of various length. In contrast to Phillips et al. (2015), we allow for autocorrelated innovations using the same data-generating processes as previously applied. To account for short lived blips in the fitted autoregressive coefficient, we follow the existing literature and impose the condition that for a bubble to exist its duration must exceed $log(T)$. For computational tractability we only consider the sample size $T = 200$. Again, to control for improvements in power due to size distortions, we compare the detection frequencies of the bootstrap test, $BSADF^B$, to those of the size-adjusted BSADF test, $BSADF^A$. Table 6 shows the results. In some cases, $BSADF^A$ and $BSADF^B$ yield very similar results. For example, in the MA(3) case with a long bubble followed by a short bubble, the detection frequencies for $BSADF^B$ are 0.19, 0.54 and 0.21 for one, two or more bubbles, respectively. For $BSADF^A$ the corresponding frequencies are 0.15, 0.57 and 0.21. However, in general the bootstrap test is slightly more conservative than the size-adjusted BSADF test with a higher (lower) detection frequency for one (two) bubble(s). This is in line with the results in Table 5 and Fig. 2.

4.6. Date stamping

An important feature of the recursive right-tailed unit root tests is that we can use these to evaluate when the bubble erupts and subsequently collapses. However, the BSADF test is generally known to provide late estimates of the start as well as end points of the bubble periods, which limits the use of the date stamping feature of the test to pinpoint exact turning points in the data-generating process. To evaluate the extent to which the original and bootstrap versions of the tests differ in terms of providing delayed signals about bubble movements, Table 7 shows the empirical start ($\hat{r}_e$) and end ($\hat{r}_c$) dates in the previously applied single-bubble base case scenario ($r_e = 0.4, r_c = 0.6, r_x = 0.7$) for $T = 200$ (Panel A in Table 5). The results are robust to other sample sizes as well as bubbles appearing earlier or later in the sample period, so for brevity these scenarios are not included in the table.

We design the analysis of delayed signals in the following way. For the bubble start we only consider bubbles that start during or after the actual bubble start date and before the actual bubble end date $\left(r_e \leq \hat{r_e} < r_c\right)$, i.e. we discard spurious bubbles. For a bubble to count as a bubble it has to be of length $log(T)$ or longer. The empirical start date is then the average start date of these non-spurious bubbles. We subsequently compute the empirical end date based on these identified bubbles. We allow the bubble to collapse after the actual collapse date. In a few cases, there are spurious collapses during the bubble period after which the bubble builds up again. To account for this, we in these cases only consider the last bubble period of length $log(T)$ or longer when computing the empirical end date. Again, we compare the bootstrap test $BSADF^B$ to the (infeasible) size-adjusted test $BSADF^A$ to provide a fair comparison free of size distortion.

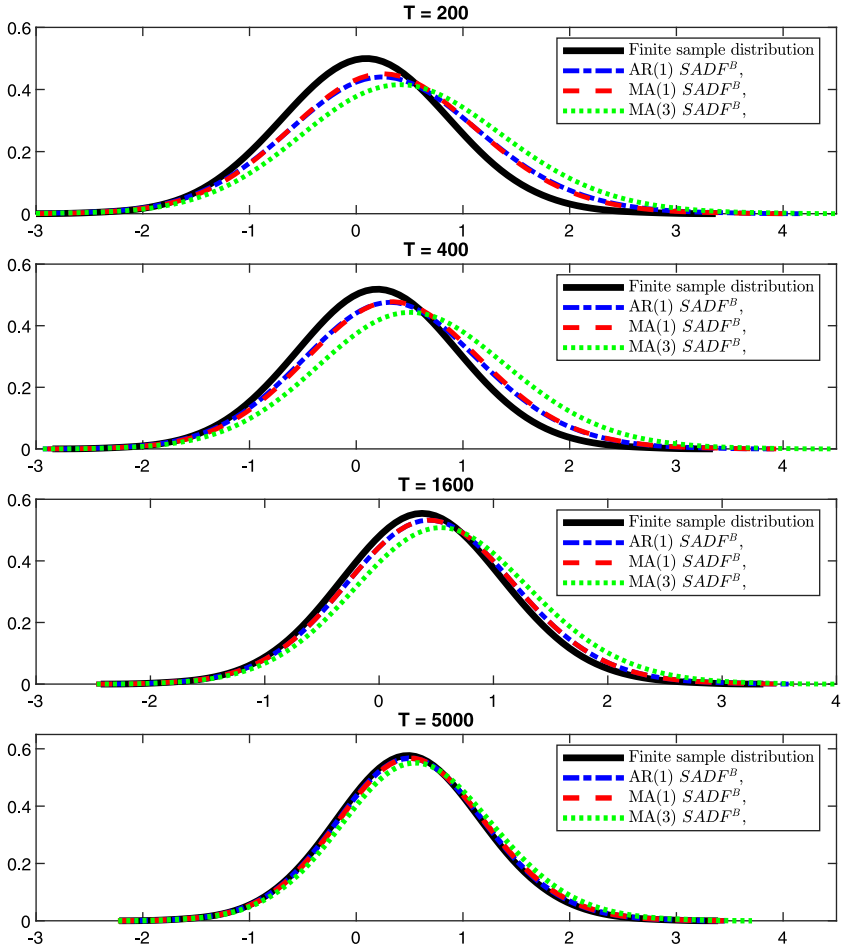

Fig. 4. Bootstrap distribution under the alternative. The figure shows the distribution of the $SADF$ test under the null and no serial correlation together the bootstrap distribution under the alternative using (4.8) as data generating process with $y_0 = 0$, $\sigma_v = 1$, $\delta_1 = 1 + c/T$, $\delta_2 = 1 - c/T$, $r_e = 0.4$, $r_c = 0.6$ and $r_x = 0.7$, and $\vartheta_1 = 0.50$, $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, and $\phi_1 = 0.5$, for the MA(1), MA(3) and AR(1) cases, respectively.

Table 7
Date stamping using $BSADF^A$ and $BSADF^B$.

ϑ_1	ϑ_2	ϑ_3	ϕ_1	A: Bubble start		B: Bubble end	
				$BSADF^A$	$BSADF^B$	$BSADF^A$	$BSADF^B$
0	**0**	**0**	**0**	0.46	0.49	0.64	0.61
0.5	**0**	**0**	**0**	0.47	0.50	0.63	0.61
0.5	**0.5**	**0.5**	**0**	0.48	0.50	0.62	0.60
0	**0**	**0**	**0.5**	0.47	0.49	0.63	0.61

The tests are applied to series generated by (4.8). The sample size is $T = 200$ with initial window set to $r_0 = 0.130$. We set $y_0 = 100$, $\sigma_v = 6.79$, $\delta_1 = 1 + T^{-\alpha}$ and $\delta_2 = 1 - T^{-\alpha}$ with $\alpha = 0.6$. Bubble start and end dates are defined as $r_e = 0.4, r_c = 0.6, r_x = 0.7$. Other parameters are given in the table. Panel A shows the empirical start date, $\hat{r}_e$, and Panel B the empirical end date of the bubble, $\hat{r}_c$. We use 4000 replications. The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

Table 7 confirms that the original test provides a delayed signal both for bubble eruption and collapse. The results are insensitive to whether innovations follow a white noise process or are autocorrelated. Across the four scenarios the average eruption date is 0.47 and the average collapse date is 0.63. Recall that the actual eruption and collapse dates are 0.4 and 0.6, respectively. With respect to $BSADF^B$, the delayed signal is slightly more pronounced with an average eruption date of 0.49, which is related to the small decrease in power as documented in Table 5. In contrast, the bootstrap test more accurately captures the bubble collapse. With an average collapse date of 0.61, $BSADF^B$ provides an almost perfect signal about the end of the bubble. This improvement in the detection of the collapse date seems to be the result of more conservative critical values.

Table 8
GSADF tests for a bubble in the housing market.

Country	T	$k = 1$	p-value	$k = 4$	p-value	BIC	p-value	$GSADF^B$	k^*	p-value
Australia	189	5.56	0.00	6.66	0.00	5.56	0.00	6.30	3.00	0.00
Canada	199	6.07	0.00	4.66	0.00	6.07	0.00	6.07	1.00	0.00
Denmark	198	3.46	0.00	3.58	0.00	3.71	0.00	3.46	1.00	0.02
Finland	198	3.65	0.00	3.18	0.00	3.38	0.00	3.65	1.00	0.02
France	198	4.60	0.00	2.44	0.02	3.23	0.00	2.73	3.00	0.28
Germany	198	4.41	0.00	3.41	0.00	3.66	0.00	3.80	2.00	0.02
Ireland	198	4.54	0.00	4.68	0.00	5.06	0.00	4.54	1.00	0.00
Italy	198	2.06	0.06	1.72	0.13	2.06	0.06	2.13	2.00	0.35
Japan	198	3.15	0.00	3.57	0.00	3.15	0.00	3.15	1.00	0.15
Netherlands	198	6.69	0.00	5.14	0.00	5.33	0.00	5.14	4.00	0.02
New Zealand	198	4.24	0.00	4.31	0.00	4.18	0.00	4.24	1.00	0.01
Norway	163	2.83	0.01	2.73	0.01	2.83	0.01	2.60	2.00	0.23
Portugal	126	3.49	0.00	3.62	0.00	4.06	0.00	3.62	4.00	0.20
Spain	194	3.53	0.00	3.21	0.00	3.02	0.00	4.26	2.00	0.04
Sweden	158	3.17	0.00	5.05	0.00	4.69	0.00	3.17	1.00	0.00
Switzerland	199	4.47	0.00	2.51	0.02	3.31	0.00	2.42	3.00	0.54
United Kingdom	199	2.67	0.01	3.52	0.00	3.82	0.00	2.67	1.00	0.14
United States	199	4.68	0.00	3.79	0.00	3.97	0.00	4.07	2.00	0.02

The table shows the results of the GSADF test for bubbles on the price–rent ratio using fixed lag-lengths of $k = 1$ and $k = 4$ and the BIC to automatically select the variable lag-length with $k_{max} = 6$. It also shows the results of the $GSADF^B$ test and the lag-length used in this test, k^*.

5. Empirical application

In this section we present an empirical application of the bootstrap tests using international housing indices. In Section 5.1 we discuss the relevance of the issue and briefly summarize how the existing literature has dealt with the issue of autocorrelated innovations when testing for bubbles in the housing market. Section 5.2 shows an empirical application of the $GSADF^B$ test and compares the results of this test with the results obtained using the GSADF test with both fixed and transient variable lag selection methods. We limit our empirical application to the GSADF and $GSADF^B$ tests since these tests have a higher detection rate than the SADF and $SADF^B$ tests. Finally, in Section 5.3 we compare the date-stamping results of the BSADF and $BSADF^B$.

5.1. Past evidence on housing bubbles

International housing markets have, following the boom and bust of house prices that contributed to the 2008–09 global financial crisis, received a lot of attention since it is not entirely clear whether or not these dynamics were associated with speculative bubbles or only the result of changing fundamentals. This has led researchers to use the SADF and GSADF tests to investigate the possibility of speculative bubbles in the housing market. Aware of the high degree of serial correlation in housing indices, researchers have attempted to accommodate the issue with the usual augmentation of the Dickey–Fuller regressions that conform the tests, using either a fixed lag-length or automatic variable lag selection methods. However, as shown in the finite sample simulation studies, these solutions tend to result in extremely oversized tests and thus in spurious bubbles.

Pavlidis et al. (2015) utilize the SADF and GSADF tests to look for episodes of exuberance in the housing markets of 22 OECD countries using the real price, price–rent and price–income ratios. Using the GSADF test with a fixed lag of $k = 4$ on price–rent ratios they reject the null of "no bubbles" in all but three of the 22 countries. In a recent paper, Shi et al. (2016) use the BSADF test to date stamp the timeline of house price bubbles in Australian capital cities using the price–rent ratio. They use the BIC to select the lag-length (with $k_{max} = 6$), and find evidence of explosive bubbles in all major Australian cities. Caspi (2015) applies the SADF and GSADF tests to the price–rent ratios of regional housing markets in Israel to test for bubbles. Caspi (2015) uses fixed lag-lengths between one and six as well as the AIC and BIC with maximum lag-length of 12, and is unable to reject the null of "no bubbles" in the majority of the regions, but he notes that the results of the Gush Dan region are highly sensitive to the lag specification.

5.2. Testing for bubbles using the GSADF and $GSADF^B$ tests

We use the official OECD data for 18 countries: Australia, Canada, Denmark, Finland, France, Germany, Ireland, Italy, Japan, the Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, the UK and the US. The data set contains seasonally adjusted quarterly observations of house price–rent ratios that generally span from the early 1970s to 2019. For a few countries the sample period starts a bit later; see Table 1 for the exact sample period for the individual countries. Given the results presented in Table 1, the presence of AR and MA components is already established. Since the average sample size among all included countries is $T = 189$, the results of the finite sample simulation study with $T = 200$ are the most relevant as points of comparison.

Table 8 presents the results of the GSADF tests for a bubble in the price–rent ratio using fixed lag-lengths of $k = 1$ and $k = 4$ and using the BIC to automatically select the transient lag-length with a maximum lag of six. The table also shows corresponding p-values, which are calculated using 5000 replications of a random walk with iid Gaussian innovations. The last three columns show the results of the $GSADF^B$ test, the lag-length used in the bootstrap test, k^*, and the bootstrap p-value, which is calculated

using $M^* = 5000$ bootstrap replications. Note that the GSADF test statistic with a fixed lag-length and the $GSADF^B$ test statistic will always be equal when $k = k^*$. This is, however, not necessarily the case for the transient variable lag-length version of the test since this version of the test allows each of the ADF regressions to select a different k.

Using a 5% significance level, we reject the null of "no bubbles" in all countries but Italy, irrespective of using a fixed lag-length of $k = 1$ or $k = 4$ or using the BIC to automatically select the transient lag-length with $k_{max} = 6$. In contrast, when we apply the $GSADF^B$ test to the same data, we fail to reject the null for France, Italy, Japan, Norway, Portugal, Switzerland and the UK. With a significance level of 1% instead of 5%, we maintain the null also in Denmark, Finland, Germany, the Netherlands, Spain and the US using the bootstrap test. With a fixed lag-length of $k = 1$ only the UK is added to the list of countries with no evidence of bubbles, while no countries are added when using the BIC to automatically select the transient lag-length. The simulation results in Section 4 suggest that these differences in conclusion across the original and bootstrap tests mainly come from the latter test being better sized although we cannot rule out that a small power loss is the culprit.

5.3. Date-stamping bubbles

To evaluate the difference between BSADF and $BSADF^B$ in terms of date-stamping, Figs. 5, 6, and 7 show the price–rent ratio for all 18 countries with shaded areas denoting bubble periods detected using a 5% significance level. For BSADF we only show the results with a fixed lag-length of $k = 1$. The online appendix contains the results with a fixed lag-length of $k = 4$ and a variable lag-length selected by the BIC with $k_{max} = 6$. To remain consistent with Phillips et al. (2015) and the simulation results in Sections 4.5 and 4.6, we only consider bubble periods of length $log(T)$ or longer. Without the $log(T)$ rule we would see many short lived bubbles, but interestingly this would be much less pronounced with the bootstrap test compared to the original test both with fixed and variable lag-length. This follows naturally from the bootstrap test being slightly more conservative than the original test. Fig. 1 also shows that an MA(3) process for innovations can well replicate the Norwegian price–rent ratio even without bubbles. Finally, we

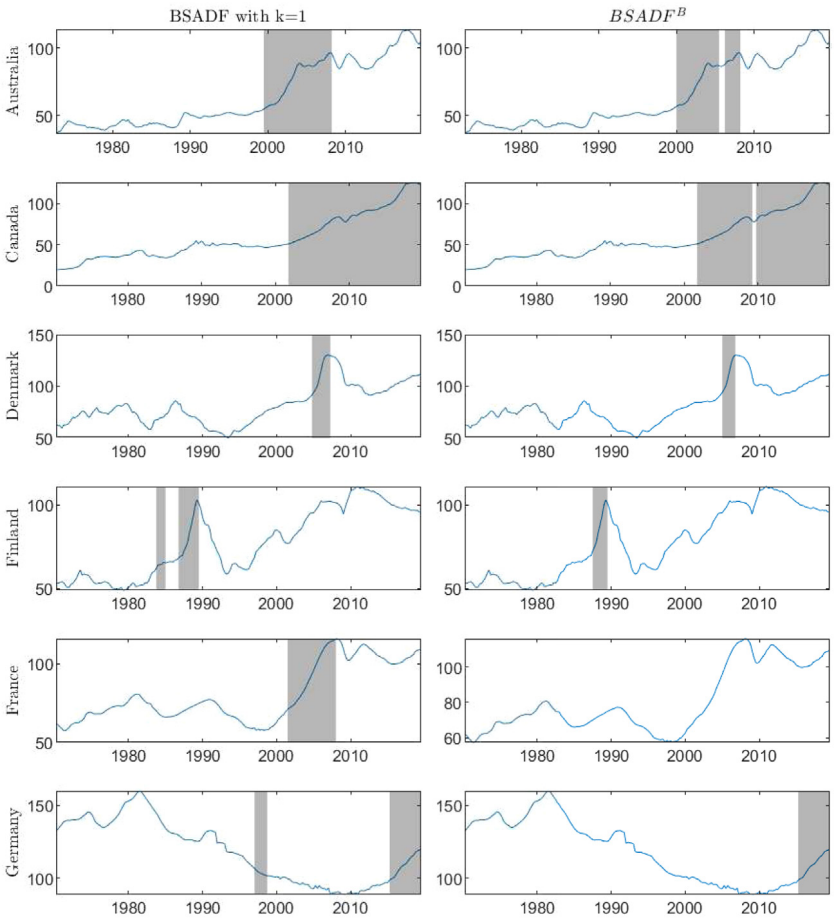

Fig. 5. Date-stamping of housing bubbles. price–rent ratio for Australia, Canada, Denmark, Finland, France and Germany. The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (left panel), and by the $BSADF^B$ test (right panel).

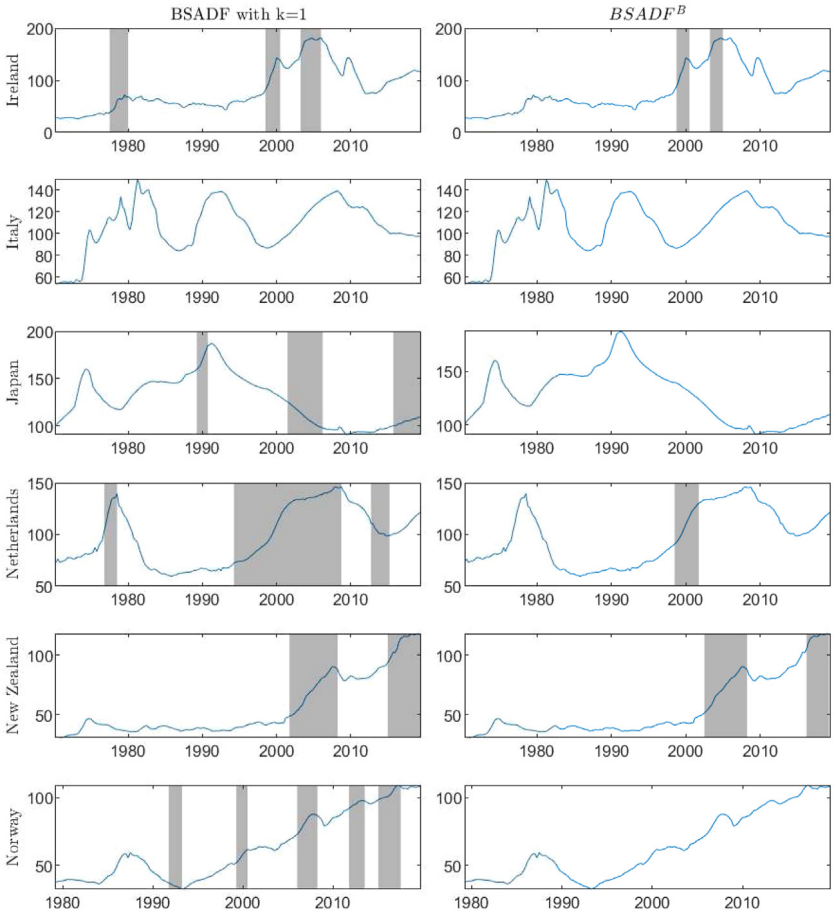

Fig. 6. Date-stamping of housing bubbles. price–rent ratio for Ireland, Italy, Japan, Netherlands, New Zealand and Norway. The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (left panel), and by the $BSADF^B$ test (right panel).

only apply the date-stamping procedure in the cases where we reject the null using the GSADF or $GSADF^B$ test, respectively. The logic here is that it only makes sense to date-stamp bubbles, when these have been detected in the first place.

From Figs. 5, 6, and 7 it is clear that there generally are much fewer dates with explosiveness in the price–rent ratio according to $BSADF^B$ compared to BSADF. In some cases the explanation might be loss of power. This could, for example, be the case for Japan and Switzerland. Both countries experienced a rapid increase in house prices around 1990 but here only the original test detects a bubble. Since the build-up in prices is relatively short-lived and we know from Section 4.3 that the power of the bootstrap test generally decreases the shorter the bubble is, we cannot rule out that the bootstrap test here erroneously maintains the null of "no bubble". Other cases point towards an oversized original test. For example, Norway where innovations follow an MA(3) process has according to BSADF suffered from multiple bubbles during the sample period, while $GSADF^B$ does not reject the null of "no bubbles". A visual inspection of the price–rent ratio does not show clear evidence of explosivity, which lends support to the conclusion reached by the bootstrap test. Other examples include the UK and last part of the sample for Switzerland, where the price–rent ratio clearly does not display explosive behavior.

There are a number of other results from the date-stamping analysis that are worth noticing. First, according to Table 7 $BSADF^B$ provides a slightly more delayed signal about bubble eruption compared to BSADF, but in contrast the bootstrap test provides a more accurate signal about bubble collapse. These results appear to be relevant for the US and Denmark. For both countries, the bubble erupts a bit later for $BSADF^B$ than for BSADF. In contrast, the bootstrap test accurately pinpoints the bubble collapse as when the explosive behavior cease to exist, while the original test continues to provide a bubble signal after the price–rent ratio flattens out or even starts to decline. This is also relevant for the Netherlands, where the bootstrap test detects a bubble that spans the period 1998Q-2001Q3, while according to the original test a bubble was among other periods present from 1994Q1 to 2008Q3. A visual inspection does not suggest an explosive price–rent ratio from around 2001 to 2008, which again points towards an oversized original test. Second, BSADF often detects downward trending explosivity. This is the case in Germany, Japan, the Netherlands, Norway and Portugal. Given the restrictions on rational bubbles (Diba and Grossman, 1988), we cannot interpret such periods as bubble periods since rational bubbles cannot be negative. This is not a flaw in the original test as the test allows for both positive

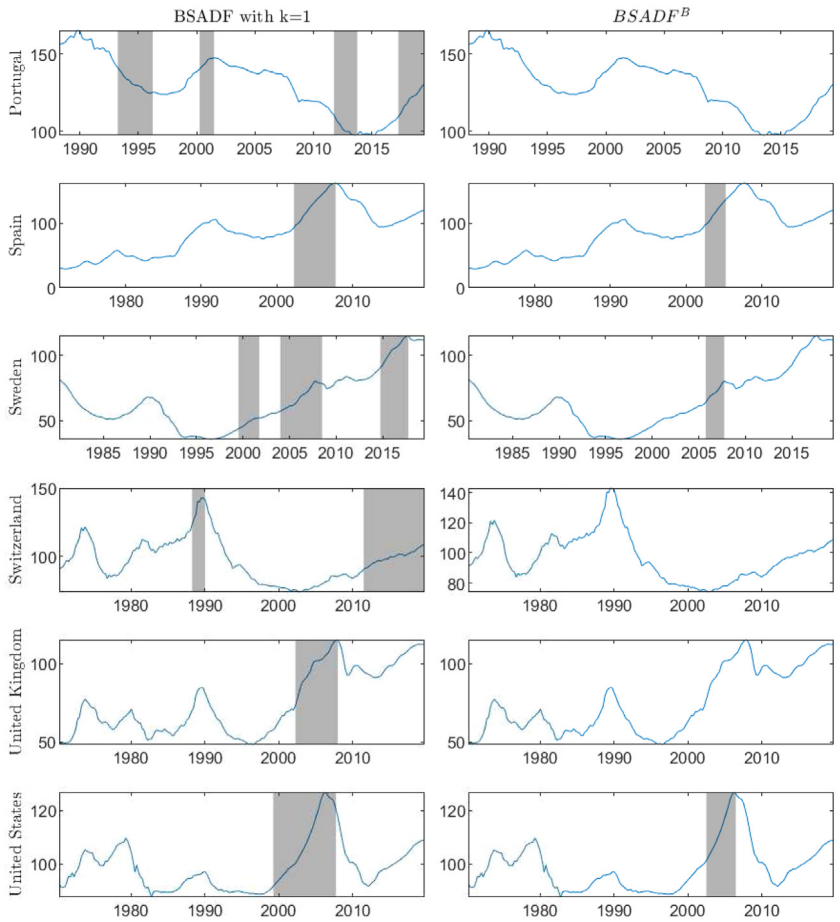

Fig. 7. Date-stamping of housing bubbles. price–rent ratio for Portugal, Spain, Sweden, Switzerland, United Kingdom and United States. The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (left panel), and by the $BSADF^B$ test (right panel).

and negative explosive behavior. It just implies that when applying the tests, we need to be careful and evaluate to what extent the overall conclusion about bubbles is caused by negative explosive behavior in which case the conclusion is invalid. Interestingly, the bootstrap test does not detect such downward trending explosive periods. Finally, evaluating the Netherlands, we could suspect that the late 1970s, which is classified as a bubble according to BSADF but not $BSADF^B$ is an example of the bootstrap test's conservatism in terms of capturing multiple bubbles, cf. Table 6. Disregarding the period with downward trending explosivity as captured by the original test, BSADF detects two bubble periods in the Netherlands and $BSADF^B$ only one. However, had we not applied the $log(T)$ rule in terms of the length of the bubble, $BSADF^B$ would also have detected a bubble in the late 1970s. The bubble is one quarter too short and hence discarded according to $log(T)$ rule. The conservatism displayed by the bootstrap test in Table 6 can thus to a large extent be explained by the $log(T)$ rule combined with shorter bubble periods, cf. Table 7.

6. Concluding remarks

Bubble testing is currently not only at the top of the research agenda, but following the surge and collapse in both stock and house prices in recent years and the subsequent financial crisis it is also at the center of attention in, for example, financial institutions and central banks and among policymakers. In this paper, we analyze an empirically important issue with the most often used bubble tests, namely the recursive right-tailed unit root tests by Phillips et al. (2011, 2015). We show that serially correlated innovations (which is often found empirically for time series used in bubble tests) can lead to severe size distortions when using either fixed or automatic (based on information criteria) lag-length selection in the auxiliary regressions underlying the test. We propose a sieve bootstrap version of the tests and show that this results in tests which control size well across a number of simulation designs both with and without highly autocorrelated innovations. More importantly, these size corrections come at a relatively low cost for the power of the tests.

Applied to the price–rent ratio in 18 OECD countries, we find weaker evidence of bubbles in the housing market using the bootstrap version of the test compared to both fixed and automatic lag-length selection. While 17 price–rent ratios are concluded to

be explosive on a 1% significance level using the BIC to select the transient lag-length, only 11 price–rent ratios display explosive behavior on a 5% level according to the bootstrap version of the test. With a 1% significant level this number drops to five countries.

CRediT authorship contribution statement

Thomas Quistgaard Pedersen: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - original draft, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Erik Christian Montes Schütte:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jempfin.2020.06.002.

References

Agiakloglou, C., Newbold, P., 1992. Empirical evidence on dickey-fuller-type tests. J. Time Series Anal. 15, 253–262.
Caspi, I., 2015. Testing for a housing bubble at the national and regional level: the case of Israel. Empir. Econ. 51, 483–516.
Cavaliere, G., Nielsen, H.B., Rahbek, A., 2018. Bootstrapping noncausal autoregressions: with applications to explosive bubble modeling. J. Bus. Econom. Statist. 1–13.
Cavaliere, G., Robert Taylor, A., 2009. Bootstrap M unit root tests. Econometric Rev. 28, 393–421.
Chang, Y., Park, J.Y., 2002. On the asymptotics of ADF tests for unit roots. Econometric Rev. 21, 431–447.
Chang, Y., Park, J.Y., 2003. A sieve bootstrap for the test of a unit root. J. Time Series Anal. 24, 370–400.
Diba, B.T., Grossman, H.I., 1988. Explosive rational bubbles in stock prices? Amer. Econ. Rev. 78, 520–530.
Engsted, T., Hviid, S.J., Pedersen, T.Q., 2016. Explosive bubbles in house prices? evidence from the OECD countries. J. Int. Financ. Mark. Inst. Money 40, 14–25.
Engsted, T., Nielsen, B., 2012. Testing for rational bubbles in a coexplosive vector autoregression. Econom. J. 15, 226–254.
Evans, G.W., 1991. Pitfalls in testing for explosive bubbles in asset prices. Amer. Econ. Rev. 81, 922–930.
Figuerola-Ferretti, I., McCrorie, J.R., 2016. The shine of precious metals around the global financial crisis. J. Empir. Financ. 38, 717–738.
Ghysels, E., Plazzi, A., Valkanov, R., Torous, W., 2013. Forecasting real estate prices. Handb. Econ. Forecast. 2, 509–580.
Gouriéroux, C., Zakoïan, J.-M., 2017. Local explosion modelling by non-causal process. J. R. Stat. Soc. Ser. B Stat. Methodol. 79 (3), 737–756.
Gutierrez, L., 2011. Bootstrapping asset price bubbles. Econ. Model. 28 (6), 2488–2493.
Harvey, D.I., Leybourne, S.J., Sollis, R., 2015. Recursive right-tailed unit root tests for an explosive bubble. J. Financ. Econ. 13, 166–187.
Harvey, D.I., Leybourne, S.J., Sollis, R., Taylor, R., 2016. Tests for explosive financial bubbles in the presence of non-stationary volatility. J. Empir. Financ. 38, 548–574.
Homm, U., Breitung, J., 2012. Testing for speculative bubbles in stock markets: a comparison of alternative methods. J. Financ. Econ. 10, 198–231.
Kivedal, B.K., 2013. Testing for rational bubbles in the US housing market. J. Macroecon. 38, 369–381.
Kraussl, R.L., Tussl, R., Lehnert, T., Martelin, N., 2016. Is there a bubble in the art market? J. Empir. Financ. 35, 99–109.
Ng, S., Perron, P., 1995. Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag. J. Amer. Statist. Assoc. 90, 253–268.
Ng, S., Perron, P., 2001. Lag length selection and the construction of unit root tests with good size and power. Econometrica 69 (6), 1519–1554.
Palm, F.C., Smeekes, S., Urbain, J.P., 2008. Bootstrap unit-root tests: comparison and extenstions. J. Time Series Anal. 29, 371–401.
Park, J.Y., 2003. Bootstrap unit root tests. Econometrica 71, 1845–1895.
Pavlidis, E., Yusupova, A., Paya, D., Peel, D., Martinez-Garcia, E., Mack, A., Grossman, V., 2015. Episodes of exuberance in housing markets: In search of the smoking gun. J. Real Estate Finance Econ. 53, 419–449.
Phillips, P.C.B., Perron, P., 1988. Testing for a unit root in time series regression. Biometrika 75, 335–346.
Phillips, P.C.B., Shi, S.P., 2018. Financial bubble implosion and reverse regression. Econometric Theory 34 (4), 705–753.
Phillips, P.C.B., Shi, S.P., Yu, J., 2015. Testing for multiple bubbles: historical episode of exuberance and the collapse in the S&P 500. Internat. Econom. Rev. 56, 1043–1078.
Phillips, P.C.B., Wu, Y., Yu, J., 2011. Explosive behavior in the 1990s NASDAQ: When did exhuberance escalate asset values? Internat. Econom. Rev. 52 (1), 201–226.
Phillips, P.C.B., Yu, J., 2011. Dating the timeline of financial bubbles during the subprime crisis. Quant. Econ. 2, 455–491.
Said, S.E., Dickey, D.A., 1984. Testing for unit roots in autoregressive-moving average models of uknown order. Biometrika 71, 599–607.
Schwert, G.W., 1989. Tests for unit roots: a monte carlo investigation. J. Bus. Econom. Statist. 7, 147–159.
Shi, S.P., Valadkhani, A., Smyth, R., Vahid, F., 2016. Dating the timeline of house price bubbles in Australian capital cities. Econ. Rec. 92, 590–605.
Zivot, E., Andrews, D.W.K., 2002. Further evidence on the great crash, the oil-price shock, and the unit-root hypothesis. J. Bus. Econ. Stat. 20 (1), 25–44.