Journal of Financial Stability 36 (2018) 66–81

Contents lists available at ScienceDirect

Journal of Financial Stability

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

Can bubble theory foresee banking crises?

Timo Virtanen[a], Eero Tölö[b,*], Matti Virén[a,c], Katja Taipalus[b]

[a] *Turku School of Economics, University of Turku, Finland*
[b] *Financial Stability and Statistics Department, Bank of Finland, Finland*
[c] *Monetary Policy and Research Department, Bank of Finland, Finland*

ARTICLE INFO

Article history:
Received 23 March 2017
Received in revised form
15 November 2017
Accepted 15 February 2018
Available online 21 February 2018

JEL classifications:
G01
G14
G21

Keywords:
Financial crises
Unit root
Combination of forecasts

ABSTRACT

We consider the effectiveness of unit root exuberance tests in predicting banking crises. Using a sample of 15 EU countries over the past three decades, our crisis dating follows the scheme of the European Systemic Risk Board. The exuberance indicators slightly outperform benchmark signaling and logit models. Variables based on credit- and debt-service are identified as better predictors than housing market variables, which in turn outperform stock market variables. The results corroborate the existing literature, which says financial crises are typically preceded by leveraged bubbles, and more specifically, that initial bubble signals from explosive growth in credit and asset prices are followed by a lift-off in debt-servicing costs as a financial crisis nears. The risk of financial crisis peaks just after the bubble bursts. Our results indicate that exuberance tests, which can be used in crisis prediction in a manner similar to conventional early warning models, may be readily incorporated into the toolkit of financial stability supervisors.

1. Introduction

Due to the high costs of 2007–2009 Great Recession and its aftermath, policymakers in many countries are keen to understand factors that predict financial crises.[1] A survey by Kauko (2014) suggests that hefty rises asset prices and indebtedness are among the most reliable predictors of factors foreshadowing banking crises. Schularick and Taylor (2012), who characterize financial crises as credit booms gone bust, further note that bubbles and their temporal arc are often apparent in hindsight from relevant time series. For policymakers, however, the challenge to financial stability policy is not characterizing the bubble or its collapse after the fact, but real-time determination of whether a bubble is developing. Timely and accurate detection of vulnerabilities enables more efficient macroprudential actions to counter financial instabilities.[2]

Identifying leveraged bubbles is an alien exercise to most economic forecasters. Beyond the long-running discussion on the rational and behavioral origins of bubbles, there is no generally accepted definition of financial crisis. Faced with such complexities, policymakers usually resort to a varied set of indicators and models. Priority goes to easy-to-interpret analytical frameworks that incorporate new information with minimal time lags.

This does not mean that we should adopt a completely a theoretical approach. In the following we test early-warning applications of the rational bubble theory elaborated by Campbell et al. (1997), Campbell and Shiller (1988a, 1988b), Craine (1993), Koustas and Serletis (2005), and others. As a benchmark, we draw upon the methodologies of Alessi et al. (2015) and Detken et al. (2014). These approaches include Bayesian methods, decision trees, and logit/probit models.[3]

Rational bubble theory reveals the existence of asset bubbles through examination of the time series behavior of stock prices and dividends. While the early literature in this area deals with stock price indices, rational bubbles tests have since been adapted for other asset classes. Escobari and Jafarinejad (2016), for example, date-stamp bubbles in REITs. A number of papers substitute rents for dividends to study house price bubbles (e.g. Anundsen et al., 2016; Efthymios et al., 2016; Pavlidis et al., 2016; Wan, 2015).

[*] Corresponding author at: Financial Stability and Statistics, Department Bank of Finland, PO Box 160, FI-00101 Helsinki, Finland.
E-mail address: eero.tolo@bof.fi (E. Tölö).

[1] Lawrence Ball (2014), for example, estimates that the cost from financial crises to 23 OECD economies in the aftermath since 2008 has averaged about 8.4% of GDP. See also Reinhart and Rogoff (2009a) for extensive review of costs and Honohan (2016) for lessons from the Irish financial crisis.

[2] See de Haan et al. (2017) for a broader perspective on effectiveness and implementation of macroprudential regulations.

[3] Both of these cited articles provide good reviews of available methods.

https://doi.org/10.1016/j.jfs.2018.02.008
1572-3089/© 2018 Elsevier B.V. All rights reserved.

Anundsen et al. (2016) even extends the method to the debt-to-GDP ratio, with income growth in a macro-setting playing a similar role to dividend growth. This is obviously true in a world where the functional distribution of income is constant.

The purpose of this study is to probe ex-ante financial crisis prediction performance of exuberance indicators on a set of 15 European advanced economies in the period 1980–2012. The performance evaluation is based on the two standard measures from the early-warning literature: Relative Usefulness (RU) and Area Under the Receiver Operating Characteristic curve (AUROC). Both measures rely on the frequency of missed crises and false alarms. Our crisis dating is taken from data of the European Systemic Risk Board (Detken et al., 2014). Moreover, we use a one- to three-year prediction horizon and two-quarter publication lags. Our robustness checks use alternative prediction horizons and two alternative crisis datasets, the Laeven and Valencia (2012) crisis dataset and the recent ECB/ESRB EU crisis database (2017).

Our approach is similar in many ways to that of Anundsen et al. (2016), who employ exuberance indicators as a conditioning variable in a logit model that predicts financial crises. We include more variables, however, and directly test the performance of the exuberance indicators instead of including them as part of a larger model.

We first compare our results to Detken et al. (2014), replicating their methodology with our dataset. We find an improvement in performance compared with the standard signaling and logit models. The performance difference is quite small for in-sample results and larger for out-of-sample results.[4]

A comparison of a broader set of methods is also included in Alessi et al. (2015), but it relies solely on in-sample comparisons and suffers from differences across samples and variables. While our exuberance indicators underperform against some data-driven methods in our initial comparison, they reveal a trade-off between in-sample performance and out-of-sample performance driven by the number of free parameters of the model. When we account for this trade-off, our exuberance indicators show broadly similar or superior performance to other methods, suggesting they are worth including in the toolkit of financial stability supervisors.

Our results corroborate other findings in the early warning literature. Periods of explosive growth in variables such as real estate price-to-income, credit-to-GDP ratio, or debt service costs are linked strongly to financial crisis. These findings comport, among others, with Anundsen et al. (2016), Jordá et al. (2015), and their conclusions that asset price bubbles are more dangerous when credit is involved. The time-ordering is in line with results obtained for a standard early-warning system based on level of the indicators (see e.g. Drehmann and Juselius, 2014).

The first bubble signals are obtained early on when credit and asset prices experience rapid growth. In this case, the alerting lead (time between the warning signal and the incoming crisis) is typically ten to twelve quarters. About eight quarters before the crisis arrives, debt-servicing costs embark on an explosive growth path. The probability of financial meltdown hits its peak after the debt bubble has been building on average for 20 quarters.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature and theory, with Section 2.1 focusing on general literature on testing for financial bubbles, and Section 2.2 detailing the linkage of rational bubbles to unit roots and explosive processes. Section 3 sets forth our empirical approach. Sections 3.1 and 3.2 explain our implementation of exuberance tests and their application to a range of variables. Section 3.3 describes the performance evaluation framework, including usefulness and AUROC measures. Section 4 reviews the data. Section 5 presents the results, with Sections 5.1 and 5.2 examining single-variable and composite exuberance indicators. Section 5.3 discusses signal timing and patterns of signals, and Section 5.4 includes robustness checks against implementation parameters and two alternative crisis datasets. Section 6 concludes.

2. Theory

2.1. Origin of bubbles and testing for bubbles

Several salient features emerge in the extensive literature on financial bubbles (e.g. Scherbina, 2013; Scherbina and Schlusche, 2014; Brunnermeier, 2008; and Brunnermeier et al., 2013). We cannot really cover more than some key features here.

The generation of a bubble is generally seen to require some *heterogeneity of agents and formation of expectations*. The typical set-up (e.g. Barberis et al., 2017) uses two sets of agents; the first set with rational expectations (fundamental traders) and the second with backward-looking expectations (extrapolators). While it is unremarkable that extrapolators tend to bid up prices and cause bubbles in such a setting, fundamental traders (somewhat surprisingly) do not necessarily arbitrage away bubbles (see Scherbina, 2013).

Bubbles also originate from *behavioral patterns* (e.g. biased self-attribution, representativeness heuristic bias, or a “conservatism bias”) that lead to deviations from optimal Bayesian ways of processing information. These hypotheses, while appealing, pose serious challenges to empirical testing in light of available data.

Testing for bubbles *itself* is a challenge. Despite years of work, no dominant strategy in setting up a proper test has emerged. Tests often try to exploit the basic time-series properties of relevant time series. Empirical tests that deal with the presumed bubble properties of financial time series include Elliot (1999) and Elliot et al. (1996), which deal with the power of unit- root tests (specifically) with different initial observations; Kim et al. (2002), Busetti and Taylor (2004), Leybourne (1995), and Leybourne et al. (2006), which deal with testing changes in the persistence of time series; and Homm and Breitung (2012), whose method considers stock market applications of unit root tests. Phillips et al. (2011, 2015) use a right-tailed unit root test for detecting bubble-type behavior in time series, as well as develop a sup ADF (SADF) test statistic and derive its (limiting) distribution. The authors apply their testing procedure to several financial times series, demonstrating reasonably good ex-post prediction performance. A similar approach relying on standard ADF tests with a rolling window is found in Taipalus (2006) and Taipalus and Virtanen (2016).

Although some studies (e.g. Corsi and Sornette, 2014) attempt a general model of bubbles, the empirical work usually relies on some form of unit root testing. The few exceptions include Banerjee et al. (2013), who use a random coefficient autoregressive model, and Franses (2016), who tests the feedback between first and second differences of time series, thereby causing the time series to explode. In addition, variance bounds tests (e.g. Gurkaynak, 2008) have been used in the context of asset price formation. The sophisticated specification test proposed by West (1987) has only been applied in a few studies. In any case, there is still plenty of room for developing better tests and testing strategies.

2.2. Explosive processes

Early bubble theory, elaborated, among others, by Campbell et al. (1997), Campbell and Shiller (1988a, 1988b), Craine (1993), Koustas and Serletis (2005), is concerned with bubbles in stock

[4] Compared with the signaling benchmark, exuberance indicators show significantly better in-sample performance with the Laeven and Valencia (2012) crisis dataset, which is used in our robustness check (see Section 5.4).

prices. Specifically, a constantly demising dividend yield is treated as a sign of worsening overpricing. As prices rise, they should eventually be realized as higher dividends. If not, the price rise is not considered to be based on fundamentals. This is apparent from the log dividend-price ratio model derived by Campbell and Shiller (1988a):

$$d_t - p_t = \mathbb{E}_t[\sum_{j=0}^{\infty} \rho^j (r_{t+j} - \Delta d_{t+j})] - \frac{c-k}{1-\rho}, \qquad (1)$$

where d is the log dividend, p the log price, and r the discount rate. Constants ρ, c, and k are parameters from log-linear approximation. The formula is a generalization of the Gordon (1962) growth model for the case where dividend growth and rates of return change over time. Moreover, the equation is itself a solution to a linearized stochastic difference equation. The solution implicitly assumes a transversality condition $\lim_{t\to\infty} \rho^t (d_t - p_t) = 0$ that requires both $\rho < 1$ and a stationary $d - p$. The stationarity of r and Δd implies that the log dividend yield must also be stationary (for details, see Cochrane, 1992; Craine, 1993).

Presence of a unit root in the log dividend yield means that agents or their expectations are not rational (assuming no other fundamental market failures). A possible interpretation is a rational bubble. Indeed, this view has spawned numerous studies in which the stationarity properties of stock prices are examined using unit root testing procedures.

More recently, the bubble-testing literature has introduced methods that detect the change of dynamics from I(1) to an explosive process in the asset price time series (e.g. Phillips et al., 2011, 2015). These methods can be understood through the present value theory of the asset price, whereby we first define the fundamental asset price as the present value (discounted by some interest rate r_f) of expected future dividends and denote it by P_t^f. We then define a bubble as any deviation from the fundamental price so that the market price of an asset can be decomposed as:

$$P_t = P_t^f + B_t, \qquad (2)$$

where B_t is the price bubble component. A rational buyer is only willing to pay P_t^f for the asset unless the expected future value of the bubble component satisfies:

$$\mathbb{E}_t(B_{t+1}) = (1 + r_f)B_t, \qquad (3)$$

i.e. the existence of a rational bubble is based on expectations of future price growth. If (3) does not hold, B_t collapses to zero.[5]

The evolution of the asset price depends on the assumed evolution of the fundamentals. A usual assumption is that the fundamentals (e.g. dividends) follow a random walk process with a negligible drift. In such case, the fundamental asset price will also follow random walk with a drift. On the other hand, when a bubble is present, it implies that the asset price process becomes explosive as $r_f > 0$. The econometric bubble test is designed to detect this change in the time series process.

3. Empirical analysis

3.1. Selecting the set of variables

The derivations in the previous section directly apply to asset price time series such as stock prices, housing prices (with rents playing the role of dividend), and land prices. It is usual practice in empirical work to control for explosive fundamentals that would temporarily justify explosive asset prices by testing the price-to-fundamentals ratio (e.g. stock price index divided by the dividends index). Sometimes the fundamentals are hard to measure. For example, it is often difficult to find reliable data on rents when dealing with real estate markets, so a proxy measure might be considered. Pavlidis et al. (2016) suggest using the ratio of house prices to personal disposable income when reliable statistical data on rents is lacking. Thus, we include data on real stock and house prices, as well as the ratios of house prices to rent and income in our set of potential predictor variables.

Anundsen et al. (2016) extend the use of exuberance tests to the debt-to-GDP ratio. Intuitively, an explosively increasing credit stock or credit-to-GDP ratio can be regarded as an economic bubble, i.e. debt used excessively to finance production and/or consumption of assets in an economy. Income growth plays a similar role to dividend growth. Moreover, we know from conventional government debt accounting models (see Wilcox, 1989) that stationary growth rate of taxable income is incompatible with a continuously increasing (nonstationary) debt-to-GDP ratio. We also include measures of credit, bank credit, and household credit, as well as their ratios to GDP to test which of these has the best predictive power.

The use of variables such as interest rates and debt-to-income ratios is also customary in financial bubble prediction.[6] Technically speaking, use of unit root methods with some of these is defensible under the stationarity argument. Clearly debt-servicing costs cannot exceed disposable income, so the ratio of these variables at least should be stationary and probably well below the value of unity (cf. Wilcox, 1989).

3.2. Econometric exuberance test

Our empirical investigation into the link between financial bubbles and crises raises the issue of whether particular bubble signals derived from an asset price time series are useful as early warning indicators of financial crises. For this, we need a test that detects the periods within the time series when asset price growth is explosive.

From the rather extensive set of econometric tests designed to detect rational bubbles, we consider the generalized sup augmented Dickey-Fuller (GSADF) test of Phillips et al. (2015) due to its proven ability to date-stamp bubble periods when multiple bubbles are present in the time series. The test is based on the ADF regression:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta y_{t-i} + \varepsilon_t. \qquad (4)$$

In contrast to the ADF test, the null hypothesis here is I(1) and the alternative is an explosive root. To improve bubble detection in cases involving multiple bubbles in the time series, the GSADF test uses a windowing scheme where the endpoint of the window is rolled forward and the ADF regression is estimated for backwards-expanding subsamples of all possible lengths starting from a minimum size r_0. (See Phillips et al., 2015 for an illustration of this windowing scheme).

The ADF statistic for a window starting at r_1 and ending at r_2 is defined as:

$$ADF_{r_1}^{r_2} = \frac{\hat{\gamma}_{r_1}^{r_2}}{s.e.(\hat{\gamma}_{r_1}^{r_2})}, \qquad (5)$$

[5] See Phillips et al. (2011, 2015) for a detailed exposition.

[6] See Frankel and Saravelos (2010) for a useful review.

and the GSADF test statistic for a given minimum window r_0 as:

$$GSADF(r_0) = \sup_{\substack{r_2 \in [r_0, 1] \\ r_1 \in [0, r_2 - r_0]}} \{ADF_{r_1}^{r_2}\}. \tag{6}$$

Here, $r_i \in [0, 1]$ and $r_2 \neq 0$. 0 and 1 correspond to the first and last data point, respectively. The backwards sup ADF (BSADF) test statistic used to date-stamp the bubble periods is defined at the point r_2 as:

$$BSADF(r_0, r_2) = \sup_{r_1 \in [0, r_2 - r_0,]} \{\mathrm{ADF}_{r_1}^{r_2}\}. \tag{7}$$

The test statistics follow a non-standard distribution. Critical values are thus obtained through Monte Carlo simulation.

The GSADF test statistic provides means for inferring the presence of one or more bubbles in the time series. In the application at hand, we are only interested in date-stamping bubbles, rather than achieving an overall conclusion.[7] Hence, it suffices to calculate the BSADF statistic for the time series under study.

The test procedure has three parameters: the minimum window length (r_0T), the lag length of the ADF equation (p), and the significance level of the test (α). Phillips et al. (2015) suggest selecting the minimum window size as $r_0 = 0.1 + 1.8/\sqrt{T}$. We use this as a starting point, but also examine other minimum window sizes to check the robustness of the test procedure.

The AR-lag length can be fixed or determined based on information criteria. Phillips et al. (2015) conclude that the GSADF test suffers from size distortions with longer lag lengths, including when information criteria such as BIC or sequential significance testing are used to determine lag length. Thus, they propose using a small, fixed lag length in empirical work. Noting their suggestion, we report our results using fixed lag lengths of 2 and 4, i.e. the two most common lag lengths suggested by data using the BIC information criteria.

With the fixed lag length, serial correlation of the errors may remain an issue that affects the validity of the results. Pedersen and Schütte (2017) investigate the size properties and effect of serially correlated errors of the GSADF test. They propose a Sieve-bootstrap version of the GSADF test to alleviate the size distortions caused by serial correlation.

The significance level remains a free parameter that can be used to adjust the sensitivity of the test. We report the results a relative usefulness metric (to be discussed in next section) using the optimal α that maximizes the relative usefulness of the indicator. Results for fixed α are provided as a robustness check. The other performance measure, AUROC (see next section), essentially distills the information from the indicator to a single performance measure considering all possible values of α.

The BSADF test may also identify short-run explosive contractions in the time series ("negative bubbles") as bubble periods. In practice, this means that the methods also elicit warning signals from e.g. stock market crashes or contractions in house prices. We do consider such signals relevant as they typically occur after the crisis is underway. Thus, when calculating usefulness values, we remove all warning signals that occur when the corresponding time series is decreasing in value.

3.3. Performance evaluation

To assess the predictive performance of the exuberance tests in predicting financial crises, we measure how frequently the bubble signals correctly precede known crises in the data. We take the approach commonly used in banking crisis early warning literature and define a pre-crisis window to start at three years before and to end one year after each crisis (see e.g. Detken et al., 2014). A good model should issue warning signals when the economy is still in this pre-crisis state.

The performance of the different variables is ranked using the *Relative Usefulness* measure in Alessi and Detken (2011), which draws upon the policy loss functions of Demirgüc-Kunt and Detragiache (2000) and Bussière and Fratzscher (2008). The loss function of Alessi and Detken (2011) is defined as follows:

$$L(\theta) = \theta T_2 + (1-\theta) T_1 = \theta \frac{C}{A+C} + (1-\theta) \frac{B}{B+D}, \tag{8}$$

where the right-hand side is the weighted average of type I and type II error rates, T_1 and T_2, respectively.[8] The weights θ and $(1-\theta)$ in the loss function reflect the policymaker's presumed preferences for type I and type II errors. A parameter value θ higher than 0.5 means that the policymaker is more averse to missing a signal of an upcoming crisis than to receiving a false alarm. As is commonly done in the literature, we set $\theta = 0.5$, but also present results for $\theta = 0.6$ and $\theta = 0.7$. Note, however, that the relative numbers of missed crises and false alarms that a given policy parameter entails is not directly related to θ, but depends on the unconditional probability for a period to be a pre-crisis period (see Sarlin, 2013). Typically, this probability is small, so that even with $\theta = 0.5$ the number of false alarms becomes much larger than the number of missed crises (in our case, about 10 times higher than the number of missed crises).

In Eq. (8), A is the number of periods in which an indicator provides a correct signal (crisis starts within 1–3 years of issuing the signal), and B the number of false alarms. C is the number of periods that miss a crisis, i.e. a signal is not generated during a defined period from the onset of the crisis (1–3 years). D denotes the number of periods in which a signal is correctly not provided. In other words, A is the number of true positives; B is the number of false positives; C is the number of false negatives; and D is the number of true negatives.

Absolute Usefulness is defined as $\min(\theta, 1-\theta) - L$. The corresponding Relative Usefulness statistic is normalized as:

$$RU = \frac{\min(\theta, 1-\theta) - L}{\min(\theta, 1-\theta)}. \tag{9}$$

Because L is non-negative, RU is bounded from above. A maximal RU of 1 means that the indicator can perfectly indicate all the pre-crisis periods and produces no false positives. 0 or less means that the indicator is not useful in distinguishing the pre-crisis periods.

As an additional measure, we calculate the AUROC for each variable. AUROC and RU may rank the indicators differently, as usefulness is calculated at a particular sensitivity level, while AUROC takes into account all sensitivity levels. The ROC curve is a commonly used tool for evaluating the performance of a binary classifier. It plots the true positive rate (TP) against the false positive rate (FP) for varying values of a threshold parameter that controls

[7] The BSADF test may indicate a bubble at some point even if the GSADF test indicates that the time series has no bubbles. Thus, short isolated bubble signals should usually be omitted.

[8] In the formula, the order of T_1 and T_2 differs a bit from Alessi and Detken (2011). This is a matter of convention in forming the null hypothesis. Here, a type I error (false positive) is the incorrect rejection of a true null hypothesis H_0. We set H_0, i.e. "no crisis within the next 3 years" so that a false positive indicates a false alarm. A type II error (false negative) incorrectly retains a false null hypothesis. Thus, our false negative here means failure to detect a crisis.

the sensitivity of the classifier. AUROC is then defined, as its name suggests, as the area that falls below this curve. In our case, the varying threshold parameter is the significance level and we calculate AUROC as:

$$AUROC = \int_0^1 TP\left(FP\left(\alpha\right)\right)FP'\left(\alpha\right)d\alpha \tag{10}$$

where α is the significance level of the BSADF test. Increasing α makes the test more sensitive and increases the number of issued signals. We calculate the true and false positive rates produced by the BSADF test at varying significance levels and approximate (10) by the trapezoidal rule.

AUROC provides a numeric measure of the predictive performance of an early warning indicator that is independent of policymaker preferences.[9] By definition, the value of AUROC lies between 0 and 1. Both extreme values mean that the classifier is "perfect," i.e. it can always predict the correct class regardless of the threshold parameter used (if AUROC is 0, the decision needs to be inverted). The value of 0.5 means that the classifier is completely uninformative, i.e. true and false positives are equally likely at all sensitivity levels.

4. Data

The empirical analysis makes use of data from the following EMU countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain. We also include Denmark, United Kingdom, and Sweden in our sample. Available data extend well back into the 1970s, when the regulatory environment was quite different from today. Administrative credit rationing was still common in many countries and the functioning of financial markets was markedly different from the present system. To limit possible bias due to structural changes in banking, we start our sample from 1980.

Table 1 presents the indicator data. Most of the indicators are based on a quarterly data series compiled by the ECB and shared within the macro-prudential analysis group (MPAG). The variables include credit-to-GDP ratios, credit aggregates, debt-servicing costs, residential real estate prices, and stock prices. Most of the data is publicly available from BIS, OECD, Dallas Fed, or stock market data providers. The debt-servicing cost variables are based on ECB calculations, but similar (albeit slightly shorter) data series are publicly available on the BIS website. To account for publication lags, all quarterly data are lagged by one quarter in the evaluation.

Evaluating the quality of the warning signals requires data on financial crises. Choosing the right period is critical in such analysis as wrong choices automatically invalidate the results.[10] Various banking crisis datasets have been provided in the earlier literature.[11] We use the crisis classification scheme of the ESRB (Detken et al., 2014), which is based extensively on expert opinion from individual central banks. It is closely related to the Heads of Research (HoR) Database compiled by a team at the Czech National Bank in collaboration with the ESCB (Babecký et al., 2012). The HoR database defines banking crisis based on bank runs, significant losses in the banking system, or significant public intervention in the banking system to prevent such losses. Detken et al. (2014) modify the HoR database by removing non-systemic crises or unrelated to financial cycle and by adding "would-be" crises, i.e. events where financial cycle would most likely have led to crisis had a policy intervention or exogenous event not prevented it.

For our robustness check, we consider two alternative crisis period definitions: the systemic banking crises database of Laeven and Valencia (2012) and the new ECB/ESRB EU crises database (Lo Duca et al., 2017). Earlier crisis databases include Caprio and Klingebiel (1996), Demirgüç-Kunt and Detragiache (1998), and Reinhart and Rogoff (2009b). The crisis dates for the three crisis datasets are shown in Table 2.

5. Empirical results

Section 5.1 provides the basic result demonstrating the usefulness of exuberance tests in signaling financial crises. Section 5.2 shows that combining signals from multiple indicators enhances our crisis predictions. Section 5.3 returns to single indicators to study the timing and pattern of signals. Finally, Section 5.4 provides robustness checks with regards to model parameters and crisis definitions.

5.1. Signals from single variable indicators

To illustrate the performance of our approach, we offer a set of signal graphs for pairs of predictor variables in Figs. 1 and 2 (bank credit-to-GDP ratios and house price-to-income ratios). The graphs show the development of the underlying variable, the bubble stamping signals from the BSADF method (full height bars), as well as the pre-crisis (half-height bars) and crisis period (short bars). Visual inspection reveals that the indicators all perform quite well in terms of signaling alarms well in advance of crisis onset. For example, the signals from credit-to-GDP (Fig. 1) successfully signal the most recent financial crisis in Denmark, Spain, France, Greece, Ireland, Portugal, and Sweden. However, the indicator also alarms for Austria, Belgium, and Italy, none of which experienced systemic crises according to the crisis dataset.[12] The visual inspection also shows that in most cases the period of explosive growth of house prices or the credit-to-GDP ratio begun many years before the outbreak of the crisis and often well before the defined pre-crisis period. This is especially true for the 2008 financial crisis; the bubble periods in earlier crises in house prices and debt seem to have been shorter.

Table 3 reports the performance of the exuberance indicators using the whole sample. Results for the in-sample optimized signaling approach are presented for comparison (see e.g. Detken et al., 2014). The policymaker's preference parameter θ is set to 0.5 and the sensitivity parameter α is optimized in sample to maximize usefulness.[13] Robustness checks with higher values of θ and fixed α are reported in Section 5.4. The exuberance indicators and benchmark signaling model have quite similar performance. Compared to the exuberance indicators, the usefulness statistics of the signaling method are slightly higher and the AUROC statistics slightly lower on average, when the whole sample is used for evaluation. However, none of the differences in AUROC statistics are statistically

[9] See e.g. Drehmann and Juselius (2014) for discussion on the use of AUROC in this context.

[10] Recent evidence is provided in Ristolainen (2017).

[11] See Tölö et al. (2018) for a comprehensive survey table showing the crisis datasets used in various studies.

[12] The banking sectors in these countries still suffered considerable losses during 2008–2012. According to the crisis definitions of Laeven and Valencia (2012), Austria, Belgium, and Italy experienced systemic banking crises from 2008 onwards. In the robustness checks below, we not that our results improve when use this crisis dataset.

[13] We calculate the usefulness values for $\alpha \in [0.01, 0.2]$ and find the maximum. In a few cases, values less than 0.8 yield a higher usefulness (at the expense of high rates of false positives), but they are highly unconventional for right-tailed hypothesis testing. Constraining the range of parameter values may have a minor deteriorating effect on performance of the exuberance indicators.

Table 1
Names and definition of the variables.

Variable	Source	Countries	Obs	Content
Bank credit-to-GDP	BIS	15	1857	Ratio of (nominal) bank credit to the private non-financial sector to (nominal) GDP
Real bank credit	BIS	15	1857	Bank credit to private non-financial sector, local currency (real) in billions
Total credit-to-GDP	BIS	15	1857	Ratio of (nominal) total credit to the private non-financial sector to (nominal) GDP
Real total credit	BIS	15	1857	Total credit to private non-financial sector, in billions of local currency (real)
Household credit-to-GDP	BIS	15	1520	Ratio of (nominal) total credit to households to (nominal) GDP
Real household credit	BIS	15	1520	Total credit to households, in billions of local currency (real)
Debt service ratio	ECB	15	1828	Debt service to income ratio, households and non-financial corporations
Household debt service ratio	ECB	15	1267	Debt service to income ratio, households
Corporate debt service ratio	ECB	15	1279	Debt service to income ratio, non-financial corporations
Residential RE price-to-income	OECD	15	1595	Residential real estate price to income index
Residential RE price-to-rent	OECD	15	1645	Residential real estate price to rent index
Real residential RE price	OECD	15	1658	Residential property real price index
House price-to-income	Dallas Fed	12	1560	House prices to income ratio (Dallas Fed international house price database)
Real house price	Dallas Fed	12	1560	Real house prices (Dallas Fed international house price database)
Real stock price index	Bloomberg	15	1769	Stock price index (real)

"Obs" is the total number of observations for each variable. "Countries" is the number of countries for which a corresponding time series is available. "Lagged" indicates that the time series is lagged by one quarter to correspond to a typical publication lag of statistics.

Table 2
Crisis dates.

Crisis dataset	ESRB (Detken et al., 2014)		Laeven and Valencia (2012)		ECB/ESRB (Lo Duca et al., 2017)	
Country	Start	End	Start	End	Start	End
Austria			2008Q1	2011Q4	2008Q1	2012Q4
Belgium			2008Q1	2011Q4	2007Q4	2012Q4
Germany					1974Q3	1974Q4
Germany	2000Q1	2003Q4			2001Q1	2003Q4
Germany			2008Q1	2011Q4	2007Q4	2012Q4
Denmark	1987Q1	1993Q4			1987Q2	1994Q4
Denmark	2008Q3	2012Q4	2008Q1	2011Q4	2008Q1	2012Q4
Spain	1978Q1	1985Q3	1977Q1	1980Q3	1978Q1	1985Q3
Spain	2009Q2	2012Q4	2008Q1	2011Q4	2009Q2	2012Q4
Finland	1991Q3	1995Q4	1991Q1	1995Q4	1991Q3	1996Q4
France	1993Q3	1995Q4			1991Q3	1995Q1
France	2008Q3	2012Q4	2008Q1	2011Q4	2008Q2	2009Q4
Great Britain	1973Q4	1975Q4			1973Q4	1975Q4
Great Britain	1990Q3	1994Q2			1991Q3	1994Q1
Great Britain	2007Q3	2012Q4	2007Q1	2011Q4	2007Q3	2009Q4
Greece	2008Q1	2012Q4	2008Q1	2011Q4	2010Q2	2012Q4
Ireland	2008Q3	2012Q4	2008Q1	2011Q4	2008Q3	2012Q4
Italy	1994Q1	1995Q4			1991Q3	1997Q4
Italy			2008Q1	2011Q4	2011Q3	2012Q4
Luxembourg			2008Q1	2011Q4	2008Q1	2010Q3
Netherlands	2002Q1	2003Q4				
Netherlands	2008Q3	2012Q4	2008Q1	2011Q4	2008Q1	2012Q4
Portugal	1999Q1	2000Q1			1983Q1	1985Q1
Portugal	2008Q4	2012Q4			2008Q3	2012Q4
Sweden	1990Q3	1993Q4	1991Q1	1995Q4	1991Q1	1997Q2
Sweden	2008Q3	2010Q4	2008Q1	2011Q4	2008Q3	2010Q4

significant.[14] Credit-to-GDP ratios, debt service ratios and house price-to-income ratios seem to work well with both methods.

In Table 4, we consider an out-of-sample evaluation, where the policymaker decides on the parameter α and threshold for the signaling method based on 1980–1999 training data given the policymaker's preference parameter $\theta = 0.5$. The performance is subsequently measured in the following period (2003–2012).[15] In the out-of-sample evaluation, the exuberance indicators are more robust than the signaling method and produce higher usefulness and AUROC statistics on average. Results for higher θ and fixed α are reported in Section 5.4 as robustness check. It turns out that optimizing α with a training period, as opposed to using fixed α, has only a small effect on the usefulness. This means our initial guess

[14] Standard errors clustered by country, which we have omitted here for the sake of brevity, are typically about 0.04–0.05.

[15] We do not use the years 2000, 2001, and 2002 in the training, because a policymaker in 2003 could not know whether to classify these years as pre-crisis or tranquil periods.

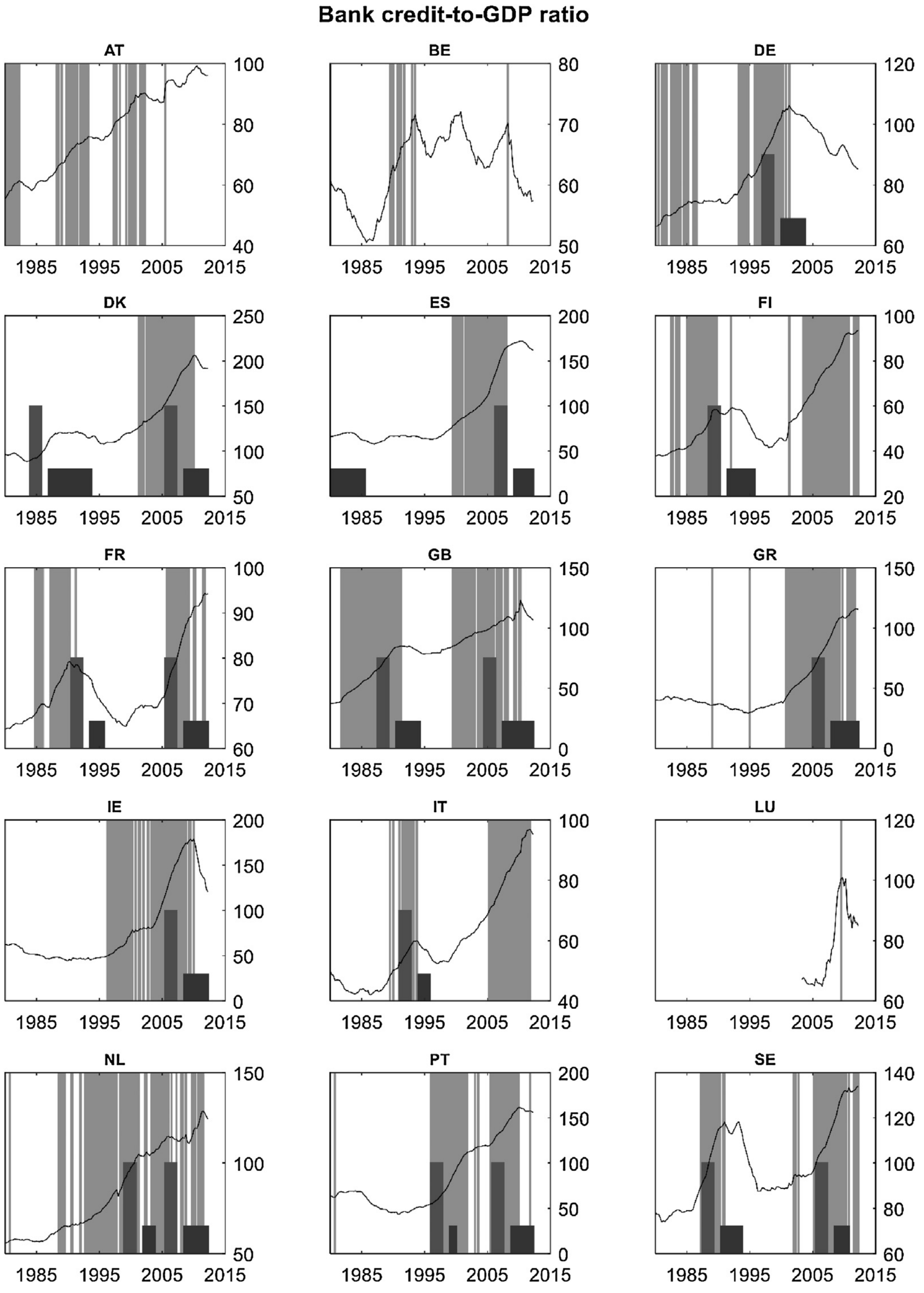

Fig. 1. The line curve is the bank credit-to-GDP ratio (BIS data). The full height shaded areas are bubble alerts from the BSADF method. Half-height shaded areas are pre-crisis periods (12–5 quarters before crisis). The dark short areas are the financial crisis periods in the ESRB crisis dataset. BSADF parameters are the value of the significance level parameter $\alpha = 0.05$, minimum window length = 24 quarters and lag length = 2.

of $\alpha = 0.05$ was quite reasonable, and that there is not enough past data to improve on it. This also highlights one benefit we claim the exuberance indicators have over other methods, namely that it is often unnecessary to rely on extensive historical data to estimate any parameters.

Summarizing the single variable results so far, what we have vaguely characterized as a “debt bubble” seems to be a good predictor of an upcoming financial crisis. Highest usefulness values are obtained with the credit-to-GDP ratios. The household debt service ratio is also among the most useful indicators, even though using the exuberance indicator with it is not directly supported by existing theory. Indicators derived from real estate prices also perform well, but generally rank lower than the debt-based variables. Usefulness of the real estate data also depends on the data source. The Dallas Fed’s International House Price Database seems to provide the most useful data source in this evaluation. While it only includes 12 of the 15 countries in our evaluation, the result is the same also when looking at the common sample. The periods of explosive growth preceding a financial crisis are generally longer than our pre-crisis window. This can be seen clearly in the short

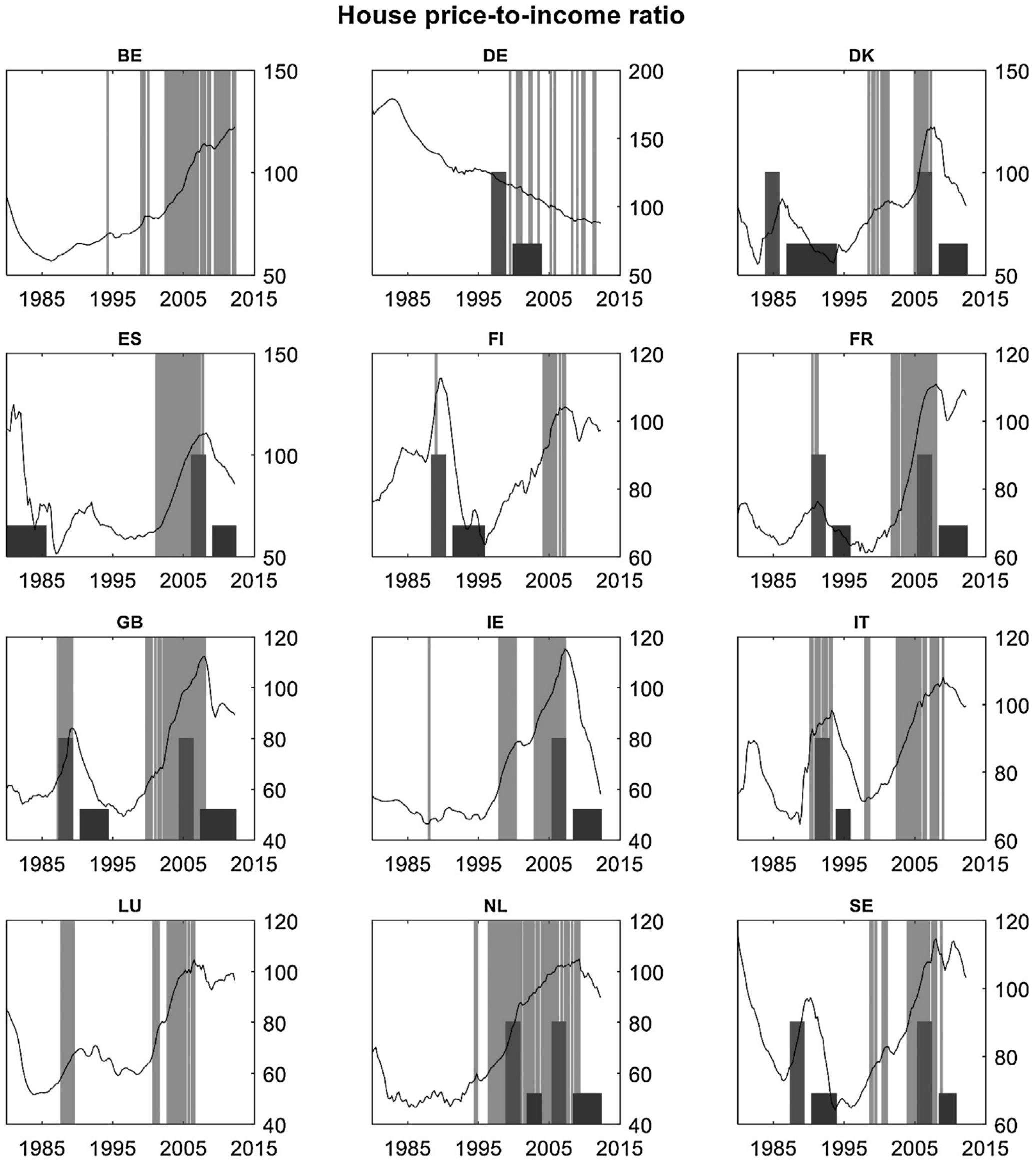

Fig. 2. The line curve is the house price-to-income ratio (Dallas Fed data). The full-height shaded areas are bubble alerts from the BSADF method. The half-height shaded areas are pre-crisis periods (12–5 quarters before crisis). The dark short areas are financial crisis periods in the ESRB crisis dataset. BSADF parameters are the value of the significance level parameter $\alpha = 0.05$, the minimum window length = 24 quarters, and the lag length = 2.

sample evaluation (Table 4). False positive rates of most variables are high owing to the explosive growth of credit and house prices that began already in the early 2000s in most countries.

5.2. Aggregating signals from multiple indicators

Hypothesizing that a policymaker could benefit from combining several indicators, we form composite exuberance indicators by taking a weighted sum of signals from N exuberance indicators:

$$A_N = w_1 I_1 + w_2 I_2 + \ldots + w_N I_N, \tag{11}$$

where w_i are the weights and each I_i is either 1 (exuberance indicator i issues an alert) or 0 (no alert). A warning signal is issued when the composite signal is above some threshold. For simplicity and to avoid the perils of overfitting, we only consider uniform weights:

$$A_N = I_1 + I_2 + \ldots + I_N. \tag{12}$$

We define a warning signal W_N so that $W_N = 1$ if $A_N \geq T$ and zero otherwise, where $T = [1, 2, \ldots, N]$ is a threshold value. An alert from a single component is called a "sub-alert," and T is the number of simultaneous sub-alerts required for the overall composite indicator to alert.

The usefulness of the composite exuberance indicator is evaluated using the same methodology as in the previous section. We estimate benchmark logit models where the explanatory variables are the transformations of those variables used in the composite model and calculate the RU and AUROC values. The benchmarks should be good as we select the components for three composite indicators based on the best performing logit model in Detken et al. (2014), their Table F3. We report the RU and AUROC with different alerting thresholds using the full 1980–2012 sample and the short sample from 2003 onwards. The results are presented in the top and middle panel of Table 5. The lower panel shows the variables included in each model. The policymaker's preference is set to $\theta = 0.5$.

Combining different exuberance indicators improves performance markedly compared to the single variable case as we require a reasonable number of sub-alerts (two or three). All three

Table 3
In-sample performance of individual exuberance indicators.

		θ = 0.5				Signaling	
Variable	AUROC	α	RU	FP	FN	RU	AUROC
Bank credit-to-GDP	0.830	0.05	0.55	0.29	0.16	0.52	0.833
Real bank credit	0.817	0.01	0.50	0.36	0.15	0.46	0.793
Total credit-to-GDP	0.804	0.03	0.56	0.29	0.15	0.47	0.774
Real total credit	0.813	0.01	0.43	0.49	0.08	0.38	0.754
Household credit-to-GDP	0.749	0.05	0.44	0.42	0.15	0.59	0.851
Real household credit	0.780	0.01	0.40	0.47	0.13	0.39	0.749
Debt service ratio	0.762	0.08	0.39	0.11	0.50	0.36	0.704
Household debt service ratio	0.812	0.15	0.52	0.23	0.25	0.41	0.765
Corporate debt service ratio	0.636	0.20	0.23	0.14	0.63	0.35	0.631
Residential RE price-to-income	0.693	0.11	0.22	0.21	0.57	0.32	0.724
Residential RE price-to-rent	0.641	0.20	0.22	0.32	0.46	0.30	0.673
Real residential RE price	0.635	0.10	0.23	0.31	0.46	0.38	0.715
House price-to-income	0.766	0.07	0.42	0.21	0.37	0.50	0.817
Real house price	0.726	0.03	0.42	0.29	0.29	0.46	0.787
Real stock price index	0.625	0.20	0.15	0.22	0.63	0.29	0.622

This table presents in-sample performance statistics for individual exuberance indicators and the benchmark signaling method. The significance level α of the BSADF test and the signaling threshold are optimized to produce maximum relative usefulness. The policymaker's preference θ with respect to false alarms (FP) and missed crises (FN) is set to 0.5.

Data cover 1980–2012. BSADF parameters are the value of the significance level parameter α, minimum window length = 24 quarters and lag length = 2. For the signaling benchmark, the variables are transformed using 2-year differences or growth rates (except debt service ratios, which are in levels), and the signaling threshold is optimized based on the full sample 1980–2012. RU ≤ 1 is the relative usefulness given the policymaker's preference parameter θ (see Section 3.2 for definition), absolute usefulness = Min(θ,1 − θ)RU, FP is False Positive Rate, FN is False Negative Rate. 0 ≤ AUROC ≤ 1 is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables, except for equity prices. Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period runs from 12 to 5 quarters before each crisis.

Table 4
Out-of-sample performance of individual exuberance indicators.

		θ = 0.5				Signaling	
Variable	AUROC	α	RU	FP	FN	RU	AUROC
Bank credit-to-GDP	0.798	0.05	0.47	0.46	0.07	0.31	0.757
Real bank credit	0.798	0.01	0.45	0.47	0.08	0.32	0.740
Total credit-to-GDP	0.702	0.04	0.30	0.57	0.13	0.24	0.664
Real total credit	0.783	0.01	0.33	0.67	0.00	0.22	0.711
Household credit-to-GDP	0.714	0.06	0.29	0.66	0.05	0.49	0.804
Real household credit	0.730	0.01	0.26	0.68	0.06	0.22	0.697
Debt service ratio	0.743	0.06	0.43	0.20	0.38	−0.13	0.413
Household debt service ratio	0.816	0.07	0.49	0.28	0.23	0.23	0.734
Corporate debt service ratio	0.653	0.20	0.25	0.16	0.59	−0.17	0.405
Residential RE price-to-income	0.496	0.11	0.04	0.49	0.47	0.15	0.602
Residential RE price-to-rent	0.485	0.19	0.03	0.54	0.43	0.10	0.550
Real residential RE price	0.528	0.06	0.14	0.50	0.36	0.12	0.540
House price to income	0.603	0.19	0.21	0.69	0.11	0.44	0.710
Real house price	0.647	0.05	0.27	0.70	0.04	0.34	0.708
Real stock price index	0.704	0.07	0.14	0.08	0.78	0.16	0.598

This table presents the out-of-sample performance statistics for individual exuberance indicators and the benchmark signaling method. Data between 1980 and 1999 are used to estimate the parameters and 2003–2012 are used for the evaluation. The policymaker's preference θ with respect to false alarms (FP) and missed crises (FN) is set to 0.5.

Data cover 2003–2012. The significance level α of the BSADF test is optimized based on training with 1980–1999 data. Other BSADF parameters are minimum window length = 24 quarters and lag length = 2. For the signaling benchmark, the variables are transformed using 2-year differences or growth rates (except debt service ratios, which are in levels), and the signaling threshold is updated each period using a training sample. RU ≤ 1 is the relative usefulness given the policymaker's preference parameter θ (see Section 3.2 for definition), absolute usefulness = Min(θ,1 − θ)RU, FP is False Positive Rate, FN is False Negative Rate. 0 ≤ AUROC ≤ 1 is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables (except for equity prices). Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period is from 12 to 5 quarters before each crisis.

models have quite similar performance, but the highest relative usefulness and AUROC are obtained with the third model. For Model 3, the variation that requires two sub-alerts produces a relative usefulness of 0.63 and AUROC of 0.841. Using the variation that requires three sub-alerts, the relative usefulness is 0.53 and AUROC 0.855. Overall, the comparison between the composite exuberance indicators and the logit model yields similar results as the previous comparison for single variables. Even if the logit model yields better relative usefulness and AUROC values in-sample, the composite exuberance indicators are more robust and outperform the logit models in the out-of-sample evaluation.

It is also helpful to compare the performance of our composite indicators to a larger set of early warning models. For this purpose, we utilize a comparison study prepared by the ECB's Macroprudential Research Network (Alessi et al., 2015). The study features a family of promising early warning modeling frameworks, including Bayesian model averaging, a Bayesian random coefficient model, decision trees, a dynamic dependent variable model, and probit/logit models. The set-up bears enough similarities to ours that we obtain a closely matched comparison.[16] The dataset is some subset of EU countries for each study and the time period is from

[16] Results are available for the 4- to 12-quarter prediction horizons on request.

Table 5
Performance statistics for composite exuberance indicators.

	Model 1				Model 2				Model 3			
Threshold	RU	FP	FN	AUROC	RU	FP	FN	AUROC	RU	FP	FN	AUROC
Full Sample (1980–2012)												
1	0.35	0.58	0.08	0.750	0.44	0.50	0.06	0.784	0.52	0.42	0.06	0.779
2	**0.59**	**0.24**	**0.17**	**0.847**	0.60	0.31	0.09	0.832	0.63	0.26	0.11	0.841
3	0.45	0.08	0.47	0.768	**0.56**	**0.13**	**0.31**	**0.850**	**0.53**	**0.12**	**0.35**	**0.855**
4	0.20	0.02	0.78	0.804	0.40	0.04	0.56	0.793	0.34	0.05	0.62	0.839
Benchmark	0.58	0.13	0.28	0.874	0.63	0.11	0.26	0.886	0.59	0.18	0.23	0.879
Small sample (2003–2012)												
1	0.28	0.70	0.01	0.646	0.31	0.68	0.01	0.651	0.29	0.71	0.00	0.645
2	**0.60**	**0.26**	**0.14**	**0.823**	0.47	0.42	0.11	0.742	0.45	0.50	0.04	0.739
3	0.44	0.06	0.50	0.710	**0.53**	**0.18**	**0.29**	**0.804**	**0.57**	**0.26**	**0.17**	**0.840**
4	0.19	0.01	0.80	0.825	0.36	0.03	0.61	0.732	0.39	0.11	0.50	0.793
Benchmark	0.28	0.64	0.09	0.752	0.35	0.61	0.04	0.820	0.45	0.48	0.07	0.792

	Variable specifications		
Model	Model 1	Model 2	Model 3
Variable 1	Bank credit-to-GDP	Bank credit-to-GDP	Bank credit-to-GDP
Variable 2	Debt service ratio	Debt service ratio	Household debt service ratio
Variable 3	House price to income	House price to income	House price to income
Variable 4	Real stock price index	Total credit-to-GDP	Total credit-to-GDP

In this table, the top (middle) panel presents the in-sample (out-of-sample) performance statistics for composite exuberance indicators and the benchmark logit model (Detken et al., 2014). The upper panel uses the full sample of data covering 1980–2012. In the middle panel, the 2003–1999 data are used to estimate the parameters and the 2003–2012 are used for the evaluation. For each composite exuberance indicator, there are four rows corresponding to each possible threshold, i.e. the number of sub-alerts considered an overall alert. For each model, the bolded row is the one that produces the highest AUROC. Bottom panel shows the components. The components of the composite indicators are selected according to Table F3 in Detken et al. (2014), which presents their four-variable model with the highest AUROC (Model 1 here). The relative usefulness for any θ can be calculated based on the reported FP and FN rates using Eq. (9), and is reported here using $\theta = 0.5$ for brevity.

In the upper (lower) panel, the significance level α of the BSADF test is optimized based on 1980–2012 (training with 1980–1999) data. Other BSADF parameters are minimum window length = 24 quarters and lag length = 2. For the logit benchmark, the variables are transformed using 2-year differences or growth rates (except debt service ratios, which are in levels), and the signaling threshold is updated each period using a training sample. RU $\leq$ 1 is the relative usefulness given the policymaker's preference parameter $\theta = 0.5$ (see Section 3.2 for definition), absolute usefulness = Min(θ,1 − θ)RU, FP is False Positive Rate, FN is False Negative Rate. 0 $\leq$ AUROC $\leq$ 1 is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables, except for equity prices. Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period runs from 12 to 5 quarters before each crisis.

Table 6
Comparison of performance statistics with the early warning models in Alessi et al. (2015).

	AUROC	FN (%)	FP (%)	Model type
Baltussen et al.	0.875	12.0	31.0	Multivariate probit model with interdependency
Bush et al.	0.730	38.0	36.0	Multivariate logit model
Antunes et al.	0.912	40.0	4.7	Dynamic panel probit model
Neudorfer, Sigmund	0.989	8.9	2.3	Bayesian random coefficient logit model
Kauko	0.870	79.3	1.4	Simple decision rule
Behn et al.	0.920	5.6	24.7	Multivariate logit model
Babecký et al.	0.892	5.6	34.8	Bayesian model averaging
Joy et al.	0.952	3.2	12.8	Decision tree
Alessi, Detken	0.925	38.0	10.0	Random forest/Decision tree
Bank credit-to-GDP (univariate, IS)	0.830	16.0	29.0	Exuberance indicator
Composite model 1 (IS)	0.847	17.0	24.0	Composite exuberance indicator
Composite model 2 (IS)	0.850	31.0	13.0	Composite exuberance indicator
Composite model 3 (IS)	0.855	35.0	12.0	Composite exuberance indicator
Bank credit-to-GDP (univariate, OOS)	0.798	7.0	46.0	Exuberance indicator
Composite model 1 (OOS)	0.823	14.0	26.0	Composite exuberance indicator
Composite model 2 (OOS)	0.804	29.0	18.0	Composite exuberance indicator
Composite model 3 (OOS)	0.840	17.0	26.0	Composite exuberance indicator

This table reports the in-sample performance statistics for nine alternative early warning models (adapted from Alessi et al., 2015) and corresponding in-sample and out-of-sample statistics for the exuberance indicators.

In Alessi et al. (2015) the training and evaluation period is from 1970 or 1980 to about 2010, the data cover 17–28 EU countries, the crisis definitions are according to the ECB HoR crisis dataset (Babecký et al., 2012), the pre-crisis period runs from 12 to 4 quarters before each crisis, and the policymaker's preference parameter is not disclosed. For in-sample (IS) exuberance indicators, the training and evaluation period is from 1980 to 2012. For the out-of-sample (OOS) exuberance indicators, the training period is from 1980 to 1999 and the evaluation period is from 2003 to 2012. In both cases, the data cover 15 EU countries, the crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014), the pre-crisis period runs from 12 to 5 quarters before each crisis, and the policymaker's preference parameter $\theta = 0.5$. For all models, publication lags are used for data based on national accounts (not for market prices). FP is False Positive Rate, FN is False Negative Rate. 0 $\leq$ AUROC $\leq$ 1 is calculated as defined in Section 3.2.

1970/80 to about 2010 for each study. Therefore, even if the definition of loss function differs and there are differences in the crisis variable, the reported AUROC measures and false rates should be comparable on a crude level.

In Alessi et al. (2015), the in-sample AUROC values fall in the interval 0.73–0.989 and are 0.90 on average (Table 6). The in-sample AUROC values for composite exuberance indicators are 0.05 smaller on average, corresponding to one standard error. Even so, it is helpful to point out some likely causes of this in-sample performance gap for AUROC. Another factor noted by Alessi et al. (2015), which has less to do with our method, is that some data-driven methods attain very high AUROC values due to overfitting (meth-

Table 7
Timing of signals from exuberance indicators.

Variable	Alerting lead	Optimal pre-crisis window start
Bank credit-to-GDP	9	4
Real bank credit	8	8
Total credit-to-GDP	4	4
Real total credit	5	7
Household credit-to-GDP	8	12
Real household credit	8	13
Debt service ratio	5	2
Household debt service ratio	8	2
Corporate debt service ratio	2	1
Residential RE price-to-income	10	10
Residential RE price-to-rent	9	11
Real residential RE price	9	9
House price-to-income	10	8
Real house price	9	8
Real stock price index	6	4

This table present two statistics that describe the timing of signals from the exuberance indicators. The alerting lead tells how many quarters before a crisis the indicator is most useful and is based on maximizing the AUROC at different prediction horizons (see Eq. (13)). The optimal pre-crisis window location defines the location for pre-crisis window that maximizes the relative usefulness with $\theta = 0.5$. For example, "optimal pre-crisis window start" of 5 means that the highest usefulness is attained when the pre-crisis window is 5–12 quarters before the crisis, which is the pre-crisis window used here.
Data cover 1980–2012. Alerting leads are based on picking the lag that maximizes AUROC and hence independent of the significance level parameter α. For determining the optimal pre-crisis window location, α is set to 0.05. Other BSADF parameters are minimum window length = 24 quarters and lag length = 2. A one-quarter publication lag is used for quarterly variables, except for equity prices. Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014).

ods may use country-specific estimated coefficients, fixed effects, or a complex fine-tuned set-up).[17] Another side of the coin is that we could use more parameters. If we allow for adjusting the indicator weights in the composite exuberance indicators of Table 5, our AUROC values increase to 0.89. Adding country fixed effects increases the AUROC to 0.91.[18] Thus, a likely cost of adding parameters is poorer out-of-sample performance or simply higher data requirements in the best case. As noted earlier, our method with fewer parameters that needs not much more than the bare bubble theory seems to be particularly robust in out-of-sample evaluation.

In summary, the results show that it is useful to consider signals from multiple indicators at the same time. If a policymaker considers only a few indicators, more weight should go to those indicators that have higher relative usefulness when evaluated alone (e.g. credit-to-GDP ratios and debt-servicing costs). However, when the number of considered indicators increases, the relatively weaker and noisy indicators such as the alerts derived from house price ratios add positively to the aggregate useful information.

5.3. *Signal timing*

To obtain more information about the timing of the early warning signals, we study the typical alerting lead of the exuberance indicators. To this end we define a dummy variable $D_{i,t}$ which flags the starting period of each financial crisis. We consider a set of models explaining the start of crisis with the lagged early warning signal parameterized by the lag *L=1,2,. . ., 20 (quarters)*:

$$P(D_{i,t} = 1 | I_{i,t-L}) = F\left(\alpha + \beta I_{i,t-L}\right), \tag{13}$$

where $F(\cdot)$ is the logistic link function and $I_{i,t}$ are the warning signals. The RHS estimation sample includes the tranquil periods and those pre-crisis periods that match the lag length *L*. This allows us to calculate AUROC for different prediction horizons as in Drehmann and Juselius (2014). The typical alerting lead is taken to be the lag length *L* of the model with the highest AUROC. A larger alerting lead means a longer time on average between indicator alerts and the onset of the financial crisis.

In addition, we experiment with how the location of the pre-crisis window affects the usefulness of different variables given policymaker's preference $\theta = 0.5$. While keeping the length of the pre-crisis window fixed in two years, we let the endpoints of the pre-crisis windows change from 1 quarter to 24 quarters before the crisis and determine the pre-crisis window location associated with the maximum value of the relative usefulness.

Table 7 shows the alerting leads and pre-crisis window endpoint locations that produce the highest usefulness value. Both measures tell a consistent story. Debt-servicing costs (and equity prices) alert relatively close to an upcoming crisis. Earliest warnings are obtained with the house price-based indicators; explosive growth of house prices on average begins two-and-a-half years before a crisis. The credit-based indicators fall somewhere within a one- to two-year lead. Hence, it appears that the credit-based indicators may benefit from having near-optimal alerting leads when calculating the RU measure. The usefulness of equity prices and debt-servicing ratios increases if prediction horizons are shorter, while the house price-based indicators benefit from longer prediction horizons.

Drehmann et al. (2011) address the issue by recommending flexibility in forecast horizons. Thus, to allow ourselves greater flexibility in the timing and pattern of signals, we construct an additional measure: the *success rate* of each variable in predicting a forthcoming crisis within a window of five years before the crisis starts. Here, the criterion for "predicted crisis" is that a warning is signaled for at least N = 6, 8, or 10 consecutive quarters within the five-year pre-crisis window, and allowing a break of one quarter at most. This measure does not consider the indicator that signals constantly and predicts every crisis, so we also need to calculate some number of false predictions. As above, a pattern of N consecutive alarms is treated as a false alert. Naturally, we include only the crises for which the warning signal can be calculated inside the pre-crisis window.

The number of predicted crises and false alarms is shown in Table 8. While the five-year pre-crisis window is laxer than the three-to-one-year window used in calculating the usefulness measure, it provides additional insight into the usefulness of different variables. The credit-to-GDP ratios seem to produce highest success rates of crisis prediction (true positives), while the number of false alerts (false negatives) remains controlled. Depending on the data used, roughly half to two-thirds of crises in our data are preceded by a bubble in the residential real estate market. House price data from the Dallas Fed's International House Price Database seems to have the highest success rate in predicting a crisis. The real estate price-to-income variables perform best in terms of the ratio of true to false alerts. In the view of this evaluation, equity prices appear to be rather poor predictors of financial crises.

Recalling Paul Samuelson's famous observation that "the stock market has forecast nine of the last five recessions," we concede the same feature is present in early warning models for banking crises. Here, the model performance is similar to earlier studies, yet the

[17] For example, one of the methods attains AUROC = 0.989. Even without reciting the methodology behind that model, it is clear that such a value cannot be reached without overfitting. Even the financial crisis datasets cannot pinpoint the crisis dates that accurately. For example, if we directly use the Laeven and Valencia (2012) crisis dates to in-sample predict the Detken et al. (2014) crisis dates, such a model has an AUROC value of only about 0.73. In theory, the best model would fitting both crisis datasets equally would have an AUROC of about (1 + 0.73)/2 = 0.87. As a rule, financial crisis forecasting only obtains very high AUROC values by rare coincidence or overfitting.

[18] When we optimize the weights for exuberance composite indicator Model 1, the in-sample AUROC is 0.89. When we add country fixed effects it increases to 0.91. For Model 2, the numbers are 0.85 and 0.88, respectively. For Model 3, the numbers are 0.87 and 0.91, respectively.

Table 8
Statistics of correctly alerted crises and false alarms.

		Criteria = 6		Criteria = 8		Criteria = 10	
Variable	Total crises	TP	FP	TP	FP	TP	FP
Bank credit-to-GDP	17	17	18	17	15	17	10
Real bank credit	17	17	23	17	21	17	20
Total credit-to-GDP	17	16	20	15	17	14	13
Real total credit	17	17	31	17	25	17	24
Household credit-to-GDP	14	13	10	12	8	12	7
Real household credit	14	14	19	14	18	14	15
Debt service ratio	16	8	6	8	2	6	1
Household debt service ratio	11	9	4	8	3	7	3
Corporate debt service ratio	10	5	5	4	4	2	3
Residential RE price-to-income	14	7	8	7	5	6	3
Residential RE price-to-rent	14	10	16	8	8	7	4
Real residential RE price	15	9	14	9	10	9	9
House price-to-income	14	10	8	10	8	7	4
Real house price	14	11	15	10	11	9	10
Real stock price index	16	2	13	1	3	1	1

This table presents the alternative statistics to assess the quality of crisis prediction. Total crises column shows how many crises in total in our data fall into the timeframe where data from the corresponding variable is available to calculate the exuberance indicators. TP columns are the number of crises where the exuberance indicator has alerted for at least 6, 8, or 10 (depending on criteria) consecutive quarters before the crisis and a crisis has occurred within 5 years from the beginning of the alert. FP columns denote the number of occasions where the exuberance indicator has alerted for at least 6, 8, or 10 quarters (depending on criteria), but no crisis has occurred within 5 years from the beginning of the alert.

Data cover 1980–2012. BSADF parameters are α=0.05, minimum window length = 24 quarters and lag length = 2. A one-quarter publication lag is used for quarterly variables (except for equity prices). Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014).

Table 9
Usefulness values for optimized significance level and varying policymaker preference.

Whole sample (1980–2012)														Signaling method			
		θ = 0.5				θ = 0.6				θ = 0.7				θ = 0.5	θ = 0.6	θ = 0.7	
Variable	AUROC	α	RU	FP	FN	α	RU	FP	FN	α	RU	FP	FN	RU	RU	RU	AUROC
Bank credit-to-GDP	0.830	0.05	0.55	0.29	0.16	0.05	0.47	0.29	0.16	0.05	0.34	0.29	0.16	0.52	0.44	0.35	0.833
Real bank credit	0.817	0.01	0.50	0.36	0.15	0.01	0.43	0.36	0.15	0.01	0.30	0.36	0.15	0.46	0.39	0.33	0.793
Total credit-to-GDP	0.804	0.03	0.56	0.29	0.15	0.04	0.49	0.30	0.14	0.04	0.37	0.30	0.14	0.47	0.44	0.41	0.774
Real total credit	0.813	0.01	0.43	0.49	0.08	0.01	0.40	0.49	0.08	0.01	0.33	0.49	0.08	0.38	0.32	0.31	0.754
Household credit-to-GDP	0.749	0.05	0.44	0.42	0.15	0.06	0.36	0.43	0.14	0.06	0.25	0.43	0.14	0.59	0.53	0.42	0.851
Real household credit	0.780	0.01	0.40	0.47	0.13	0.01	0.33	0.47	0.13	0.05	0.27	0.60	0.05	0.39	0.33	0.26	0.749
Debt service ratio	0.762	0.08	0.39	0.11	0.50	0.08	0.14	0.11	0.50	0.20	−0.28	0.17	0.48	0.36	0.17	0.10	0.704
Household debt service ratio	0.812	0.15	0.52	0.23	0.25	0.18	0.40	0.25	0.23	0.18	0.21	0.25	0.23	0.41	0.27	0.13	0.765
Corporate debt service ratio	0.636	0.20	0.23	0.14	0.63	0.20	−0.08	0.14	0.63	0.20	−0.60	0.14	0.63	0.35	0.29	0.22	0.631
Residential RE price-to-income	0.693	0.11	0.22	0.21	0.57	0.20	−0.10	0.26	0.56	0.20	−0.56	0.26	0.56	0.32	0.21	0.19	0.724
Residential RE price-to-rent	0.641	0.20	0.22	0.32	0.46	0.20	−0.01	0.32	0.46	0.20	−0.39	0.32	0.46	0.30	0.17	0.11	0.673
Real residential RE price	0.635	0.10	0.23	0.31	0.46	0.17	0.02	0.36	0.41	0.20	−0.33	0.37	0.41	0.38	0.17	0.15	0.715
House price-to-income	0.766	0.07	0.42	0.21	0.37	0.07	0.24	0.21	0.37	0.20	−0.04	0.28	0.33	0.50	0.39	0.31	0.817
Real house price	0.726	0.03	0.42	0.29	0.29	0.03	0.28	0.29	0.29	0.05	0.04	0.32	0.28	0.46	0.31	0.24	0.787
Real stock price index	0.625	0.20	0.15	0.22	0.63	0.20	−0.16	0.22	0.63	0.20	−0.68	0.22	0.63	0.29	0.18	0.06	0.622

Short sample (2003–2012)														Signaling method			
		θ = 0.5				θ = 0.6				θ = 0.7				θ = 0.5	θ = 0.6	θ = 0.7	
Variable	AUROC	α	RU	FP	FN	α	RU	FP	FN	α	RU	FP	FN	RU	RU	RU	AUROC
Bank credit-to-GDP	0.798	0.05	0.47	0.46	0.07	0.05	0.43	0.46	0.07	0.05	0.37	0.46	0.07	0.31	0.25	0.20	0.757
Real bank credit	0.798	0.01	0.45	0.47	0.08	0.01	0.41	0.47	0.08	0.01	0.34	0.47	0.08	0.32	0.38	0.28	0.740
Total credit-to-GDP	0.702	0.04	0.30	0.57	0.13	0.04	0.24	0.57	0.13	0.04	0.14	0.57	0.13	0.24	0.20	0.20	0.664
Real total credit	0.783	0.01	0.33	0.67	0.00	0.01	0.33	0.67	0.00	0.01	0.33	0.67	0.00	0.22	0.20	0.22	0.711
Household credit-to-GDP	0.714	0.06	0.29	0.66	0.05	0.06	0.27	0.66	0.05	0.06	0.23	0.66	0.05	0.49	0.46	0.23	0.804
Real household credit	0.730	0.01	0.26	0.68	0.06	0.05	0.21	0.74	0.03	0.06	0.17	0.75	0.03	0.22	0.26	0.21	0.697
Debt service ratio	0.743	0.06	0.43	0.20	0.38	0.20	0.21	0.31	0.32	0.20	−0.06	0.31	0.32	−0.13	−0.66	−1.47	0.413
Household debt service ratio	0.816	0.07	0.49	0.28	0.23	0.07	0.37	0.28	0.23	0.20	0.42	0.38	0.09	0.23	0.16	0.10	0.734
Corporate debt service ratio	0.653	0.20	0.25	0.16	0.59	0.20	−0.05	0.16	0.59	0.20	−0.54	0.16	0.59	−0.17	−0.68	−1.53	0.405
Residential RE price-to-income	0.496	0.11	0.04	0.49	0.47	0.20	−0.19	0.52	0.44	0.20	−0.56	0.52	0.44	0.15	0.07	0.06	0.602
Residential RE price-to-rent	0.485	0.19	0.03	0.54	0.43	0.20	−0.12	0.54	0.39	0.20	−0.45	0.54	0.39	0.10	−0.19	−0.36	0.550
Real residential RE price	0.528	0.06	0.14	0.50	0.36	0.20	0.05	0.60	0.24	0.20	−0.15	0.60	0.24	0.12	−0.13	0.00	0.540
House price-to-income	0.603	0.19	0.21	0.69	0.11	0.19	0.15	0.69	0.11	0.20	0.06	0.69	0.11	0.44	0.20	0.25	0.710
Real house price	0.647	0.05	0.27	0.70	0.04	0.07	0.22	0.73	0.04	0.20	0.14	0.78	0.04	0.34	0.17	0.16	0.708
Real stock price index	0.704	0.07	0.14	0.08	0.78	0.20	−0.12	0.12	0.67	0.20	−0.68	0.12	0.67	0.16	0.13	0.12	0.598

This table presents in-sample and out-of-sample performance statistics for individual exuberance indicators. In the upper panel, the significance level α is optimized using the whole data sample. In the lower panel, the significance level has been optimized using the data sample between 1980 and 1999 and the evaluation is based on data between 2003 and 2012. Corresponding relative usefulness and AUROC values for the signaling method are provided as comparison. Three different values of the policymaker's preference parameter θ are considered.

Data cover 2003–2012. The significance level α of the BSADF test is optimized based on training with the whole sample (upper panel) and 1980–1999 data (lower panel). Other BSADF parameters are minimum window length = 24 quarters and lag length = 2. For the signaling benchmark, the variables are transformed using 2-year differences or growth rates (except debt service ratios, which are in levels), and the signaling threshold is updated each period using a training sample. RU ≤ 1 is the relative usefulness given the policymaker's preference parameter θ (see Section 3.2 for definition), absolute usefulness = Min(θ,1 − θ)RU, FP is False Positive Rate, FN is False Negative Rate. 0 ≤ AUROC ≤ 1 is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables (except for equity prices). Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period runs from 12 to 5 quarters before each crisis.

Table 10
Usefulness values for fixed significance level and varying policymaker preference.

					θ = 0.5	θ = 0.6	θ = 0.7
Variable	AUROC	α	FP	FN	RU	RU	RU
Whole sample (1980–2012)							
Bank credit-to-GDP	0.830	0.05	0.29	0.16	0.55	0.47	0.34
Real bank credit	0.817	0.05	0.47	0.11	0.42	0.37	0.27
Total credit-to-GDP	0.804	0.05	0.32	0.14	0.54	0.47	0.36
Real total credit	0.813	0.05	0.57	0.08	0.36	0.32	0.25
Household credit-to-GDP	0.749	0.05	0.42	0.15	0.44	0.36	0.24
Real household credit	0.780	0.05	0.60	0.05	0.34	0.32	0.27
Debt service ratio	0.762	0.05	0.08	0.54	0.38	0.11	−0.34
Household debt service ratio	0.812	0.05	0.15	0.35	0.50	0.32	0.02
Corporate debt service ratio	0.636	0.05	0.08	0.76	0.16	−0.22	−0.85
Residential RE price-to-income	0.693	0.05	0.17	0.63	0.19	−0.12	−0.65
Residential RE price-to-rent	0.641	0.05	0.21	0.64	0.15	−0.17	−0.70
Real residential RE price	0.635	0.05	0.25	0.53	0.21	−0.06	−0.50
House price-to-income	0.766	0.05	0.19	0.39	0.42	0.22	−0.11
Real house price	0.726	0.05	0.32	0.28	0.40	0.27	0.04
Real stock price index	0.625	0.05	0.15	0.76	0.09	−0.29	−0.92
Short sample (2003–2012)							
Bank credit-to-GDP	0.798	0.05	0.46	0.07	0.47	0.43	0.37
Real bank credit	0.798	0.05	0.62	0.01	0.36	0.36	0.35
Total credit-to-GDP	0.702	0.05	0.59	0.13	0.28	0.22	0.12
Real total credit	0.783	0.05	0.75	0.00	0.25	0.25	0.25
Household credit-to-GDP	0.714	0.05	0.65	0.05	0.31	0.28	0.24
Real household credit	0.730	0.05	0.74	0.03	0.23	0.21	0.18
Debt service ratio	0.743	0.05	0.18	0.39	0.43	0.23	−0.09
Household debt service ratio	0.816	0.05	0.25	0.25	0.50	0.38	0.17
Corporate debt service ratio	0.653	0.05	0.07	0.75	0.18	−0.20	−0.82
Residential RE price-to-income	0.496	0.05	0.46	0.51	0.03	−0.23	−0.66
Residential RE price-to-rent	0.485	0.05	0.41	0.58	0.00	−0.29	−0.77
Real residential RE price	0.528	0.05	0.48	0.38	0.15	−0.04	−0.35
House price-to-income	0.603	0.05	0.60	0.13	0.28	0.22	0.11
Real house price	0.647	0.05	0.70	0.04	0.27	0.25	0.22
Real stock price index	0.704	0.05	0.07	0.81	0.13	−0.28	−0.95

This table presents in-sample and out-of-sample performance statistics for the individual exuberance indicators. Significance level α is fixed to 0.05. Three different values of the policymaker's preference parameter θ are considered.
BSADF parameters are the value of the significance level parameter α, minimum window length = 24 quarters and lag length = 2. RU ≤ 1 is the relative usefulness given the policymaker's preference parameter θ (see Section 3.2 for definition), absolute usefulness = Min(θ,1 − θ)RU, FP is False Positive Rate, FN is False Negative Rate. Since α is fixed, note that the FP and FN are independent of θ. 0 ≤ AUROC ≤ 1 is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables, except for equity prices. Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period runs from 12 to 5 quarters before each crisis.

ratio of correct crisis prediction to false crisis prediction is about unity for the credit ratio variables. Even the best variables such as debt service ratios predict over seven of the past five crises.

5.4. Robustness checks

In the following, we report number of robustness checks against alternative parameterizations and alternative crisis definitions.

5.4.1. Alternative policy preferences θ and significance level α

The most important parameter that largely determines when the exuberance indicator issues a warning signal is the significance level of the BSADF test. The significance level should be chosen according to the policymaker's preferences θ for the trade-off with respect to the false alarms and missed crises. Previously, we assumed $\theta = 0.5$. To motivate and test the robustness of that choice, here we report results also $\theta = 0.6$ and $\theta = 0.7$. A policymaker with $\theta > \theta'$ is relatively more averse to missed crises than a policymaker with θ'. In practice, when $\theta = 0.5$ the number of actual false positive signals may already be an order of magnitude higher than the number of false negatives. In the absolute sense, the policymaker is already averse to missing a crisis with $\theta = 0.5$. Thus, a higher value of θ seems unrealistic. Still, it is customary to check the performance for higher θ.

Table 9 presents the in-sample and out-of-sample performance for the exuberance indicators for alternative values of θ such that the significance level α is optimized for each θ. Corresponding statistics for the signaling method is included for comparison. For both exuberance indicators and the signaling method, relative usefulness decreases as θ is increased. Similar to the $\theta = 0.5$ case, the similar performance of the two methods in-sample, and the relative advantage of the exuberance indicators out-of-sample, broadly carry on to higher values of θ. However, the performance of the exuberance indicators for higher values of θ is somewhat limited by our imposed cap α = 0.2.

We have so far only optimized the significance level α of the BSADF test based on the training period. However, a statistician would simply calculate the signals for α = 0.05 (and possibly for α = 0.01 and α = 0.1). Table 10 shows the in-sample and out-of-sample performance for the exuberance indicators when we make this simple choice for α. Compared to Table 9, optimization based on past data yields only negligible benefits and is sometimes counterproductive. Hence, selecting significance level α based on usual statistical practice seems to be adequate as long as θ is close to 0.5.

5.4.2. Alternative window and lag lengths

The warning signals also depend on the choice of minimum window length and lag length of the BSADF test. In Table 11, we report the sensitivity analysis along these dimensions, while keeping the significance level fixed at α = 0.05.

Our original choice of window and lag lengths, based on recommendations in the Phillips et al. (2015) paper, seem to produce results that are reasonably good compared to other parameterizations. A longer window length ($r_0 = 36$) in our sample only slightly improves the results in terms of AUROC, while longer lag length in the ADF equation makes the results slightly worse. Thus, it seems

Table 11
In-sample performance with alternative lag and windows lengths.

		Window = 12, lags = 2				Window = 24, lags = 2				Window = 36, lags = 2		
Variable	AUROC	RU	FP	FN	AUROC	RU	FP	FN	AUROC	RU	FP	FN
Bank credit-to-GDP	0.809	0.51	0.32	0.17	0.830	0.55	0.29	0.16	0.845	0.54	0.26	0.19
Real bank credit	0.807	0.38	0.52	0.10	0.817	0.42	0.47	0.11	0.824	0.44	0.45	0.11
Total credit-to-GDP	0.809	0.57	0.34	0.08	0.804	0.54	0.32	0.14	0.812	0.55	0.30	0.16
Real total credit	0.828	0.35	0.59	0.06	0.813	0.36	0.57	0.08	0.821	0.38	0.54	0.08
Household credit-to-GDP	0.672	0.28	0.49	0.24	0.749	0.44	0.42	0.15	0.815	0.51	0.37	0.12
Real household credit	0.721	0.28	0.64	0.08	0.780	0.34	0.60	0.05	0.807	0.39	0.57	0.03
Debt service ratio	0.759	0.42	0.13	0.45	0.762	0.38	0.08	0.54	0.779	0.36	0.05	0.58
Household debt service ratio	0.816	0.48	0.21	0.31	0.812	0.50	0.15	0.35	0.825	0.46	0.15	0.39
Corporate debt service ratio	0.637	0.28	0.12	0.61	0.636	0.16	0.08	0.76	0.675	0.18	0.07	0.75
Residential RE price-to-income	0.631	0.17	0.22	0.61	0.693	0.19	0.17	0.63	0.735	0.24	0.15	0.61
Residential RE price-to-rent	0.602	0.18	0.26	0.57	0.641	0.15	0.21	0.64	0.656	0.21	0.15	0.64
Real residential RE price	0.628	0.25	0.31	0.44	0.635	0.21	0.25	0.53	0.665	0.29	0.19	0.52
House price-to-income	0.738	0.40	0.24	0.37	0.766	0.42	0.19	0.39	0.765	0.43	0.17	0.41
Real house price	0.734	0.40	0.36	0.24	0.726	0.40	0.32	0.28	0.728	0.40	0.28	0.31
Real stock price index	0.697	0.21	0.19	0.60	0.625	0.09	0.15	0.76	0.619	0.08	0.14	0.79

		Window = 12, lags = 4				Window = 24, lags = 4				Window = 36, lags = 4		
Variable	AUROC	RU	FP	FN	AUROC	RU	FP	FN	AUROC	RU	FP	FN
Bank credit-to-GDP	0.747	0.45	0.39	0.16	0.815	0.48	0.32	0.20	0.834	0.49	0.28	0.23
Real bank credit	0.724	0.35	0.55	0.10	0.808	0.41	0.47	0.13	0.822	0.45	0.44	0.11
Total credit-to-GDP	0.749	0.47	0.42	0.12	0.792	0.49	0.35	0.16	0.807	0.48	0.32	0.20
Real total credit	0.696	0.31	0.62	0.07	0.789	0.35	0.57	0.08	0.808	0.38	0.54	0.08
Household credit-to-GDP	0.605	0.28	0.53	0.19	0.720	0.39	0.45	0.16	0.783	0.48	0.39	0.13
Real household credit	0.655	0.26	0.69	0.05	0.778	0.35	0.62	0.03	0.775	0.37	0.60	0.03
Debt service ratio	0.745	0.45	0.18	0.37	0.742	0.39	0.09	0.52	0.759	0.35	0.05	0.60
Household debt service ratio	0.771	0.47	0.29	0.23	0.799	0.48	0.18	0.34	0.816	0.45	0.16	0.39
Corporate debt service ratio	0.586	0.23	0.17	0.60	0.614	0.20	0.09	0.71	0.657	0.18	0.07	0.75
Residential RE price-to-income	0.553	0.08	0.28	0.64	0.623	0.13	0.19	0.68	0.666	0.17	0.16	0.68
Residential RE price-to-rent	0.569	0.17	0.31	0.52	0.643	0.13	0.24	0.63	0.649	0.18	0.17	0.65
Real residential RE price	0.570	0.16	0.38	0.46	0.620	0.17	0.29	0.54	0.675	0.28	0.22	0.50
House price-to-income	0.595	0.23	0.33	0.44	0.674	0.28	0.23	0.49	0.688	0.27	0.21	0.53
Real house price	0.599	0.24	0.40	0.36	0.636	0.28	0.31	0.42	0.639	0.29	0.26	0.45
Real stock price index	0.693	0.29	0.24	0.48	0.615	0.07	0.14	0.79	0.591	0.01	0.12	0.87

This table presents the in-sample performance statistics for individual exuberance indicators. Upper (lower) panel shows results for shorter (longer) lag length parameter. Three alternative specifications (12, 24, and 36) for minimum window length are considered.

Data cover 1980–2012. $\alpha = 0.05$. $RU \leq 1$ is the relative usefulness given the policymaker's preference parameter $\theta = 0.5$ (see Section 3.2 for definition), absolute usefulness = $\text{Min}(\theta, 1 - \theta)$RU, FP is False Positive Rate, FN is False Negative Rate. $0 \leq \text{AUROC} \leq 1$ is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables (except for equity prices). Crisis definitions are according to the ESRB crisis dataset (Detken et al., 2014). The pre-crisis period runs from 12 to 5 quarters before each crisis.

Table 12
Alternative crisis definitions.

		Laeven and Valencia (2012) crisis dating							Lo Duca et al. (2017) ECB/ESRB EU crisis dating						
					$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	Signaling				$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$	Signaling
Variable	α	AUROC	FP	FN	RU	RU	RU	AUROC	AUROC	FP	FN	RU	RU	RU	AUROC
Bank credit-to-GDP	0.05	0.782	0.30	0.27	0.43	0.29	0.07	0.723	0.728	0.33	0.30	0.37	0.21	−0.04	0.715
Real bank credit	0.05	0.770	0.47	0.18	0.35	0.26	0.11	0.704	0.707	0.50	0.22	0.28	0.16	−0.02	0.680
Total credit-to-GDP	0.05	0.776	0.34	0.26	0.40	0.27	0.06	0.705	0.718	0.34	0.34	0.32	0.15	−0.14	0.690
Real total credit	0.05	0.754	0.56	0.10	0.34	0.29	0.21	0.677	0.720	0.60	0.18	0.22	0.13	−0.03	0.657
Household credit-to-GDP	0.05	0.788	0.44	0.23	0.33	0.22	0.03	0.793	0.679	0.47	0.27	0.26	0.12	−0.10	0.735
Real household credit	0.05	0.783	0.61	0.13	0.26	0.20	0.10	0.709	0.678	0.65	0.19	0.16	0.06	−0.09	0.621
Debt service ratio	0.05	0.708	0.08	0.54	0.38	0.11	−0.34	0.707	0.687	0.09	0.61	0.30	0.00	−0.51	0.743
Household debt service ratio	0.05	0.702	0.15	0.49	0.36	0.11	−0.29	0.689	0.676	0.18	0.52	0.30	0.04	−0.39	0.705
Corporate debt service ratio	0.05	0.632	0.09	0.82	0.09	−0.32	−1.00	0.570	0.670	0.09	0.74	0.17	−0.20	−0.81	0.623
Residential RE price-to-income	0.05	0.814	0.13	0.38	0.49	0.30	−0.01	0.769	0.702	0.18	0.63	0.19	−0.12	−0.65	0.675
Residential RE price-to-rent	0.05	0.815	0.18	0.44	0.38	0.15	−0.22	0.739	0.651	0.22	0.65	0.13	−0.19	−0.73	0.638
Real residential RE price	0.05	0.821	0.22	0.39	0.39	0.20	−0.12	0.757	0.633	0.27	0.60	0.13	−0.17	−0.66	0.668
House price-to-income	0.05	0.851	0.17	0.26	0.57	0.44	0.23	0.816	0.750	0.19	0.49	0.32	0.08	−0.33	0.762
Real house price	0.05	0.846	0.28	0.15	0.56	0.49	0.36	0.784	0.698	0.33	0.36	0.31	0.13	−0.17	0.723
Real stock price index	0.05	0.552	0.16	0.86	−0.02	−0.44	−1.16	0.752	0.567	0.17	0.81	0.03	−0.37	−1.04	0.626

The in-sample performance statistics for the individual exuberance indicators and the benchmark signaling method with two alternative crisis-dating schemes are reported here. Three alternative specifications for policymaker's preference θ with respect to false alarms (FP) and missed crises (FN) are considered.

Data cover 1980–2012. BSADF parameters include the value of $\alpha = 0.05$, minimum window length = 24 quarters and lag length = 2. For the signaling benchmark, the variables are transformed using 2-year differences or growth rates (except debt service ratios, which are in levels). $RU \leq 1$ is the relative usefulness given the policymaker's preference parameter θ (see Section 3.2 for definition), absolute usefulness = $\text{Min}(\theta, 1 - \theta)$RU, FP is False Positive Rate, FN is False Negative Rate. Note that since α is fixed, the FP and FN are independent of θ. $0 \leq \text{AUROC} \leq 1$ is calculated as defined in Section 3.2. A one-quarter publication lag is used for quarterly variables (except for equity prices). The pre-crisis period runs from 12 to 5 quarters before each crisis.

reasonable to follow the original recommendations when selecting the parameters of the test.

An earlier working paper version of this study also considers rolling window ADF (RADF) tests where the right-tailed ADF test uses fixed window length. That version finds largely similar results to those as presented here (see Virtanen et al., 2017). This earlier paper also considers available higher frequency data, producing largely similar results. This further confirms our finding that the unit root based methods are robust against even quite significant changes in the windowing systems.

5.4.3. *Alternative crisis definitions*

Finally, we test the robustness or results against alternative definitions of crises using two additional crisis databases: the widely-used crisis dataset of Laeven and Valencia (2012), and a dataset produced recently by the joint efforts of the ECB and ESRB (Lo Duca et al., 2017).

When the systemic banking crisis database of Laeven and Valencia (2012) is used to define crises, the results change a bit (Table 12, top panel). In this case, real estate variables have the highest usefulness values, while credit-to-GDP variables are somewhat inferior. On average, the AUROC increases by a small amount. For this crisis definition, the exuberance indicators seem to perform significantly better than the signaling method even in the in-sample comparison. The improvement seems to be mainly caused by the fact that the Laeven and Valencia dataset includes Austria, Belgium, and Italy as countries that experienced a systemic banking crisis at 2008. Also, the timing of the crises changes slightly in many cases and some crises that are marked in the MPAG database are missing from Laeven and Valencia. The good news is that generally the same variables; i.e. house price-to-income ratios and credit-to-GDP ratios remain the best bubble predictors.

Turning to Table 12, lower panel, we see that the crisis dates of new ECB/ESRB EU crises database (Lo Duca et al., 2017) have generally lower usefulness values than with the other datasets. This seems to be a feature of the new crisis dataset. Lo Duca et al. (2017) themselves report similarly low values for their indicators. The average of the AUROC values is the same for the signaling method and exuberance indicators. Still, credit-to-GDP ratios, debt service ratios and house price-to-income ratio have the highest usefulness values.

6. Concluding remarks

This study found that an early warning indicator or set of indicators based on exuberance indicators of Phillips et al. (2015) can help in predicting financial crises with the caveat that the indicators are based on relevant time-series information and computed using an appropriate set of parameter values e.g. window length and number of lags. Although the choice of these values creates a certain amount of specification uncertainty, the approach has several unquestionable advantages compared to conventional data-driven prediction systems. Exuberance indicators are easy to compute and flexible, and may be used with different time frequencies. They do not need to rely on historical data or panel of countries to estimate the early warning. In principle, there is no upper or even lower limit for the size of the data in terms of the number of indicators. They also allow repeated tests and accumulation of information and present a step towards leveraging the full time series information of relevant variables.

For the policymaker, exuberance indicators provide policy-relevant information about the risk of a financial crisis occurring in the future. Our results bolster the case for monitoring credit-to-GDP ratios, house price-to-income (or price-to-rent) ratios and debt service ratios for explosive growth that continues for extended periods, because as such occurrences historically have often been associated with systemic banking crises and other financial crises. Exuberance indicators usually alert early enough to give policymakers time to make and implement decisions before the crisis erupts. In this context, possible relevant policies include adjustments of the countercyclical capital buffer (Jokivuolle et al., 2015).

As for extensions and future research, it seems evident that combining the information from multiple exuberance indicators and using a more flexible model would further improve the prediction performance, especially if the duration of the explosive growth is considered as well. Constructing such models by means of optimization using historical data, however, means betting on the probability that future crises will unfold in the same way as past crises – an assumption that may not hold. Fortunately, the irony of Reinhart and Rogoff's "this time is different" policymaker excuse is that historical regularities seem to be strikingly persistent.

Acknowledgements

Useful comments from two anonymous referees and the participants in the ECB's Macro-Prudential Analysis Group (MPAG) WS-B, the IIF's 2016 *International Symposium on Forecasting* (ISF) in Santander, the 2016 *RiskLab/BOF/ESRB Conference on Systemic Risk Analytics* in Helsinki, and the *International Conference of Computational and Financial Econometrics* (CFE 2017) in London, are gratefully acknowledged. Also thanks to Gregory Moore for proofreading the manuscript. Timo Virtanen acknowledges the financial support from OP Group Research Foundation.

References

Alessi, L., Detken, C., 2011. Quasi real time early warning indicators for costly asset price boom/bust cycles: a role for global liquidity. Eur. J. Polit. Econ. 27 (3), 520–533.

Alessi, L., Antunes, A., Babecky, J., Baltussen, S., Behn, M., Bonfim, D., Bush, O., Detken, C., Frost, J., Guimaraes, R., Havranek, T., Joy, M., Kauko, K., Mateju, J., Monteiro, N., Neudorfer, B., Peltonen, T., Rodrigues, P., Rusnak, M., Schudel, W., Sigmund, M., Stremmel, H., Smidkova, K., van Tilburg, R., Vasicek, B., Zigraiova, D., 2015. Comparing Different Early Warning Systems: Results from a Horse Race Competition Among Members of the Macro-prudential Research Network. MPRA Paper n. 62194. University Library of Munich, Germany.

Anundsen, A., Gerdrup, K., Hansen, F., Kragh-Sorensen, K., 2016. Bubbles and crises: the role of house prices and credit. J. Appl. Econom. 31, 1291–1311.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K., Vašíček, B., 2012. Banking, debt and currency crises: Early warning indicators for developed countries. ECB Working Paper 1485.

Ball, L., 2014. Long-term damage from the Great Recession in OECD countries. NBER Working Paper 20185.

Banerjee, A., Chevillon, G., Kratz, M., 2013. Detecting and forecasting large deviations and bubbles in a near-explosive random coefficient model. ESSEC Business School, ESSEC Working Paper 1314.

Barberis, N., Greenwood, R., Jin, L., Shleifer, A., 2018. Extrapolation and bubbles. J. Financ. Econ. (forthcoming).

Brunnermeier, M., Eisenbach, T., Sannikov, Y., 2013. Macroeconomics with financial frictions: a survey. In: Advances in Economics and Econometrics, Tenth World Congress of the Econometric Society. Cambridge University Press, New York, pp. 1–93.

Brunnermeier, M., 2008. Bubbles. In: Durlauf, S., Blume, L. (Eds.), New Palgrave Dictionary of Economics. , 2nd edition. Palgrave Macmillan, pp. 1–17.

Busetti, F., Taylor, A., 2004. Tests of stationarity against a change in persistence. J. Econom. 123, 33–66.

Bussière, M., Fratzscher, M., 2008. Low probability, high impact: policy making and extreme events. J. Policy Model. 30 (1), 111–121.

Campbell, Y., Shiller, R., 1988a. The dividend-price ratio and expectations of future dividends and discount factors. Rev. Financ. Stud. 1 (3), 195–228.

Campbell, Y., Shiller, R., 1988b. Stock prices, earnings and expected dividends. J. Finance 43 (3), 661–676.

Campbell, J.Y., Lo, A.W., McKinlay, A.C., 1997. The Econometrics of Financial Markets. Princeton University Press, NJ.

Caprio, G., Klingebiel, D., 1996. Bank Insolvencies, Cross-Country Experience. World Bank Policy Research Working Paper n. 1620.

Cochrane, J., 1992. Explaining the variance of price-dividend ratios. Rev. Financ. Stud. 5 (2), 243–280.

Corsi, F., Sornette, D., 2014. Follow the money: the monetary roots of bubbles and crashes. Int. Rev. Financ. Anal. 32, 47–59.

Craine, R., 1993. Rational bubbles—a test. J. Econ. Dyn. Control 17, 829–846.

de Haan, J., Nijskens, R., Wagn, W., 2017. Macroprudential regulation: from theory to implementation. J. Financ. Stabil. 28, 182.

Demirgüc-Kunt, A., Detragiache, E., 1998. The determinants of banking crises in developed countries. IMF Staff Papers 45 (1), 81–109.

Demirgüc-Kunt, A., Detragiache, E., 2000. Monitoring banking sector fragility: a multivariate logit approach. World Bank Econ. Rev. 14, 287–307.

Detken, K., Weeken, O., Alessi, L., Bonfim, D., Boucinha, M., Castro, C., Frontczak, S., Giordana, G., Giese, J., Jahn, N., Kakes, J., Klaus, B., Lang, J., Puzanova, N., Welz, P., 2014. Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options. ERSB Occasional Paper Series n. 5, June 2014.

Drehmann, M., Juselius, M., 2014. Evaluating early warning indicators of banking crises: satisfying policy requirements. Int. J. Forecast. 30, 759–780.

Drehmann, M., Borio, M., Tsatsaronis, K., 2011. Anchoring countercyclical capital buffers: the role of credit aggregates. Int. J. Cent. Bank. 7, 189–240.

Efthymios, P., Yusupova, A., Paya, I., Peel, D., Martinez-Garcia, E., Mack, A., 2016. Episodes of exuberance in housing markets: In search of the smoking gun. J. Real Estate Finance Econ. 53, 419–449.

Elliot, G., Rothenberg, T.J., Stock, J.H., 1996. Efficient tests for an autoregressive unit root. Econometrica 64, 813–836.

Elliot, G., 1999. Efficient tests for a unit root when the initial observation is drawn from its unconditional distribution. Int. Econ. Rev. 40, 767–783.

Escobari, D., Jafarinejad, M., 2016. Date stamping bubbles in real estate investment trusts. Q. Rev. Econ. Finance 60, 224–230.

Frankel, J., Saravelos, G., 2010. Are Leading Indicators of Financial Crises Useful for Assessing Country Vulnerability? Evidence from the 2008–09 Global Crisis. NBER Working Paper 16047, June.

Franses, P., 2016. A simple test for a bubble based on growth and acceleration. Comput. Stat. Data Anal. 100 (August), 160–169.

Gordon, H., 1962. The Investment, Financing and Valuation of the Corporation. Irwin Homewood, IL.

Gurkaynak, R., 2008. Econometric tests of asset price bubbles: taking stock. J. Econ. Surv. 22, 166–186.

Homm, U., Breitung, J., 2012. Testing for speculative bubbles in stock markets: a comparison of alternative methods. J. Financ. Econom. 10 (1), 198–231.

Honohan, P., 2016. Debt and austerity: post-crisis lessons from Ireland. J. Financ. Stabil. 24, 149–157.

Jokivuolle, E., Pesola, J., Viren, M., 2015. Why is credit-to-GDP a good measure for setting countercyclical capital buffers? J. Financ. Stabil. 18, 118–126.

Jordá, Ò., Schularick, M., Taylor, A., 2015. Leveraged bubbles. J. Monet. Econ. 76 (Supplement), S1–S20.

Kauko, K., 2014. How to foresee banking crises? A survey of the empirical literature. Econ. Syst. 38 (3), 289–308.

Kim, T.-H., Leybourne, S., Newbold, P., 2002. Unit root tests with a break in innovation variance. J. Econom. 109, 365–387.

Koustas, Z., Serletis, A., 2005. Rational bubbles or persistent deviations from market fundamentals? J. Bank. Finance 29, 2523–2539.

Laeven, L., Valencia, F., 2012. Systemic Banking Crises Database: An Update. IMF Working Paper n. 163/12.

Leybourne, S., Kim, T., Taylor, A., 2006. Regression-based test for a change in persistence. Oxf. Bull. Econ. Stat. 68 (5), 595–621.

Leybourne, S., 1995. Testing for unit roots using forward and reverse Dickey-Fuller regression. Oxf. Bull. Econ. Stat. 57, 559–571.

Lo Duca, M., Koban, A., Basten, M., Bengtsson, E., Klaus, B., Kusmierczyk, P., Lang, J., Detken, C., Peltonen, T., 2017. A new database for financial crises in European countries. ECB Occasional Paper n. 194.

Pavlidis, E., Yusupova, A., Paya, I., Peel, D., Martínez-García, E., Mack, A., Grossman, V., 2016. Episodes of exuberance in housing markets: In search of the smoking gun. J. Real Estate Finance Econ. 53 (4), 419–449.

Pedersen, T., Schütte, E., 2017. Testing for Explosive Bubbles in the Presence of Autocorrelated Innovations. CREATES Research Paper 2017-9.

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

Phillips, P., Shi, P., Yu, J., 2015. Testing for multiple bubbles: historical episodes of exuberance and collapse in the S&P500. Int. Econ. Rev. 56 (4), 1043–1078.

Reinhart, C., Rogoff, K., 2009a. The aftermath of financial crises. Am. Econ. Rev. 99 (2), 466–472.

Reinhart, C.M., Rogoff, K.S., 2009b. This Time is Different: Eight Centuries of Financial Folly. Princeton University Press.

Ristolainen, K., 2017. Essays on early warning indicators of banking crises. University of Turku, Series E, n. 14. (published doctoral dissertation).

Sarlin, P., 2013. On policymakers' loss functions and the evaluation of early warning systems. ECB Working Paper n. 1509.

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

Scherbina, A., 2013. Asset price bubbles: A selective survey. IMF Working Paper WP/13/45.

Schularick, M., Taylor, A., 2012. Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008. Am. Econ. Rev. 102 (2), 1029–1061.

Tölö, E., Laakkonen, H., Kalatie, S., 2018. Evaluating indicator for use in setting the countercyclical capital buffer. Int. J. Cent. Bank. 14 (2), 51–112.

Taipalus, K., Virtanen, T., 2016. Predicting asset bubbles with unit root methods. Bank of Finland, unpublished mimeo.

Taipalus, K., 2006. Bubbles in the Finnish and US equities markets. Bank of Finland Studies, Series E, n. 3.

Virtanen, T., Tölö, E., Virén, M., Taipalus K., 2017. Use of unit root methods in early warning of financial crises, ESRB Working paper n. 45.

Wan, J., 2015. Household savings and housing prices in China. World Econ. 38, 172–192.

West, K., 1987. A specification test for speculative bubbles. Q. J. Econ. 102, 553–580.

Wilcox, D., 1989. The sustainability of government deficits: implications for present value borrowing constraints. J. Money Credit Bank. 54, 1837–1847.