Testing for Explosive Bubbles in the Presence of Autocorrelated Innovations

Thomas Quistgaard Pedersen and
Erik Christian Montes Schütte

CREATES Research Paper 2017-9

Testing for Explosive Bubbles in the Presence of Autocorrelated Innovations[*]

Thomas Quistgaard Pedersen[†] Erik Christian Montes Schütte[‡]

February 14, 2017

Abstract

We analyze an empirically important issue with the 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 more or less perfectly sized test statistics even in the presence of 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 less strong evidence of bubbles compared to existing evidence.

JEL Classification: C58, G12

Keywords: Right-tailed unit root tests, GSADF, size and power properties, sieve bootstrap, international housing market

[*] 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. We thank Jörg Breitung, Bent Jesper Christensen, Tom Engsted, Niels Haldrup and participants at a CREATES seminar and the 10th International Conference in Computational and Financial Econometrics for valuable comments.

[†] Department of Economics and Business Economics and CREATES, Aarhus University, Denmark, Email: tqpedersen@econ.au.dk.

[‡] Department of Economics and Business Economics and CREATES, Aarhus University, Denmark, Email: christianms@econ.au.dk.

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 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), Homm and Breitung (2012), Engsted and Nielsen (2012), Phillips et al. (2015), and Harvey et al. (2015b). 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) and Phillips et al. (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) and Phillips et al. (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 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. (2015b) consider the case with non-stationary volatility and show that the SADF test is severely over-sized.

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) and Phillips et al. (2015), which to our knowledge has not yet been addressed in the literature. 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. A large literature deals with the effect of serially correlated innovations and lag-length selection in standard unit root tests; for example, Schwert (1989) and Ng and Perron (1995, 2001). 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

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

GSADF tests are i) we test against the explosive and not the stationary alternative, and ii) the test statistics are computed as the supremum of a sequence of unit root tests. The importance of analyzing the effect of serially correlated innovations and lag-length selection is emphasized by the fact that it is not possible to reject the presence of autoregressive and moving average components in the first difference of many of the time series used in bubble tests, such as the price-dividend ratio in case of stock markets and the price-rent ratio in case of housing markets. Furthermore, the recursive right-tailed unit root tests entail performing unit root tests on many subsamples of the relevant data, some of which will be very small and likely to lead to size problems.

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) and Phillips et al. (2015). Size distortions decrease with the sample size, 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 is not there. For example, Phillips et al. (2015) provide an empirical illustration based on the monthly U.S. price-dividend ratio over the period 1871-2010 and with no lags in the auxiliary regression used to compute the test statistic. With a similar sample size and one moving average component with a magnitude corresponding to empirical findings, the size of the GSADF test is 0.36, when using no lags in the auxiliary regression.[2] In case of one moving average component, the tests are only reasonably sized with one lag in the auxiliary regression. In practice, it is difficult to determine the order of moving average components with test procedures that require multiple unit root tests on subsamples of the data. One potential solution is the use of information criteria such as the Bayesian Information Criterion. However, we show that this also leads to severe size distortions. Instead, motivated by Park (2003), Chang and Park (2003) and Palm et al. (2008) we suggest the use of a sieve bootstrap to restore the size of the recursive right-tailed unit root tests in the presence of serially correlated innovations. We show that such a bootstrap procedure leads to almost perfectly sized tests and that these size corrections come at a relatively low cost in terms of power.

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 automatic transient lag-length selection. 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 all 17 countries in our sample using a 1% significance level. In contrast, with the bootstrap test only 9 out of 17 countries are subject to bubbles using a 5% significance level. With a 1% significance level the number drops to 5 countries.

[2] This choice of lag-length is motivated by a small simulation study conducted by Phillips et al. (2015) based on serially uncorrelated innovations. They generally find that the SADF and GSADF tests have best size properties with either no or one lag in the auxiliary regression.

The rest of the paper is organized as follows. Section 2 gives a brief review of the rational bubble model, which is the theoretical backbone of rational bubble tests, while section 3 provides an overview of the recursive right-tailed test procedures, we consider in this paper. Section 4 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 5 describes the sieve bootstrap version of the tests and contains a simulation study analyzing their size and power. Section 6 provides an empirical application of the sieve bootstrap tests and section 7 some concluding remarks.

2 The rational bubble model

We begin by considering the standard asset pricing model, in which the value of an asset is determined by the discounted price of the asset in the next period, P_{t+1}, plus its dividend or service flow, X_{t+1}, (fundamentals henceforth). Assuming constant expected return, $R > 0$, the price of the asset is given by

$$P_t = \frac{1}{1+R}\mathbb{E}_t(P_{t+1} + X_{t+1}), \tag{2.1}$$

where $\mathbb{E}_t$ is the expectations operator conditional on time t information. Using forward recursive substitution and the law of iterated expectations, we arrive at the following general solution for (2.1)

$$P_t = \sum_{i=1}^{\infty} \left(\frac{1}{1+R}\right)^i \mathbb{E}_t X_{t+i} + B_t, \tag{2.2}$$

where the first term is a fundamental value component, $P_t^f = \sum_{i=1}^{\infty} \left(\frac{1}{1+R}\right)^i \mathbb{E}_t X_{t+i}$, and, $B_t = \left(\frac{1}{1+R}\right) \mathbb{E}_t B_{t+1}$, is a rational bubble component reflecting self-fulfilling expectations, i.e. the bubble only exists in period t because it is expected to exist in the next period. We can express the data generating process for the bubble as

$$B_{t+1} = (1+R)B_t + \xi_{t+1}, \tag{2.3}$$

where ξ_{t+1} is a zero-mean rational forecast error. Consequently, since $R > 0$, B_t will be an explosive process.[3]

If we impose the ("no-bubble") transversality condition, $\underset{T\to\infty}{lim}(1+R)^{-T}\mathbb{E}_t P_{t+T} = 0$, when using forward recursive substitution in (2.1), $P_t = P_t^f$. This means that the properties of the time series with and without the bubble will be different and in some sense, testing for bubbles is implicitly testing the transversality condition. To see why this is the case, we can start by making the common, and often plausible assumption, that the process driving fundamentals is a random walk with drift,

[3] Note that (2.3) also means that the presence of a bubble does not result in riskless arbitrage opportunities.

$$X_{t+1} = \mu + X_t + \varepsilon_{t+1}, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_X^2). \tag{2.4}$$

It then follows that

$$P_t^f = \frac{(1+R)\mu}{R^2} + \frac{X_t}{R}, \tag{2.5}$$

which implies that if X_t is a random walk with drift, P_t^f will also be a random walk with drift. Thus, one intuitive way to test for bubbles is to see if prices are best characterized by an explosive autoregressive process or by a random walk. There is, however, one caveat to using prices when testing for bubbles. This is because prices can become explosive, even without a bubble, if fundamentals are transiently explosive. For this reason, empirical researchers often work with the ratio of prices to fundamentals:

$$\frac{P_t}{X_t} = \sum_{i=1}^{\infty} \left(\frac{1}{1+R}\right)^i \frac{\mathbb{E}_t X_{t+i}}{X_t} + \frac{B_t}{X_t}, \tag{2.6}$$

which will only be explosive if there is a bubble, $B_t > 0$. In other words, the price-fundamentals ratio, P_t/X_t, automatically controls for explosiveness in fundamentals.

More generally, if there is no bubble ($B_t = 0$) and fundamentals follow a random walk (with drift), there are two possibilities: i) P_t and X_t have a common stochastic $I(1)$ trend and as a result, the price-fundamentals ratio is stationary (Craine, 1993); and ii) P_t and X_t do not cointegrate, which implies that the price-fundamentals ratio is an $I(1)$ process. The bubble tests we consider here build on the null hypothesis that the relevant series is an $I(1)$ process. With the price-fundamentals ratio as test series, this can be seen as a relatively conservative assumption since a rejection of the $I(1)$ null against explosiveness will have even higher discriminatory power against $I(0)$ behavior.[4]

3 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 look for exuberance in the stock market. They utilize unit root tests but instead of testing the unit root null against the stationarity alternative, they look at the right-tail of the distribution and test against the explosive alternative. A problem with this approach is pointed out by Evans (1991) who uses simulations to show that unit root tests have low power when trying to detect periodically collapsing bubbles.

Phillips et al. (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

[4] In practice it is commonly found that the price-fundamentals series, whether in the stock market or in the housing market, have a unit root. Indeed, using a battery of unit root tests, we were unable to reject the null of a unit root (against the alternative of stationarity) in the price-dividend and price-rent series presented below.

run, where the largest of these test statistics is used to test for explosiveness. They named this method the Supremum Augmented Dickey Fuller (SADF) test and show that it does not only result in much greater power - even in the presence of periodically collapsing bubbles - but also allows us to pinpoint the start and ending date of the bubble. Phillips et al. (2015) develop a generalized version of the SADF test (GSADF) by allowing both the starting and ending date of the sample window to vary. They find that the GSADF test has an even greater power in detecting periodically collapsing bubbles. We describe the SADF and GSADF tests in more detail in the following section.

3.1 The SADF

Consistent with the ideas presented in the preceding section, the null hypothesis of the SADF test is that the series in question follows a random walk with asymptotically negligible drift,

$$y_t = dT^{-\eta} + \theta y_{t-1} + \varepsilon_{t}, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2_{r1,r2}), \tag{3.1}$$

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 regressions underpinning the test. The methodology proposed by Phillips et al. (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

$$\Delta y_t = \alpha_{r_1,r_2} + \beta_{r_1,r_2} y_{t-1} + \sum_{i=1}^{k} \psi^i_{r_1,r_2} \Delta y_{t-i} + \varepsilon_t \tag{3.2}$$

where k is the lag order and the subscripts r_1,r_2 indicate that the fractional regression window used starts at r_1 and ends at r_2. More specifically, since the SADF fixes the start point at 0 the first regression will have a sample size of $\lfloor Tr_0 \rfloor$, where $\lfloor . \rfloor$ denotes the floor function, and expand 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 $ADF^{r_2}_{r_1}$ and the supremum of this sequence will define the SADF test statistic

$$SADF(r_0) = \underset{r_2 \in [r_0, 1]}{sup} \{ADF^{r_2}_0\} \tag{3.3}$$

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

3.2 The GSADF

The GSADF test builds upon the idea of its precursor but in contrast to the SADF, this test allows both the starting, r_1, and ending, r_2, points of the sample window to vary. Thus, for a given r_0, a double recursion scheme is performed 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 same data-generating process for the null (3.1) and empirical regression model (3.2), the GSADF 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

$$GSADF(r_0) = \underset{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}}{sup} \left\{ADF_{r_1}^{r_2}\right\}. \tag{3.4}$$

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

3.3 The date-stamping of bubbles

As mentioned, one of the advantages of these recursive types of tests is that they allow us to pinpoint the origin and collapse of the bubble. The date-stamping algorithm works by performing SADF tests in a backward expanding sample sequence, where the ending 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 will define the BSADF

$$BSADF_{r_2}(r_0) = \underset{r_1 \in [0, r_2 - r_0]}{sup} \left\{ADF_{r_1}^{r_2}\right\} \tag{3.5}$$

We can infer whether or not observation τ is part of the bubble by comparing the $BSADF$ test statistic for that observation to its corresponding critical value (based on a sample size of Tr_2). Although our finite sample simulation study does not explicitly cover the BSADF test, we do analyze the size and power properties of the SADF 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. In our empirical application we compare the performance of the BSADF test and a bootstrap version of it.

4 Testing for bubbles with autocorrelated innovations

The null hypothesis (3.1) of the SADF and GSADF tests, assumes that the time series under consideration is generated by a pure unit-root process with an asymptotically negligible drift. Nonetheless, the sensitivity of these tests to a violation of the assumption of serially uncorrelated innovations has to our knowledge not yet been explored. If the time series has a data-generating process where innovations are serially correlated instead of white noise, the critical values simulated under (3.1) can be misleading. This section examines the effect that autoregressive and moving average components of varying magnitude have on the SADF and GSADF tests at different lag-lengths in the auxiliary regressions. Section 4.1 describes the motivation and basic framework. Section 4.2 explores the effects of autocorrelated errors when the auxiliary regressions use a fixed lag-length, while section 4.3 analyzes the effectiveness of information criteria in controlling the size of the tests under these circumstances.

4.1 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. This model was extended by Said and Dickey (1984) who augmented the Dickey-Fuller regression with lagged differences of the series, resulting in the augmented Dickey-Fuller (ADF) regression presented in (3.2). The theoretical underpinning of this regression is that an ARMA(p,q) model of unknown order can be satisfactorily approximated by an AR(k) process, where $k = O(T^{1/3})$. Chang and Park (2002) show that this approximation will hold for a general class of linear models and that, assuming $k \to \infty$ as $T \to \infty$, the ADF test statistic will have the same limiting distribution as the simple Dickey-Fuller t-statistic. The issue then becomes one of selecting the right autoregressive lag-length, k, in (3.2) 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 over-sized if k is too small, while low power ensues if k is too large. 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; Ng and Perron 1995). While the effects of heteroskedastic innovations on the SADF and GSADF tests have already been analyzed by Phillips et al. (2015) and Harvey et al. (2015b), the effects of serially correlated error terms and the impact of varying truncation lags under these circumstances have not previously been examined, at least to our knowledge. The relevance of such an analysis is emphasized by the fact that the series usually employed to test for bubbles, namely price-dividend and price-rent ratios, where the latter pertains to housing markets, are commonly found to contain non-negligible autoregressive and moving average components.

Using the BIC to determine the presence and order of ARMA components on the first differences of the annual price-dividend series 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 series 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 and price-rent indices) will 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 differences of price-rent ratios, $\Delta(P_t/X_t)$, based on housing data collected for the OECD, we find that the most common models are either MA(3) or AR(1) models. Table 4.1 shows the autoregressive and moving average coefficients and t-statistics for selected indices.

Table 4.1: ARMA(p,q) model fit and coefficients for price-rent ratios

Country	Period	T	ARMA(p,q) coefficient / *(t-statistic)*					
			constant	AR(1)	AR(2)	MA(1)	MA(2)	MA(3)
Australia	1972Q2 - 2016Q2	175	0.43	–	–	0.74	0.80	0.56
			(1.90)	–	–	*(11.67)*	*(13.30)*	*(8.81)*
Canada	1970Q2 - 2016Q2	185	0.60	0.45	–	–	–	–
			(0.45)	*(6.69)*	–	–	–	–
Germany	1970Q2 - 2016Q2	185	-0.14	0.38	0.33	–	–	–
			(-0.51)	*(5.38)*	*(4.68)*	–	–	–
Denmark	1970Q2 - 2016Q2	185	0.22	–	–	0.66	0.63	0.38
			(0.70)	–	–	*(9.76)*	*(9.00)*	*(5.65)*
Spain	1971Q2 - 2016Q2	181	0.32	0.62	0.22	–	–	–
			(0.65)	*(8.41)*	*(3.03)*	–	–	–
Finland	1970Q2 - 2016Q2	185	0.56	–	–	0.67	0.68	0.53
			(2.35)	–	–	*(10.62)*	*(10.99)*	*(8.43)*
France	1970Q2 - 2016Q2	185	0.20	0.73	–	0.09	0.47	–
			(1.20)	*(11.47)*	–	*(1.11)*	*(6.53)*	–
United Kingdom	1968Q3 - 2016Q2	192	0.30	–	–	0.70	0.66	0.39
			(1.09)	–	–	*(9.05)*	*(10.22)*	*(5.85)*
Italy	1970Q2 - 2016Q2	185	0.19	–	–	0.75	0.55	0.32
			(0.40)	–	–	*(10.70)*	*(6.89)*	*(4.47)*
Ireland	1970Q2 - 2016Q2	185	0.35	0.61	–	–	–	–
			(0.59)	*(10.40)*	–	–	–	–
Japan	1970Q2 - 2016Q2	185	-0.01	0.75	–	0.13	0.37	–
			(-0.01)	*(12.21)*	–	*(1.53)*	*(4.78)*	–
Norway	1979Q2 - 2016Q3	150	0.56	–	–	0.48	0.55	0.30
			(2.35)	–	–	*(6.08)*	*(7.28)*	*(3.77)*
Netherlands	1970Q2 - 2016Q2	185	0.17	0.60	0.19	-0.42	0.43	–
			(0.38)	*(3.94)*	*(1.29)*	*(2.82)*	*(2.97)*	–
New Zealand	1970Q2 - 2016Q2	185	0.56	–	–	0.60	0.74	0.53
			(2.35)	–	–	*(9.31)*	*(12.02)*	*(7.67)*
Sweden	1980Q2 - 2016Q2	145	0.30	0.78	–	–	–	–
			(0.69)	*(14.95)*	–	–	–	–
Switzerland	1970Q2 - 2016Q2	185	0.04	0.76	–	-0.58	0.37	–
			(0.99)	*(10.41)*	–	*(-6.66)*	*(5.04)*	–
United States	1970Q2 - 2016Q2	185	0.04	-0.01	0.54	0.53	-0.13	0.34
			(0.40)	*(6.34)*	*(6.60)*	*(6.05)*	*(6.05)*	*(6.05)*

ARMA(p,q) fit for the first differences in price-rent indices. The optimal p and q were selected by the BIC using a maximum AR order of p_{max} =2 and maximum MA order of $q_{max} = 5$.

In our simulation study of the impact of autocorrelated innovations on right-tailed unit root tests, we start from the null hypothesis given in (3.1) and calculate finite sample critical values for the SADF and GSADF 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 $r_0 = 0.01 + 1.8/\sqrt{T}$. To isolate the effects of AR and MA components on innovations, when analyzing the size of the tests, we utilize the same data-generating process and 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

$$\begin{gathered} y_t = dT^{-\eta} + \theta y_{t-1} + \nu_t \\ \nu_t = \phi_1 \nu_{t-1} + \varepsilon_t + \vartheta_1 \varepsilon_{t-1} + \vartheta_2 \varepsilon_{t-2} + \vartheta_3 \varepsilon_{t-3} \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma_\nu). \end{gathered} \tag{4.1}$$

The parameter combinations for ν_t that we consider are presented in Table 4.2. The size of these coefficients are motivated by the empirical findings in the price-dividend and price-rent ratios. We begin by considering the case where ν_t follows an MA(1) process, which is mainly relevant for stock data and then consider the case where innovations follow an MA(3) or AR(1) process, which are relevant for housing market indices. All the simulated processes for ν_t are invertible and have standard normal innovations, i.e. $\sigma_\nu = 1$. We also consider the case where ν_t is a white noise process. This case is interesting since choosing the wrong autoregressive lag-length in (3.2) can also distort the size of the test, even without the presence of autocorrelation.

Table 4.2: ARMA coefficients considered for ν_t

Model	ϕ_1	ϑ_1	ϑ_2	ϑ_3
MA(1)	-	0.8	-	-
MA(1)	-	0.5	-	-
MA(1)	-	0.3	-	-
White noise	-	-	-	-
MA(3)	-	0.8	0.8	0.8
MA(3)	-	0.5	0.5	0.5
MA(3)	-	0.8	0.5	0.3
AR(1)	0.8	-	-	-
AR(1)	0.5	-	-	-

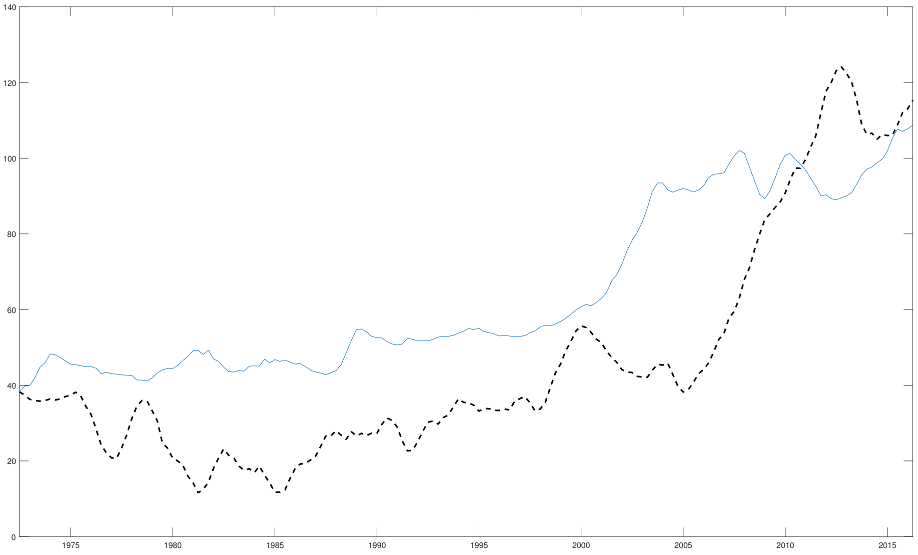
Once we have calculated the critical values, we then generate S replications using (4.1) and count the proportion of test statistics exceeding them. For the SADF we use $S = 4000$ and for the GSADF we use $S = 2000$. If the tests are unaffected by serially correlated innovations or an incorrectly specified 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%.

Figure 4.1 shows a simulated process using (4.1) together with the price-rent ratio of Australia from 1970Q1 to 2016Q2. The parameter values for the MA(3) process are set to match the coefficients of the estimated MA(3) model presented in Table 4.1. We set $d = \eta = \theta = 1$, $y_0 = 38.2$ and $\sigma_\nu = 1.39$ where the latter two are set to match the initial value of the price-rent ratio in Australia and the standard deviation of the first differences in the same series. The figure shows that the simulated unit root process with MA innovations is realistically capturing

[5] Although not shown in Table 4.1, the R^2 of the regression for Australia is 0.54, which also implies a fairly good fit.

the dynamics that drive the price-rent ratios since both series display a fairly similar and rather persistent behavior.[5] 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. 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.

Figure 4.1: Simulated unit root process with MA(3) innovations and the price-rent ratio in Australia



The figure shows a simulated unit root process with MA(3) innovations (dashed line) and the price-rent ratio in Australia from 1970Q1 to 2016Q2 (solid line). The data generating process for the simulated series is (4.1) with $d = \eta = \theta = 1$, $\phi_1 = 0$, $\vartheta_1 = 0.74$, $\vartheta_2 = 0.80$, $\vartheta_3 = 0.56$, $y_0 = 38.2$ and $\sigma_\nu = 1.39$.

4.2 Fixed lag-length

In this section we perform a size analysis for the SADF and GSADF using a fixed 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". We consider sample sizes $T = \{100, 200, 400, 1600\}$ and let k take integer values between 0 and 6.

Table 4.3 reports the empirical size of the SADF and GSADF tests for the MA(1) case. Our results for the white noise case are similar to what Phillips et al. (2015) report in their simulation study, namely that when the series in question does not suffer from serial correlation increasing the lag-length results in positive size distortions.[6] These distortions are much worse

[6] Our results deviate slightly from Phillips et al. (2015) since we use finite sample critical values, while Phillips

for the GSADF test than for the SADF, but they decrease for both tests as we let the sample size grow. For instance, given a nominal size of 5% and $T = 100$, setting $k = 6$ results in an incorrect rejection of the null in 74% of the cases using the GSADF and 20% using the SADF. If $T = 1600$ the empirical size of the GSADF test drops to 20% while the SADF is almost perfectly sized at 6%.

If the time series has a positive moving average component (in accordance with what is typically found empirically), fixing the lag-length k at zero results in size distortions that increase both with the magnitude of ϑ_1 and the sample size T. In fact, for large sample sizes setting $k = 0$ results in severe size distortions, even for MA components that appear relatively inconsequential, such the ones we empirically find in the price-dividend ratio. For example, when $T = 1600$ and $k = 0$, an MA component with $\vartheta_1 = 0.30$ results in an incorrect rejection of the null in 22% and 36% of the cases for SADF and GSADF, respectively. This example is particularly interesting since both the sample size and point estimate of the MA component of the price-dividend series used by Phillips et al. (2015) in their empirical application are very similar ($T = 1680, \hat{\vartheta}_1$=0.28). Given that Phillips et al. (2015) set the lag order $k = 0$, some of their empirical results should be interpreted with caution. Indeed, a visual inspection of the date-stamping of bubble periods (using the BSADF test statistic) presented in their article suggests that some of the bubbles are spurious since the price-dividend series is clearly trending downwards when the test statistic crosses the 95% critical value sequence. Phillips et al. (2015) conjecture that these false positives are due to volatility changes. While we do not dispute that volatility might play a role, our results suggest that serial correlation in the error terms can also be the culprit.

In general, when $\vartheta_1 > 0$, the GSADF comes closest to its nominal size with $k = 1$, while the SADF is more appropriately sized with either $k = 1$ or $k = 3$ depending on the magnitude of ϑ_1 and size of the sample.[7] The GSADF test is particularly sensitive to incorrect specifications of the approximating autoregression, since almost any choice different than $k = 1$ can yield misleading conclusions. For a positive ϑ_1, any heavily parametrized specification will result in over-rejections of the null, although these distortions decrease as we let the sample size grow. These results imply 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.[8]

Although we do not show the results, we find that using the tests on series with negatively autocorrelated error terms, i.e. $\vartheta_1 < 0$, yields diametrically opposed results. In such a case the test is undersized for low autoregressive truncations and the severity of the problem increases with sample size. At any rate, this scenario does not seem to have much empirical relevance

et al. (2015) use asymptotic critical values.

[7] It is interesting to note that odd numbered lag-lengths perform better than even numbered ones, this pattern is precisely the opposite of what Ng and Perron (1995) report for ADF tests for stationarity.

[8] 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. Our results for tests against explosivity are diametrically opposed since selecting a large k results in size distortions.

since we did not find any price or price-fundamentals series with a negative MA component.

Table 4.3: Empirical size for MA(1) innovations and fixed lag-length

SADF

	ϑ_1	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
T=100	**0.8**	0.28	0.03	0.18	0.10	0.20	0.18	0.27
r0=0.190	**0.5**	0.24	0.04	0.14	0.11	0.16	0.17	0.23
	0.3	0.16	0.06	0.10	0.12	0.14	0.17	0.22
	0	0.05	0.06	0.09	0.10	0.13	0.16	0.20
T=200	**0.8**	0.29	0.01	0.15	0.06	0.14	0.09	0.16
r0=0.130	**0.5**	0.25	0.03	0.12	0.08	0.12	0.12	0.15
	0.3	0.18	0.04	0.09	0.09	0.10	0.11	0.13
	0	0.04	0.06	0.07	0.08	0.10	0.11	0.13
T=400	**0.8**	0.33	0.01	0.15	0.05	0.13	0.07	0.13
r0=0.100	**0.5**	0.27	0.02	0.10	0.06	0.09	0.09	0.10
	0.3	0.19	0.04	0.08	0.07	0.08	0.09	0.10
	0	0.05	0.06	0.07	0.07	0.08	0.09	0.09
T=1600	**0.8**	0.38	0.01	0.14	0.03	0.10	0.04	0.09
r0=0.055	**0.5**	0.31	0.02	0.09	0.04	0.07	0.06	0.07
	0.3	0.22	0.03	0.06	0.05	0.06	0.06	0.06
	0	0.05	0.05	0.05	0.06	0.06	0.06	0.06

GSADF

	ϑ_1	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
T=100	**0.8**	0.36	0.06	0.37	0.28	0.56	0.60	0.86
r0=0.190	**0.5**	0.29	0.08	0.28	0.31	0.48	0.59	0.82
	0.3	0.20	0.10	0.23	0.30	0.46	0.57	0.79
	0	0.05	0.09	0.18	0.26	0.39	0.53	0.74
T=200	**0.8**	0.48	0.05	0.37	0.21	0.48	0.43	0.64
r0=0.130	**0.5**	0.37	0.06	0.25	0.23	0.39	0.43	0.57
	0.3	0.25	0.08	0.19	0.23	0.33	0.42	0.53
	0	0.05	0.08	0.15	0.20	0.30	0.38	0.51
T=400	**0.8**	0.54	0.02	0.33	0.12	0.38	0.29	0.49
r0=0.100	**0.5**	0.43	0.04	0.24	0.19	0.30	0.34	0.42
	0.3	0.30	0.07	0.18	0.21	0.28	0.32	0.40
	0	0.05	0.08	0.13	0.17	0.24	0.29	0.37
T=1600	**0.8**	0.70	0.01	0.29	0.07	0.27	0.13	0.30
r0=0.055	**0.5**	0.55	0.02	0.18	0.10	0.19	0.19	0.25
	0.3	0.36	0.04	0.11	0.11	0.16	0.18	0.22
	0	0.05	0.07	0.10	0.12	0.14	0.17	0.20

The tests are applied to series generated by (4.1) with $\phi_1 = \vartheta_2 = \vartheta_3 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

Tables 4.4 and 4.5 show the size of the SADF and GSADF tests when fixed lag-lengths are used and ν_t follows an MA(3) and AR(1) process, respectively. In both cases, we find that the size distortions for both tests increase significantly in comparison to the MA(1) case, especially if the lag-length is set to $k = 0$.

Table 4.4: Empirical size for MA(3) innovations and fixed lag-length

SADF

	ϑ_1	ϑ_2	ϑ_3	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
T=100	**0.8**	**0.8**	**0.8**	0.57	0.07	0.08	0.12	0.35	0.22	0.26
r0=0.190	**0.5**	**0.5**	**0.5**	0.52	0.13	0.09	0.09	0.26	0.26	0.27
	0.8	**0.5**	**0.3**	0.51	0.06	0.11	0.14	0.22	0.19	0.27
T=200	**0.8**	**0.8**	**0.8**	0.63	0.04	0.04	0.08	0.29	0.12	0.13
r0=0.130	**0.5**	**0.5**	**0.5**	0.58	0.09	0.05	0.05	0.21	0.18	0.15
	0.8	**0.5**	**0.3**	0.57	0.03	0.07	0.11	0.16	0.11	0.16
T=400	**0.8**	**0.8**	**0.8**	0.70	0.03	0.03	0.07	0.29	0.08	0.08
r0=0.100	**0.5**	**0.5**	**0.5**	0.65	0.08	0.03	0.04	0.20	0.15	0.11
	0.8	**0.5**	**0.3**	0.63	0.02	0.07	0.09	0.14	0.08	0.11
T=1600	**0.8**	**0.8**	**0.8**	0.79	0.01	0.01	0.04	0.28	0.04	0.04
r0=0.055	**0.5**	**0.5**	**0.5**	0.75	0.06	0.02	0.02	0.17	0.11	0.06
	0.8	**0.5**	**0.3**	0.73	0.01	0.04	0.07	0.11	0.05	0.07

GSADF

	ϑ_1	ϑ_2	ϑ_3	$k=0$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
T=100	**0.8**	**0.8**	**0.8**	0.84	0.23	0.28	0.38	0.81	0.81	0.92
r0=0.190	**0.5**	**0.5**	**0.5**	0.74	0.31	0.30	0.32	0.66	0.78	0.90
	0.8	**0.5**	**0.3**	0.75	0.15	0.30	0.43	0.64	0.72	0.90
T=200	**0.8**	**0.8**	**0.8**	0.93	0.19	0.21	0.32	0.76	0.60	0.66
r0=0.130	**0.5**	**0.5**	**0.5**	0.86	0.29	0.22	0.23	0.60	0.63	0.70
	0.8	**0.5**	**0.3**	0.85	0.11	0.23	0.33	0.53	0.52	0.67
T=400	**0.8**	**0.8**	**0.8**	0.96	0.10	0.11	0.22	0.73	0.42	0.44
r0=0.100	**0.5**	**0.5**	**0.5**	0.93	0.25	0.14	0.14	0.55	0.53	0.52
	0.8	**0.5**	**0.3**	0.93	0.07	0.18	0.28	0.46	0.37	0.49
T=1600	**0.8**	**0.8**	**0.8**	1.00	0.03	0.03	0.13	0.72	0.20	0.18
r0=0.055	**0.5**	**0.5**	**0.5**	0.99	0.16	0.04	0.04	0.48	0.38	0.26
	0.8	**0.5**	**0.3**	0.99	0.02	0.09	0.17	0.33	0.18	0.29

The tests are applied to a series generated by (4.1) with $\phi_1 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

In the MA(3) case there appears to be no clear pattern in terms of which lag-length achieves the best size, since this might vary between $k = 1$ and $k = 3$ depending on the size of the moving average coefficients and the sample size. For sample sizes usually found in housing market indices, i.e. $T = \{100, 200\}$, the SADF appears to have empirical sizes close to nominal size at lag-lengths between $k = 1$ and $k = 3$, but this depends heavily on the moving average coefficients. For the AR(1) case, the best size for both tests is achieved when $k = 1$. However, contrary to what we would expect, both tests remain highly oversized, even if the autoregressive lag is correctly specified. For example, with $T = 200$ and $\phi_1 = 0.8$, setting $k = 1$ results in a rejection of the null in 14% and 46% of the cases for SADF and GSADF, respectively.

Although MA(3) and AR(1) innovations result in generally oversized tests, the asymptotic validity of the lag augmentation becomes clear with large sample sizes since empirical size improves significantly with $k > 0$ when $T = 1600$. Paradoxically, when innovations are autocorrelated, choosing $k = 0$ results in severe size deterioration as T increases. This deterioration is considerably large with MA(3) and AR(1) components, and it can lead to 100% rejection rates when using the GSADF.

Table 4.5: Empirical size for AR(1) innovations and fixed lag-length

SADF

	ϕ_1	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
T=100	**0.8**	0.76	0.17	0.20	0.22	0.26	0.29	0.34
r0=0.190	**0.5**	0.42	0.10	0.12	0.14	0.17	0.20	0.26
T=200	**0.8**	0.83	0.14	0.16	0.17	0.19	0.21	0.23
r0=0.130	**0.5**	0.48	0.08	0.09	0.10	0.13	0.14	0.16
T=400	**0.8**	0.88	0.12	0.13	0.14	0.15	0.16	0.17
r0=0.100	**0.5**	0.56	0.08	0.09	0.09	0.11	0.11	0.12
T=1600	**0.8**	0.94	0.08	0.09	0.09	0.09	0.09	0.10
r0=0.055	**0.5**	0.65	0.06	0.07	0.07	0.08	0.08	0.08

GSADF

	ϕ_1	$k = 0$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
T=100	**0.8**	0.97	0.47	0.57	0.65	0.75	0.84	0.94
r0=0.190	**0.5**	0.60	0.20	0.32	0.40	0.54	0.69	0.86
T=200	**0.8**	1.00	0.46	0.54	0.62	0.70	0.75	0.84
r0=0.130	**0.5**	0.73	0.20	0.29	0.35	0.45	0.53	0.65
T=400	**0.8**	1.00	0.42	0.47	0.53	0.60	0.64	0.73
r0=0.100	**0.5**	0.82	0.15	0.22	0.27	0.34	0.40	0.47
T=1600	**0.8**	1.00	0.26	0.31	0.34	0.39	0.43	0.46
r0=0.055	**0.5**	0.96	0.12	0.15	0.17	0.21	0.23	0.27

The tests are applied to a series generated by (4.1) with $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

4.3 Variable lag-length

In the preceding section we showed that serially correlated innovations or an incorrect autoregressive truncation results in size distortions for both the SADF and GSADF tests. While it seems that $k = 1$ works well for the case when ν_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. From a practical point of view, the autoregressive and moving average orders of the data will not be known with certainty. Furthermore, given that both tests work by running regressions on different subsets of the sample, it seems intuitive 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 the choice of k, for each of the auxiliary regressions, to be data-dependent. The most common of these data-dependent rules are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and sequential testing for significance (STS).[9]

The idea behind information based rules is to select k by minimizing an objective function that trades off reductions in the sum of squared residuals against parsimony. More concretely, these rules choose k according to the following criterion,[10]

$$k = \underset{k_{min} \leq k \leq k_{max}}{arg\,min} \; IC(k), \qquad IC(k) = ln(\hat{\sigma}_k^2) + \frac{(k+1)C_T}{(T-k_{max})}, \tag{4.2}$$

where $\hat{\sigma}_k^2 = (T-k_{max})^{-1}\sum_{t=k_{max}+1}^{T} \hat{\varepsilon}_{tk}^2$ and C_T denotes a penalty function. This penalty function is specified as $C_T = 2$ for the AIC and $C_T = ln(T-k_{max})$ for the BIC. For brevity we only show the results of using the BIC since this method performs relatively better than the AIC or STS.[11] The results for the two latter, are available upon request. As in the preceding section we consider all parameter combinations presented in Table 4.2. We let the maximum number of lags k_{max}, take integer values between 1 and 6. For brevity we only consider $T = \{100, 200, 400\}$ which are values that have the most pertinence for empirical applications.

In the case where there is no autocorrelation in the innovations, Table 4.6 shows that the BIC is relatively good at controlling the size of the SADF, even when allowing k_{max} to be large. However, this is not the case for the GSADF, where allowing $k_{max} > 1$ results in size distortions that increase monotonically with k_{max}. In general, when there are no autoregressive and moving average components, these transient automatic lag-selection methods are much better at controlling the size of the SADF test than that of the GSADF test. The relative difference in performance is, in some cases, quite large. For example, when $T = 200$ and $\vartheta_1 = 0$, using the BIC with $k_{max} = 6$ results in an empirical size of 6% for the SADF and 24% for the GSADF. For the MA(1) case (Table 4.6), the BIC performs relatively well for the SADF as long as the maximum lag order is kept at $k_{max} = 1$ or $k_{max} = 3$, although the latter only works for large sample sizes. This relatively good performance does not apply to the GSADF test, which is universally oversized.

[9] The sequential test for significance consists of a general-to-specific approach that starts with a model with $k = k_{max}$ lags and sequentially reduces the lag order if the t-statistic of the last lag is insignificant at a given significance level.

[10] We adopt the formulation suggested by Ng and Perron (2005) where the effective number of observations is held fixed across models.

[11] We also considered the modified versions of the AIC and BIC proposed by Ng and Perron (2001) as well as using the non-parametric Phillips and Perron (1988) test instead of using (3.2) as the auxiliary regression, but this did not yield better results.

Table 4.6: Empirical size for MA(1) innovations and variable lag-length selection using the BIC

SADF

	ϑ_1	$k_{max}=1$	$k_{max}=2$	$k_{max}=3$	$k_{max}=4$	$k_{max}=5$	$k_{max}=6$
T=100	**0.8**	0.07	0.18	0.16	0.19	0.19	0.23
r0=0.190	**0.5**	0.10	0.14	0.14	0.16	0.17	0.19
	0.3	0.12	0.13	0.14	0.15	0.15	0.17
	0	0.06	0.07	0.07	0.08	0.09	0.11
T=200	**0.8**	0.04	0.15	0.11	0.14	0.14	0.15
r0=0.130	**0.5**	0.06	0.11	0.11	0.12	0.12	0.12
	0.3	0.09	0.11	0.11	0.11	0.12	0.12
	0	0.05	0.06	0.06	0.06	0.06	0.07
T=400	**0.8**	0.02	0.14	0.08	0.12	0.11	0.12
r0=0.100	**0.5**	0.04	0.10	0.10	0.10	0.10	0.10
	0.3	0.08	0.09	0.09	0.09	0.09	0.10
	0	0.06	0.06	0.06	0.06	0.06	0.06

GSADF

	ϑ_1	$k_{max}=1$	$k_{max}=2$	$k_{max}=3$	$k_{max}=4$	$k_{max}=5$	$k_{max}=6$
T=100	**0.8**	0.25	0.41	0.42	0.58	0.63	0.86
r0=0.190	**0.5**	0.24	0.34	0.37	0.48	0.58	0.78
	0.3	0.20	0.27	0.33	0.41	0.52	0.72
	0	0.10	0.17	0.22	0.31	0.42	0.65
T=200	**0.8**	0.24	0.40	0.38	0.49	0.50	0.60
r0=0.130	**0.5**	0.25	0.32	0.34	0.40	0.43	0.49
	0.3	0.21	0.26	0.28	0.32	0.34	0.39
	0	0.09	0.14	0.17	0.21	0.24	0.28
T=400	**0.8**	0.14	0.35	0.31	0.42	0.42	0.47
r0=0.100	**0.5**	0.22	0.30	0.31	0.34	0.36	0.38
	0.3	0.23	0.26	0.28	0.30	0.31	0.32
	0	0.08	0.10	0.12	0.14	0.15	0.17

The tests are applied to series generated by (4.1) with $\phi_1 = \vartheta_2 = \vartheta_3 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

The MA(3) case (Table 4.7) using the BIC also underscores big differences in performance between the SADF and GSADF, since the GSADF is severely oversized for any combination of MA(3) coefficients or sample size in the analysis. However, for the SADF the results are a bit more nuanced, since using the BIC can result in relatively low size distortions. Particularly, for $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.8$ and $\vartheta_1 = 0.8$, $\vartheta_2 = 0.5$, $\vartheta_3 = 0.3$, it seems that k_{max} between 1 and 3 can achieve empirical sizes close to nominal sizes. The AR(1) case (4.8) is generally oversized for both tests, and although size distortions decrease with T, spurious rejections of the null render both tests ineffective at most sample sizes used in empirical applications. Overall, it seems that allowing for flexibility by using an information criterion such as the BIC to choose lag-length is an ineffective way to control size.

Table 4.7: Empirical size for MA(3) innovations and variable lag-length selection using the BIC

SADF

	ϑ_1	ϑ_2	ϑ_3	$k_{max} = 1$	$k_{max} = 2$	$k_{max} = 3$	$k_{max} = 4$	$k_{max} = 5$	$k_{max} = 6$
T=100	**0.8**	**0.8**	**0.8**	0.09	0.11	0.14	0.28	0.26	0.29
r0=0.190	**0.5**	**0.5**	**0.5**	0.17	0.17	0.18	0.25	0.27	0.30
	0.8	**0.5**	**0.3**	0.07	0.11	0.13	0.16	0.17	0.21
T=200	**0.8**	**0.8**	**0.8**	0.06	0.06	0.08	0.26	0.17	0.18
r0=0.130	**0.5**	**0.5**	**0.5**	0.15	0.14	0.14	0.22	0.22	0.22
	0.8	**0.5**	**0.3**	0.04	0.07	0.08	0.10	0.10	0.11
T=400	**0.8**	**0.8**	**0.8**	0.03	0.03	0.05	0.25	0.11	0.11
r0=0.100	**0.5**	**0.5**	**0.5**	0.10	0.08	0.08	0.19	0.18	0.18
	0.8	**0.5**	**0.3**	0.02	0.05	0.05	0.07	0.07	0.07

GSADF

	ϑ_1	ϑ_2	ϑ_3	$k_{max} = 1$	$k_{max} = 2$	$k_{max} = 3$	$k_{max} = 4$	$k_{max} = 5$	$k_{max} = 6$
T=100	**0.8**	**0.8**	**0.8**	0.50	0.54	0.61	0.84	0.88	0.94
r0=0.190	**0.5**	**0.5**	**0.5**	0.59	0.61	0.63	0.75	0.84	0.92
	0.8	**0.5**	**0.3**	0.40	0.47	0.54	0.68	0.75	0.90
T=200	**0.8**	**0.8**	**0.8**	0.38	0.41	0.49	0.77	0.76	0.79
r0=0.130	**0.5**	**0.5**	**0.5**	0.60	0.62	0.64	0.74	0.77	0.80
	0.8	**0.5**	**0.3**	0.28	0.36	0.42	0.53	0.56	0.63
T=400	**0.8**	**0.8**	**0.8**	0.17	0.20	0.27	0.70	0.63	0.64
r0=0.100	**0.5**	**0.5**	**0.5**	0.51	0.53	0.54	0.67	0.69	0.70
	0.8	**0.5**	**0.3**	0.13	0.22	0.27	0.36	0.36	0.39

The tests are applied to series generated by (4.1) with $\phi_1 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

Table 4.8: Empirical size for AR(1) innovations and variable lag-length selection using the BIC

SADF

	ϕ_1	$k_{max} = 1$	$k_{max} = 2$	$k_{max} = 3$	$k_{max} = 4$	$k_{max} = 5$	$k_{max} = 6$
T=100	**0.8**	0.20	0.23	0.24	0.26	0.28	0.32
r0=0.190	**0.5**	0.16	0.18	0.19	0.20	0.21	0.24
T=200	**0.8**	0.15	0.17	0.17	0.18	0.19	0.20
r0=0.130	**0.5**	0.12	0.13	0.14	0.15	0.15	0.16
T=400	**0.8**	0.12	0.12	0.12	0.13	0.13	0.13
r0=0.100	**0.5**	0.09	0.09	0.10	0.10	0.11	0.11

GSADF

	ϕ_1	$k_{max} = 1$	$k_{max} = 2$	$k_{max} = 3$	$k_{max} = 4$	$k_{max} = 5$	$k_{max} = 6$
T=100	**0.8**	0.72	0.78	0.83	0.88	0.93	0.97
r0=0.190	**0.5**	0.49	0.53	0.57	0.64	0.71	0.85
T=200	**0.8**	0.61	0.67	0.73	0.78	0.82	0.86
r0=0.130	**0.5**	0.53	0.55	0.57	0.61	0.63	0.67
T=400	**0.8**	0.43	0.48	0.52	0.55	0.57	0.61
r0=0.100	**0.5**	0.42	0.48	0.50	0.53	0.55	0.56

The tests are applied to series generated by (4.1) with $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0$. We use 4,000 replications for the SADF and 2,000 for the GSADF. Nominal size is 5%.

5 The sieve-bootstrap SADF and GSADF tests

In the preceding section we showed that using either a fixed lag-length or some automatic lag selection method will generally not be an effective strategy when controlling the size of SADF and GSADF in the presence of serially correlated innovations. This problem is particularly serious for the GSADF test, which is severely oversized when innovations follow MA(3) or AR(1) processes. Motivated by Park (2003), Chang and Park (2003) and Palm et al. (2008) who propose sieve bootstrap versions of the ADF test, we consider the same sieve bootstrap algorithm but apply it to the SADF and GSADF tests.[12] Chang and Park (2003) and Palm et al. (2008) show that the asymptotic distribution of the ADF bootstrap test is the same under the null as the asymptotic distribution of the original ADF test. They also use simulations to 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 tests. In this section, we begin by describing the sieve bootstrap algorithm, then investigate the empirical size and power of the bootstrap versions of the tests in the case where innovations are serially correlated.

5.1 The algorithm

Our algorithm is conformed by the following steps:

Step 1 . Estimate by OLS the ADF regression, to obtain estimates $\hat{\psi}_{i,T}$ and residuals:

$$\hat{\varepsilon}_{t,T} = y_t - \hat{\alpha} - \hat{\beta} y_{t-1} - \sum_{i=1}^{k^*} \hat{\psi}_{i,T} \Delta y_{t-i} \qquad t = k^* + 1, ..., T \tag{5.1}$$

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

Step 2. Generate an iid sample of bootstrap errors, $\varepsilon^*_{t,T}$, by drawing randomly with replacement from:

$$\hat{\varepsilon}_{t,T} - (T - k^*)^{-1} \sum_{t=1+k^*}^{T} \hat{\varepsilon}_{t,T} \tag{5.2}$$

Step 3. Construct u^*_t recursively from ε^*_t as,

$$u^*_t = \sum_{i=1}^{k^*} \hat{\psi}_{i,T} u^*_{t-i,T} + \varepsilon^*_t \tag{5.3}$$

To have a full bootstrap sample of size T and eliminate any initialization effect, we draw

[12] The algorithm proposed by Park (2003) and Chang and Park (2003), differs slightly from the one proposed by Palm et al. (2008) since the former imposes a unit root restriction in the regression of step 1 whereas the latter estimates the autoregressive coefficient $\hat{\beta}$ in (5.1). We found that the sieve bootstrap proposed by Palm et al. (2008) has slightly better size properties in the case of the SADF and GSADF.

$(T-k^*)+b$ bootstrap errors from step 2, and then discard the first $b-k^*$ values of u_t^*. Having u_t^*, we can build y_t^* as $y_t^* = y_{t-1}^* + u_t^*$, $t = 1, .., T$ with $y_0^* = 0$.

Step 4. Using y_t^*, we calculate the bootstrap test statistic we are interested in (i.e. the SADF or GSADF)

$$SADF^*(r_0) = \underset{r_2 \in [r_0,1]}{sup} \; ADF_0^{*r_2} \tag{5.4}$$

or

$$GSADF^*(r_0) = \underset{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}}{sup} \left\{ ADF_{r_1}^{*r_2} \right\}. \tag{5.5}$$

The lag-length in the auxiliary ADF regressions that conform these bootstrap test statistics should be fixed at $k = k^*$.

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

$$cv(q)^{SADF*} := max \left\{ x : M^{*-1} \sum_{m=1}^{M^*} I\left(SADF_m^*(r_0) < x\right) \leq q \right\} \tag{5.6}$$

or

$$cv(q)^{GSADF^*} := max \left\{ x : M^{*-1} \sum_{m=1}^{M^*} I\left(GSADF_m^*(r_0) < x\right) \leq q \right\}. \tag{5.7}$$

Step 6. Calculate the actual test statistic, $SADF^B(r_0)$ or $GSADF^B(r_0)$ with y_t, using a lag order equal to k^*. Reject the null of "no-bubbles" if this test statistic is larger than the bootstrap critical value calculated above.[13]

5.2 Finite sample simulations

To analyze the empirical size properties of the sieve bootstrap $SADF^B$ and $GSADF^B$ tests, we use (4.1) as the data-generating process with parameter combinations presented in Table 4.2. Sample sizes analyzed are $T = \{100, 200, 400\}$. We use 4,000 simulations when analyzing the size and power of the $SADF^B$ and 2,000 for the $GSADF^B$. For each simulated series, we use $M^* = 899$ bootstrap replications to calculate the critical values, $cv(q)^{SADF^*}$ and $cv(q)^{GSADF^*}$. We set the maximum lag-length of the bootstrap tests to $k_{max} = int[8(T/100)^{1/4}]$, and let the optimal lag-length, k^*, be selected by the BIC.

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

5.3 Size of the $SADF^B$ and $GSADF^B$

Table 5.1 shows the empirical size, using a nominal size of 5%, of the $SADF^B$ and $GSADF^B$ for the MA(1), MA(3) and AR(1) cases. The bootstrap procedure is very effective at controlling the size of the tests when there are autoregressive and moving average components in the series and these size corrections hold even for relatively small sample sizes, i.e. $T = 100$.

Table 5.1: Empirical size for SADF and GSADF bootstrap tests

	MA(1)			MA(3)					AR(1)		
T	ϑ_1	$SADF^B$	$GSADF^B$	ϑ_1	ϑ_2	ϑ_3	$SADF^B$	$GSADF^B$	ϕ_1	$SADF^B$	$GSADF^B$
100	**0.8**	0.05	0.05	**0.8**	**0.8**	**0.8**	0.05	0.05	**0.8**	0.05	0.06
	0.5	0.04	0.04	**0.5**	**0.5**	**0.5**	0.07	0.07	**0.5**	0.06	0.05
	0.3	0.06	0.05	**0.8**	**0.5**	**0.3**	0.03	0.03	-	-	-
	0	0.05	0.05	-	-	-	-	-	-	-	-
200	**0.8**	0.05	0.06	**0.8**	**0.8**	**0.8**	0.04	0.04	**0.8**	0.05	0.06
	0.5	0.05	0.05	**0.5**	**0.5**	**0.5**	0.08	0.08	**0.5**	0.05	0.05
	0.3	0.04	0.04	**0.8**	**0.5**	**0.3**	0.03	0.03	-	-	-
	0	0.05	0.05	-	-	-	-	-	-	-	-
400	**0.8**	0.05	0.05	**0.8**	**0.8**	**0.8**	0.05	0.05	**0.8**	0.05	0.05
	0.5	0.06	0.06	**0.5**	**0.5**	**0.5**	0.08	0.08	**0.5**	0.05	0.06
	0.3	0.04	0.04	**0.8**	**0.5**	**0.3**	0.03	0.03	-	-	-
	0	0.05	0.05	-	-	-	-	-	-	-	-

The tests are applied to 4,000 series for the $SADF^B$ and 2,000 series for the $GSADF^B$ generated by (4.1). Initial windows are set to $r_0 = \{0.190, 0.130, 0.100\}$ for $T = \{100, 200, 400\}$, respectively. The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

For the MA(1) case, all results lie within one percentage point from nominal size, and there is no deterioration of size as the moving average coefficient increases. The results for the MA(3) case are also robust, and there is never more than 3 percentage points in difference between empirical and nominal size, although there is variation depending on the combination of coefficients. Nonetheless, when considering that MA(3) innovations resulted in such a degree of size distortions that the SADF and GSADF test were rendered almost useless, the performance of the $SADF^B$ and $GSADF^B$ tests seems comparatively impressive. Since the AR(1) case does not involve any approximation error from a moving average representation to an autoregressive one, it is almost perfectly sized with only 1% deviations from nominal size as a worst case scenario. These results are within the margin of error that we can expect from random sampling given the number of simulations that we use.

5.4 Power of the sieve-bootstrap $SADF^B$ and $GSADF^B$

The preceding section showed that the sieve bootstrap algorithm we propose is very successful in restoring the size of both tests in the presence of autocorrelated innovations. In this section, we evaluate the empirical detection rate of the tests. In order to do this, we use the mildly explosive single bubble process proposed by Phillips and Yu (2009) and Phillips et al. (2015), but allowing the series to have autocorrelated innovations:

$$
Bubble\, process: \qquad y_t = \begin{cases} y_{t-1} + \nu_t, & t = 1, ..., \tau_e - 1 \\ \delta_{1,T} y_{t-1} + \nu_t, & t = \tau_e, ..., \tau_f \\ y_t^c, & t = \tau_f + 1 \\ y_{t-1} + \nu_t, & t = \tau_f + 2, ..., T \end{cases} \tag{5.8}
$$

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

Following Phillips et al. (2015) we let $\delta_{1,T} = 1 + cT^{-\alpha}$, with $c > 0$, $0 < \alpha < 1$ and $y_t^c = y_{\tau e} + O_p(1)$. Under this bubble process, the series starts as a random walk and continues to be so until $\tau_e = \lfloor Tr_e \rfloor$, where the series becomes explosive with a local to unity autoregressive coefficient, $\delta_{1,T}$, and this explosivity continues until observation $\tau_f = \lfloor Tr_f \rfloor$. At observation $\tau_f + 1$ the bubble collapses to a value of y_t^c, which represents the fundamental value of the series plus a random innovation. After the collapse, the series continues its martingale behavior until the end of the series. For ease of comparison, we follow Phillips et al. (2015) and set $y_0 = 100$, $\sigma = 6.79$, $c = 1$, $\alpha = 0.6$. We use $T = \{100, 200, 400\}$ which result in autoregressive coefficients during the bubble period of $\delta_{1,T,} = \{1.06, 1.04, 1.03\}$.[14] We let the bubble start at $\tau_e/T = 0.40$ and end at $\tau_f/T = 0.55$.

Tables 5.2, 5.3 and 5.4 report the power of the $SADF^B$ and $GSADF^B$ together with the power of the SADF and GSADF test. In addition, we also report the (infeasible) size-adjusted power of the SADF and GSADF in the case where ν_t follows an AR or MA process.[15] We note that the presence of autocorrelated innovations, particularly in the AR(1) case, can also affect the magnitude of the bubble given by (5.8), and thus have a positive effect on power when compared to the white noise case. However, this effect is also present in the size-adjusted power figures, making these figures the most relevant point of comparison.

If we begin by considering the case where there are no autoregressive or moving average components (Table 5.2), there appears to be a small decrease in empirical power for the bootstrap tests in small samples (i.e. $T = 100$). This power loss disappears as the sample size increases, since the power of the bootstrap tests is virtually the same as that of the SADF and GSADF tests when $T \geq 200$. For the MA(1) case (Table 5.2), the power of the bootstrap tests decreases with the magnitude of the moving average coefficient. This is as expected since the augmented test (with $k^* > 0$) will have a finite sample distribution that is shifted to the right. Nonetheless, the power of the $SADF^B$ and $GSADF^B$ is greater than the size-adjusted power of the original tests. This higher detection rate is particularly noteworthy when it comes to the $GSADF^B$ in small samples, which outperforms the GSADF test by an average of 9 percentage points when

[14] Note that this bubble process keeps the empirical power of the tests relatively constant as T increases allowing us to focus on the differences in power (between the bootstrap and original tests) that arise purely because of a larger T.

[15] 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 (4.1) instead of (3.1) as null hypothesis.

$\vartheta_1 > 0$ and the sample size is $T = 100$.

Table 5.2: Empirical power for $SADF^B$ and $GSADF^B$ tests, MA(1) case

	ϑ_1	$SADF$	$GSADF$	$SADF^B$	$GSADF^B$
T=100	**0.8**	0.47	0.46	0.55	0.59
r0=0.190	**0.5**	0.55	0.54	0.60	0.63
	0.3	0.58	0.61	0.65	0.67
	0	0.72	0.75	0.68	0.70
T=200	**0.8**	0.59	0.58	0.63	0.66
r0=0.130	**0.5**	0.64	0.61	0.66	0.70
	0.3	0.68	0.69	0.69	0.73
	0	0.73	0.74	0.72	0.73
T=400	**0.8**	0.71	0.71	0.71	0.71
r0=0.100	**0.5**	0.72	0.73	0.74	0.76
	0.3	0.74	0.75	0.76	0.77
	0	0.78	0.80	0.78	0.80

The tests are applied to 4,000 series for the $SADF^B$ and 2,000 series for the $GSADF^B$ generated by (5.8). The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

As we would expect, given the size distortions we show above, the size-adjusted power of the SADF and GSADF test decreases considerably for the MA(3) case (Table 5.3). However, although the power of the $SADF^B$ and $GSADF^B$ does decrease in comparison to the white noise case, it does not decrease as much as the (size-adjusted) power of the SADF and GSADF. As a result, the bootstrap tests achieve a much higher empirical power than the original tests. For example, given a sample size of $T = 100$ and MA coefficients of $\vartheta_1 = \vartheta_2 = \vartheta_3 = 0.5$, the power of the $SADF^B$ and $GSADF^B$ is 19 and 26 percentage points higher than the size-adjusted power of the SADF and GSADF, respectively. Finally, it is worth noting that as the sample size increases, the power of the $SADF^B$ and $GSADF^B$ in the presence of MA components approaches the power of the SADF and GSADF in the case of no MA components.

For the AR(1) case (Table 5.4), the power advantages of the bootstrap tests over the size-adjusted power of the original tests are even more striking. This is specially the case when $T = 100$, where the $SADF^B$ and $GSADF^B$ have an average power advantage (across both autoregressive cases) of roughly 25 and 32 percentage points over the original tests, respectively.

Table 5.3: Empirical power for $SADF^B$ and $GSADF^B$ tests, MA(3) case

	ϑ_1	ϑ_2	ϑ_3	$SADF$	$GSADF$	$SADF^B$	$GSADF^B$
T=100	**0.8**	**0.8**	**0.8**	0.35	0.34	0.54	0.59
r0=0.190	**0.5**	**0.5**	**0.5**	0.38	0.38	0.57	0.64
	0.8	**0.5**	**0.3**	0.37	0.38	0.54	0.61
T=200	**0.8**	**0.8**	**0.8**	0.53	0.48	0.62	0.69
r0=0.130	**0.5**	**0.5**	**0.5**	0.59	0.53	0.67	0.72
	0.8	**0.5**	**0.3**	0.53	0.51	0.62	0.67
T=400	**0.8**	**0.8**	**0.8**	0.67	0.67	0.73	0.76
r0=0.100	**0.5**	**0.5**	**0.5**	0.69	0.68	0.75	0.77
	0.8	**0.5**	**0.3**	0.68	0.66	0.71	0.74

The tests are applied to 4,000 series for the $SADF^B$ and 2,000 series for the $GSADF^B$ generated by (5.8). The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

Table 5.4: Empirical power for $SADF^B$ and $GSADF^B$ tests, AR(1) case

	ϕ_1	$SADF$	$GSADF$	$SADF^B$	$GSADF^B$
T=100	**0.8**	0.34	0.35	0.67	0.75
r0=0.190	**0.5**	0.45	0.43	0.61	0.67
T=200	**0.8**	0.50	0.51	0.73	0.79
r0=0.130	**0.5**	0.57	0.57	0.68	0.72
T=400	**0.8**	0.68	0.67	0.81	0.82
r0=0.100	**0.5**	0.72	0.70	0.75	0.76

The tests are applied to 4,000 series for the $SADF^B$ and 2,000 series for the $GSADF^B$ generated by (5.8). The bootstrap critical values are calculated using 899 replications. Nominal size is 5%.

Overall, although there is a slight decrease in power for small samples when there are no MA or AR components, the $SADF^B$ and $GSADF^B$ have higher detection rates than the size-adjusted power of the SADF and GSADF in the presence of serially correlated innovations. More importantly, this power disadvantage is only present with small samples and it is negligible in sample sizes that are relevant for most empirical applications.

6 Empirical application

In this section we present an empirical application of our bootstrap tests using international housing indices. In section 6.1 we discuss the relevance of the issue and briefly summarize how past researchers have dealt with the issue of serial correlation in innovations when testing for bubbles in the housing market. In section 6.2 we describe the data used in the analysis. Section 6.3 shows an empirical application of the $GSADF^B$ test and compares the results of this test with the results obtained when using the GSADF 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 $SADF^B$. Finally, in section 6.4 we compare the date-stamping results of the BSADF and $BSADF^B$ for a small subset of countries in which the null was rejected.

6.1 Past evidence on housing bubbles

International housing markets have received a lot of attention following the boom and bust that contributed to the 2008-09 global financial crisis 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 we showed in our 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 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 with a fixed lag of $k = 4$ on price-rent ratio they reject the null of “no-bubbles” in all but 3 of the 22 countries. In a recent contribution, 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 to the price-rent ratio of regional housing markets in Israel to test for bubbles. Caspi (2015) uses fixed lag-lengths between 1 and 6 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 also he notes that the results of the Gush Dan region are highly sensitive to the lag specification that is used.

6.2 The data

We use the official OECD data for 17 countries: Australia, Canada, Denmark, Germany, Spain, Finland, France, the UK, Italy, Ireland, Japan, the Netherlands, New Zealand, Norway, Sweden, Switzerland and the US. The data set contains seasonally adjusted quarterly observations of house price-rent ratios that span from 1970Q1 to 2016Q2, except for: Australia (begins in 1972Q2), Spain (begins in 1971Q1), UK (begins in 1968Q2), Norway (1979Q1 to 2016Q3) and Sweden (begins in 1981Q1). Note that this is the same data as the one presented in Table 4.1. Hence, the presence of AR and MA components is already established. Since the average sample size among all countries included is $T = 180$, the results of our finite sample simulation study with $T = 200$ are the most relevant as points of comparison.

6.3 Testing for bubbles using the GSADF and $GSADF^B$

Table 6.1 presents the results of the GSADF tests for a bubble on 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 6. The p-values, which are calculated using 5,000 replications of a random walk with iid Gaussian innovations, are also shown. The last three columns show the results of the $GSADF^B$, the lag-length used in the bootstrap test, k^*, and the bootstrap p-value, which is calculated using $M^* = 5,000$ bootstrap replications. Note that the GSADF test statistic with a fixed lag-length and the $GSADF^B$ will always be equal when $k = k^*$. This is, however, not necessarily the case for the variable lag-length version of the test since this version of the test allows each of the ADF regressions to select a different k.

When using a fixed lag-length of $k = 1$, we find evidence of a bubble at a 1% significance level in all countries but the UK and Italy where the test statistics are only significant at the 5% and 10% level, respectively. The results when using a fixed lag-length of 4 are similar. In this case we find evidence of bubbles at the 1% level in all countries except France and Italy, which are significant at the 5% and 10% level, respectively. When using the BIC to automatically select the transient lag-length with a $k_{max} = 6$, we reject the null of “no bubbles” for all the countries at the 1% level. However, when we apply the $GSADF^B$ to the same data set, we fail to reject the null even at the 10% level for Germany, France, the UK, Italy, Japan, Norway and Switzerland. For Denmark, Spain and Finland the null is rejected at the 5% level and for Sweden at the 10%. It is interesting to note that, in contrast to the original GSADF test, the bootstrap version of the test does not reject the null (at conventional significance levels) in countries such as Germany and Italy where a visual analysis of the data also does not seem to suggest the presence of explosive bubbles in the sample period.

Table 6.1: 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	176	5.56	0.000	6.66	0.000	12.53	0.000	5.56	1	0.000
Canada	186	6.01	0.000	4.63	0.000	10.25	0.000	6.01	1	0.000
Germany	186	2.94	0.004	3.17	0.002	4.73	0.000	2.71	2	0.250
Denmark	186	3.40	0.001	3.54	0.001	11.28	0.000	3.40	1	0.026
Spain	182	3.50	0.001	3.21	0.003	6.31	0.000	4.20	2	0.041
Finland	186	3.75	0.000	4.20	0.000	7.09	0.000	3.75	1	0.012
France	186	4.66	0.000	2.48	0.016	8.65	0.000	3.20	3	0.148
U.K	193	2.67	0.010	3.48	0.000	5.24	0.000	2.67	1	0.141
Italy	186	2.01	0.058	1.89	0.081	5.91	0.000	2.08	2	0.340
Ireland	186	4.65	0.000	4.68	0.000	8.41	0.000	4.65	1	0.001
Japan	186	3.18	0.002	4.00	0.000	7.48	0.000	3.18	1	0.150
Netherlands	186	6.65	0.000	5.35	0.000	10.11	0.000	5.35	4	0.010
New Zealand	186	4.31	0.000	4.31	0.000	7.82	0.000	4.31	1	0.004
Norway	151	3.09	0.003	2.80	0.007	4.71	0.000	2.74	2	0.199
Sweden	146	3.17	0.003	5.00	0.000	6.25	0.000	3.17	1	0.090
Switzerland	186	4.48	0.000	2.77	0.006	4.59	0.000	2.48	3	0.540
USA	186	5.95	0.000	3.29	0.002	13.62	0.000	4.93	2	0.004

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^*.

6.4 Date-stamping bubbles

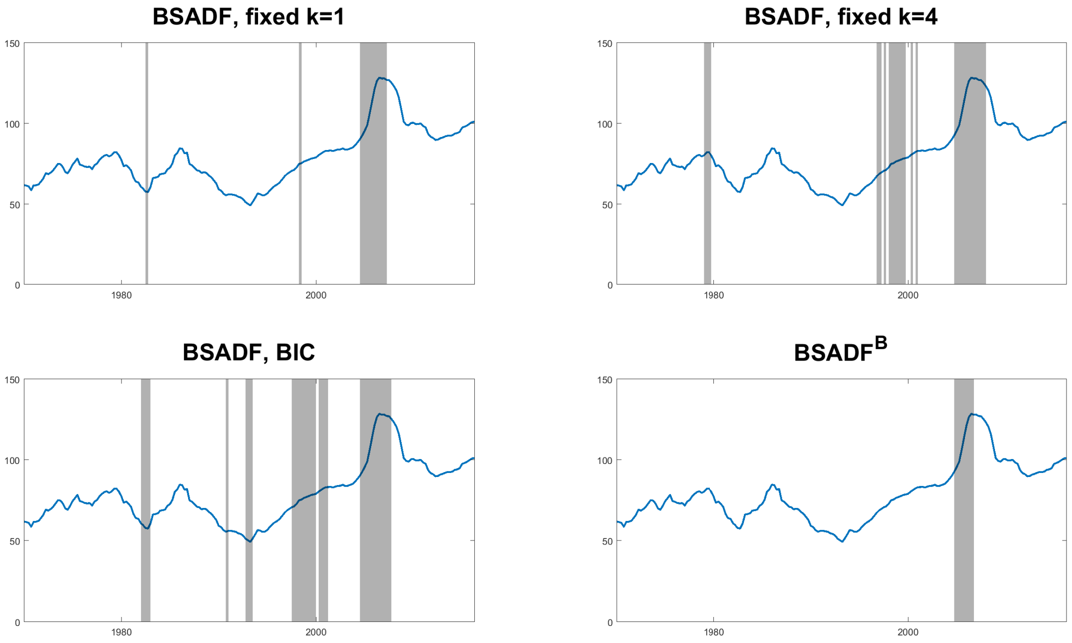
Figures 6.1, 6.2, and 6.3 show the date-stamping of the speculative bubbles for Denmark, the Netherlands and the US using the BSADF and $BSADF^B$ tests. For the BSADF test we use a fixed lag-length of $k = 1$ and $k = 4$ in accordance with Pavlidis et al. (2015), and a variable lag-length selected by the BIC with $k_{max} = 6$, which matches the methodology of Shi et al. (2016). The objective of this exercise does not lie in the accurate identification of bubble periods but in a comparison of the bootstrap test to the original BSADF test under different lag specifications. This can be particularly illustrating since we can see if the periods identified as explosive by the BSADF tests actually appear to be so by visual inspection. In line with the previous section, we calculate the critical value sequence using the 95% quantile of 5,000 replications.[16]

When looking at the figures below it seems clear that the BSADF test performs less than optimally and it identifies periods in which the series does not seem to have an explosive autoregressive root as bubble periods. In the case of Denmark (Figure 6.1), it seems that $k = 1$ performs relatively well, but it still identifies the second quarter of 1982 as being explosive even though the series is at a trough. The worst performance comes from using the BIC as an automatic method to select k, since this identifies periods that are clearly not explosive as bubbles and it appears to have a tendency towards finding spurious periods of explosiveness. We also note that, in contrast to the BSADF, which seems to suffer a slight delay, the $BSADF^B$ is very

[16] The lag-length in the $BSADF^B$ test, k^*, is the same as the one presented in Table 6.1.

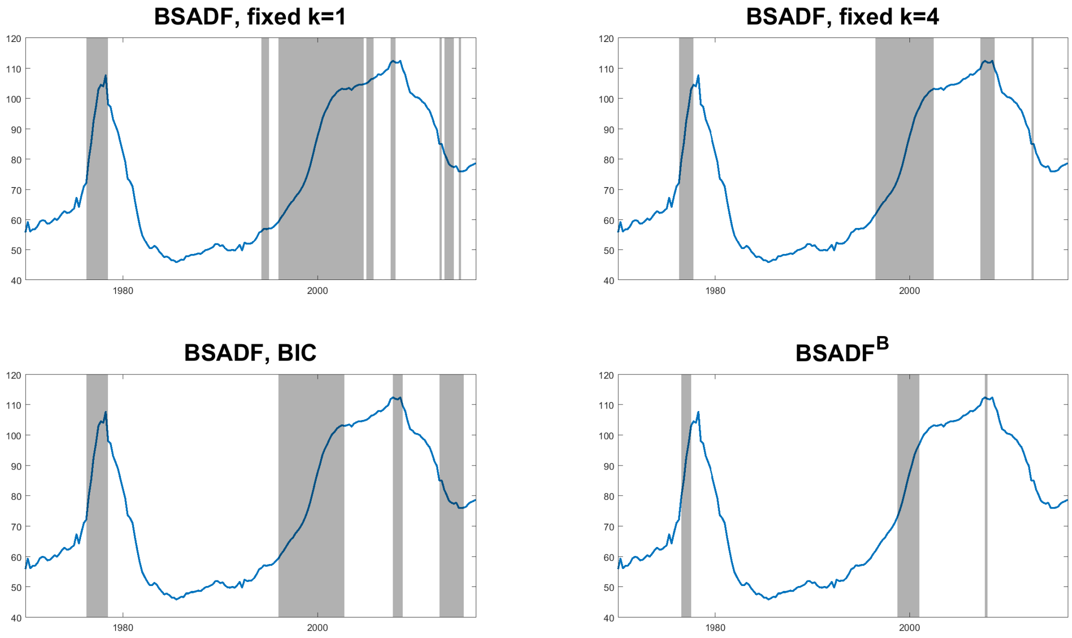
effective at date stamping the collapse of the bubble. The difference in the date-stamping results for the Netherlands (Figure 6.2) are even more striking since the original test identifies the end of the sample period as being explosive even though the series is downward trending, irrespective of which method is used to select k. All methods identify a period of apparent stability in the late 2007 as being explosive. This is possibly the result of a small delay in the detection of the break point since this period is preceded by what could be explosiveness. Finally, the results for the US (Figure 6.3) also show that, in contrast to the BSADF, the bootstrap version of the test is quite effective at date-stamping the infamous housing bubble of the early to mid 2000s without also incorrectly defining other periods of relative stability as bubbles. Overall, the $BSADF^B$ appears to be able to capture periods of explosive autoregressive growth without incurring false positives due to serially correlated innovations.

Figure 6.1: Date-stamping of housing bubbles in Denmark



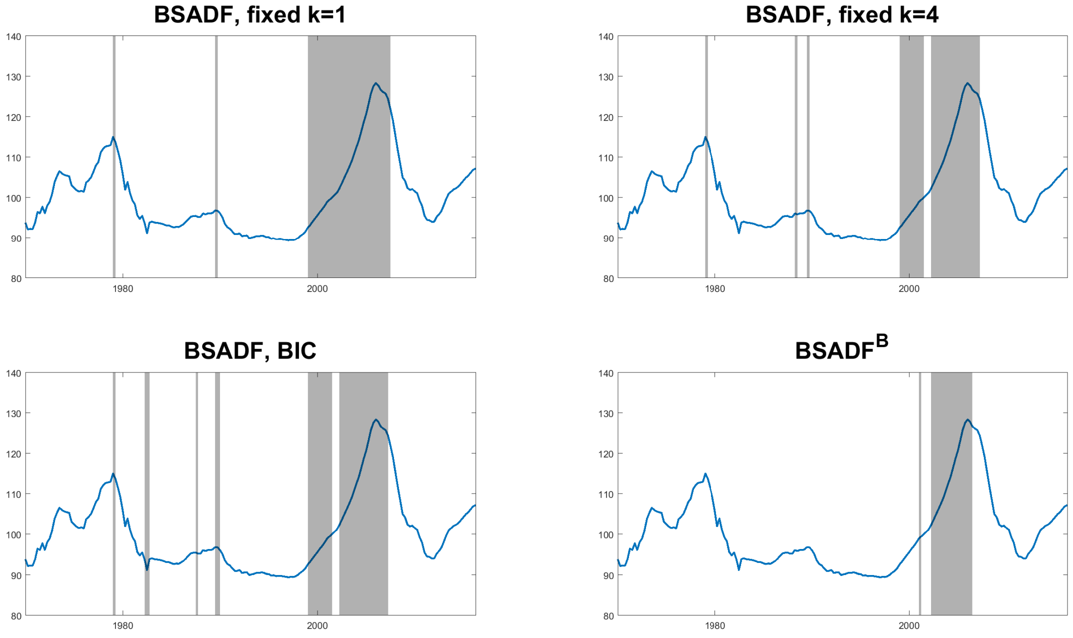
Price-rent ratio in Denmark (1970Q1-2016Q2). The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (top-left), $k = 4$ (top-right), letting the BIC automatically choose the transient lag with $k_{max} = 6$ (bottom-left) and by the $BSADF^B$ test (bottom-right).

Figure 6.2: Date-stamping of housing bubbles in the Netherlands



Price-rent ratio in the Netherlands (1970Q1-2016Q2). The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (top-left), $k = 4$ (top-right), letting the BIC automatically choose the transient lag with $k_{max} = 6$ (bottom-left) and by the $BSADF^B$ test (bottom-right).

Figure 6.3: Date-stamping of housing bubbles in the US



Price-rent ratio in the US (1970Q1-2016Q2). The shaded areas indicate the periods identified as a bubble by the BSADF test using a fixed lag-length of $k = 1$ (top-left), $k = 4$ (top-right), letting the BIC automatically choose the transient lag with $k_{max} = 6$ (bottom-left) and by the $BSADF^B$ test (bottom-right).

Overall, our results show that using critical values that are robust to the presence of autocorrelated innovations leads to much weaker evidence of speculative bubbles in international housing markets. Since the empirical power of the $GSADF^B$ is relatively unaffected by AR or MA components when $T \approx 200$, our results suggest that many of the cyclical upswings in the price-rent ratio have been spuriously interpreted as explosive bubbles and that once serial correlation in the innovations is taken into account the evidence for explosivity in many of the countries is not present anymore. Nonetheless, we still find evidence of speculative bubbles in 9 out 17 countries, and this evidence is strongly significant for six of the countries.

7 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) and Phillips et al. (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 more or less perfectly sized test statistics. More importantly, these size corrections come at relatively low cost for the power of the tests.

Applied to the price-rent ratio in 17 OECD countries, we find less strong evidence of bubbles in the housing market using the bootstrap version of the test compared to both a fixed and automatic lag-length selection. While all 17 price-rent ratios are concluded to be explosive on a 1% significance level using the BIC to select the transient lag-length, only 9 price-rent ratios display explosive behavior on a 5% level according to the bootstrap version of the test.

References

Agiakloglou, C. and P. Newbold (1992): "Empirical evidence on Dickey-Fuller-type tests." *Journal of Time Series Analysis*, 15, 253–262.

Caspi, I. (2015): "Testing for a housing bubble at the national and regional level: the case of Israel." *Empirical Economics*, Forthcoming.

Chang, Y. and J. Y. Park (2002): "On the asymptotics of ADF tests for unit roots." *Econometric Reviews*, 21, 431–447.

Chang, Y. and J. Y. Park (2003): "A sieve bootstrap for the test of a unit root." *Journal of Time Series Analysis*, 24, 370–400.

Craine, R. (1993): "Rational bubbles: A test." *Journal of Economic Dynamics and Control*, 17, 829–846.

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

Engsted, T., S. J. Hviid, and T. Q. Pedersen (2016): "Explosive bubbles in house prices? Evidence from the OECD countries." *Journal of International Financial Markets, Institutions and Money*, 40, 14–25.

Engsted, T. and B. Nielsen (2012): "Testing for rational bubbles in a coexplosive vector autoregression." *Econometrics Journal*, 15, 226–254.

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

Figuerola-Ferretti, I. and J. R. McCrorie (2016): "The shine of precious metals around the global financial crisis." *Journal of Empirical Finance*, 38, 717–738.

Ghysels, E., A. Plazzi, R. Valkanov, and W. Torous (2013): "Forecasting real estate prices." *Handbook of Economic Forecasting*, 2, 509–580.

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

Harvey, I. H., S. J. Leybourne, R. Sollis, and R. Taylor (2015b): "Tests for explosive financial bubbles in the presence of non-stationary volatility." *Journal of Empirical Finance*, Forthcoming.

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

KIVEDAL, B. K. (2013): "Testing for rational bubbles in the US housing market." *Journal of Macroeconomics*, 38, 369–381.

KRAUSSL, R. L., R. TUSSL, T. LEHNERT, AND N. MARTELIN (2016): "Is there a bubble in the art market?" *Journal of Empirical Finance*, 35, 99–109.

NG, S. AND P. PERRON (1995): "Unit root tests in ARMA models with data dependent methods for the selection of the truncation lag." *Journal of the American Statistical Association*, 90, 253–268.

NG, S. AND P. PERRON (2001): "Lag length selection and the construction of unit root tests with good size and power." *Econometrica*, 69, 1619–1554.

NG, S. AND P. PERRON (2005): "A Note on the Selection of Time Series Models," *Oxford Bulletin of Economics and Statistics*, Volume 67, 115–134.

PALM, F. C., S. SMEEKES, AND J. P. URBAIN (2008): "Bootstrap unit-root tests: comparison and extenstions." *Journal of Time Series Analysis*, 29, 371–401.

PARK, J. Y. (2003): "Bootstrap unit root tests," *Econometrica*, 71, 1845–1895.

PAVLIDIS, E., A. YUSUPOVA, D. PAYA, D. PEEL, E. MARTINEZ-GARCIA, A. MACK, AND V. GROSSMAN (2015): "Episodes of exuberance in housing markets: In search of the smoking gun." *Journal of Real Estate Finance and Economics*, Forthcoming.

PERRON, P. (1988): "Trends and random walks in macroeconomic time series: further evidence from a new approach." *Journal of Economic Dynamics and Control*, 12, 297–332.

PHILLIPS, P. C. B. AND P. PERRON (1988): "Testing for a unit root in time series regression." *Biometrika*, 75, 335–346.

PHILLIPS, P. C. B., S. P. SHI, AND J. YU (2015): "Testing for multiple bubbles: historical episode of exuberance and the collapse in the S&P 500." *International Economic Review*, 56, 1043–1078.

PHILLIPS, P. C. B., Y. WU, AND J. YU (2011): "Explosive behavior in the 1990s NASDAQ: When did exhuberance escalate asset values?" *International Economic Review*, 52, 201–226.

PHILLIPS, P. C. B. AND J. YU (2009): "Limit theory for dating the origination and collapse of mildly explosive periods in time series data." *Discussion Paper. Singapore Management University.*

PHILLIPS, P. C. B. AND J. YU (2011): "Dating the timeline of financial bubbles during the subprime crisis." *Quantitative Economics*, 2, 455–491.

SAID, S. E. AND D. A. DICKEY (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." *Journal of Business & Economic Statistics*, 7, 147–159.

Shi, S. P., A. Valadkhani, R. Smyth, and F. Vahid (2016): "Dating the timeline of house price bubbles in Australian capital cities." *Economic Record*, 92, 590–605.

Research Papers 2016

CREATES
Center for Research in Econometric
Analysis of Time Series

2016-25: Gustavo Fruet Dias, Marcelo Fernandes and Cristina M. Scherrer: Improving on daily measures of price discovery

2016-26: Martin M. Andreasen, Tom Engsted, Stig V. Møller and Magnus Sander: Bond Market Asymmetries across Recessions and Expansions: New Evidence on Risk Premia

2016-27: Kim Christensen, Ulrich Hounyo and Mark Podolskij: Testing for heteroscedasticity in jumpy and noisy high-frequency data: A resampling approach

2016-28: Kim Christensen, Roel Oomen and Roberto Renò: The Drift Burst Hypothesis

2016-29: Hossein Asgharian, Charlotte Christiansen, Rangan Gupta and Ai Jun Hou: Effects of Economic Policy Uncertainty Shocks on the Long-Run US-UK Stock Market Correlation

2016-30: Morten Ørregaard Nielsen and Sergei S. Shibaev: Forecasting daily political opinion polls using the fractionally cointegrated VAR model

2016-31: Carlos Vladimir Rodríguez-Caballero: Panel Data with Cross-Sectional Dependence Characterized by a Multi-Level Factor Structure

2016-32: Lasse Bork, Stig V. Møller and Thomas Q. Pedersen: A New Index of Housing Sentiment

2016-33: Joachim Lebovits and Mark Podolskij: Estimation of the global regularity of a multifractional Brownian motion

2017-01: Nektarios Aslanidis, Charlotte Christiansen and Andrea Cipollini: Predicting Bond Betas using Macro-Finance Variables

2017-02: Giuseppe Cavaliere, Morten Ørregaard Nielsen and Robert Taylor: Quasi-Maximum Likelihood Estimation and Bootstrap Inference in Fractional Time Series Models with Heteroskedasticity of Unknown Form

2017-03: Peter Exterkate and Oskar Knapik: A regime-switching stochastic volatility model for forecasting electricity prices

2017-04: Timo Teräsvirta: Sir Clive Granger□s contributions to nonlinear time series and econometrics

2017-05: Matthew T. Holt and Timo Teräsvirta: Global Hemispheric Temperatures and Co–Shifting: A Vector Shifting–Mean Autoregressive Analysis

2017-06: Tobias Basse, Robinson Kruse and Christoph Wegener: The Walking Debt Crisis

2017-07: Oskar Knapik: Modeling and forecasting electricity price jumps in the Nord Pool power market

2017-08: Malene Kallestrup-Lamb and Carsten P.T. Rosenskjold: Insight into the Female Longevity Puzzle: Using Register Data to Analyse Mortality and Cause of Death Behaviour Across Socio-economic Groups

2017-09: Thomas Quistgaard Pedersen and Erik Christian Montes Schütte: Testing for Explosive Bubbles in the Presence of Autocorrelated Innovations