Use of Unit Root Methods in Early Warning of Financial Crises

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

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

Current version 5.10.2016

Abstract

Unit root methods have long been used in detection of financial bubbles in asset prices. The basic idea is that fundamental changes in the autocorrelation structure of relevant time series imply the presence of a rational price bubble. We provide cross-country evidence for performance of unit-root-based early warning systems in ex-ante prediction of financial crises in 15 EU countries over the past three decades. We then combine the identified early warning signals from multiple time series into a composite indicator. We also show that a mix of data with different frequencies may be useful in providing timely warning signals. Our results suggest and an early warning tool based on unit root methods provides be a valuable accessory in financial stability supervision.

Key words: Financial crises; unit root; combination of forecasts

JEL codes: G01, G14, G21

*Correspondence to: Eero Tölö, Financial Stability and Statistics Department, Bank of Finland, PO Box 160, FI-00101 Helsinki, Finland. E-mail:eero.tolo@bof.fi.

Disclaimer and acknowledgements: The views presented are those of the authors and do not necessarily represent the views of the Bank of Finland. We gratefully acknowledge the comments of participants in the ECB's Macro-prudential analysis group (MPAG) WS-B and at the IIF's 2016 International Symposium on Forecasting (ISF) in Santander. Our gratitude also goes to Gregory Moore for proof-reading the manuscript.

1. Introduction

Given the costs sustained by many countries in the wake of the Great Recession, the current interest in developing early warning mechanisms for financial crises is hardly surprising. Lawrence Ball (2014), for example, estimates that the cost to 23 OECD economies in the aftermath since 2008 has averaged about 8.4% of GDP. In the case of Greece, that burden has exceeded 30%. Even if fiscal and structural problems and exchange market imbalances account for some of this cost and only part of these costs are attributable to financial or banking crises, the implications are substantial (Reinhart and Rogoff, 2009).

A good starting point for financial crisis study is identifying the elements of a prototype crisis. As a rule, we can say crises emerge after hefty increases in asset prices, indebtedness, and output. Further, time-series behavior in retrospect often looks like a bubble. Thus, the basic issue for policymakers becomes one of predicting asset bubbles.

Our analysis builds on the main findings of several significant asset-bubble papers: Campbell, Lo and McKinlay (1997), Campbell and Shiller (1988), Craine (1993), and Koustas and Serletis (2005).

The Campbell and Shiller (1988) paper on rational bubbles notes that dividend-price ratio behavior can reveal an existing rational asset bubble. A constantly demising dividend-price ratio may be a sign of worsening overpricing, while rising prices should at some point be realized as higher dividends. If they are not, the price rise is not based on fundamentals. This becomes apparent from the following pricing formula:[1]

$$d_t - p_t = -\frac{\alpha}{1-\rho} + \mathbb{E}_t\left[\sum_{j=0}^{\infty} \rho^j \left(-\Delta \mathrm{d}_{t+j+1} + r_{t+j+1}\right)\right] \tag{1}$$

where d denotes dividends, p price of asset, r the asset returns (all in logs). α and ρ are parameters from the linear approximation of the asset return formula. ρ can be interpreted as the discount rate. According to the formula, stationarity of Δd and r implies that the log dividend yield must also be stationary (for details, see Cochrane, 1992; and Craine, 1993).

In other words, the presence of a unit root in the log dividend-to-price ratio means that agents or their expectations are not rational (assuming no other fundamental market failures). A possible interpretation here is that there is a rational bubble in the asset price. Indeed, this view has spawned several studies where the stationarity properties of stock prices are examined using unit root testing procedures. Although some studies (e.g. Corsi and Sornette, 2014) attempt a general model of bubbles, the empirical work usually relies on unit root testing.[2]

For our purposes, the asset pricing equation defines a "market fundamental" price for an asset, i.e. the price justified by future dividends and possible other returns. Any deviations from

[1] The formula is a generalization of the Gordon (1962) growth model for the case where dividend growth and rates of return change over time.
[2] Banerjee *et al.* (2013) and Franses (2013) offer novel techniques in bubble detection. Banerjee *et al.* (2013) use a random coefficient autoregressive model, while Franses (2013) tests the feedback between first and second differences of time series, causing the time series to explode.

this fundamental price could possibly constitute an asset price bubble. Denoting this fundamental price by p_t^f, we can formulate the asset price at time t as

$$p_t = p_t^f + b_t \tag{2}$$

where b_t is the price bubble component. Assuming that all agents (buyers and sellers of the asset) are rational and that they have the same information about the fundamental price, the bubble component must either be zero or follow a submartingale process (Phillips *et al.*, 2013):

$$\mathbb{E}_t(b_{t+1}) = (1 + r_f)b_t \tag{3}$$

where r_f is the risk-free interest rate used to discount future earnings. This means that when a bubble is present, the asset price process changes from I(1) (or even stationary) to an explosive process.[3] Methods used in this paper are designed to detect this change in the time series dynamics.

Although the Campbell-Schiller formula (1) deals with stock prices, it is relevant for other asset prices as well. For example, in examining house prices, we can substitute rents for dividends. This approach can even be extended to the debt-to-GDP ratio, where income growth in a macro-setting plays a similar role to aggregate dividend growth. This is obviously true in a world where the functional distribution of income is constant. Moreover, we know from conventional government debt accounting models that stationary growth rate of income (tax returns) is incompatible with a continuously increasing (nonstationary) debt-to-GDP ratio (e.g. Wilcox, 1989).

The literature on detecting bubbles with unit root methods is well-known in the forecasting community. Empirical tests that deal with the presumed bubble properties of financial time series include Elliot (1996) and Elliot *et al.* (1999), 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.* (2004), 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 and 2013) 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, and demonstrate reasonably good *ex post* prediction performance. A similar approach relying on standard unit root tests with a rolling window is found in Taipalus (2006a) and Taipalus and Virtanen (2016).

This study investigates the *ex ante* predictive performance of two asset bubble tests in signaling the risk of financial crisis. The two tests are a backward SADF test (PSY test) and the rolling window ADF test used in Taipalus (2006a). In practical sense, the main difference between the two methods is the application of time windows in the testing procedure. While

[3] The term "bubble" is rather vague characterization of historical phenomena that have preceded financial or economic crises. It is thus hardly surprising that there is no consensus on the mathematical representation of a bubble. For our purposes, Eq. (3) is probably a good approximation.

Taipalus (2006a) uses a rolling window of fixed length, Phillips *et al.* (2013) uses the supremum of tests with all window sizes larger than a specified minimum window length. The procedure of Phillips *et al.* (2013) no doubt appeals conceptually,[4] but the fixed rolling window recognizes the practical objective of simply spotting the fundamental change in the data-generating process of the underlying time series, i.e. the point at which the "normal" regime in the time series process shifts to a different regime. In this respect, both Monte Carlo evidence and tests with actual time series suggest that standard unit root tests perform *ex post* about as well as the PSY test (see Taipalus, 2006a, and Taipalus and Virtanen, 2016). The rolling window ADF test procedure is transparent and signal interpretation is straightforward – especially in a setting were the test is carried out routinely from period to period using the same parametrization. Moreover, the test avoids the complexities of PSY limit theory, allowing instead the use of standard critical values for the ADF test.

Our testing sample consists of 15 EU countries over the period 1980–2014. The data include a broad set of financial and macroeconomic variables from measures such as credit-to-GDP ratio and debt-servicing costs to asset prices (see the next section for details). The performance evaluation is based on standard measures taken from the early warning literature that rely on the number of type I and type II errors produced by the signals.

The results show that the unit root based methods are successful in predicting financial crises in both in-sample and out-of-sample evaluation. The two methods yield quite similar results, on average beating our benchmark conventional signaling method. The credit-to-GDP ratios and debt-servicing costs are among the best-performing indicators according to the relative usefulness metric, but house price and stock price based indicators are also shown to be useful. Furthemore, policymakers can benefit from combining single-value indicators into composite indices. Using two variables, the obvious choices are the credit-to-GDP ratio and debt-servicing costs. When the number of components is increased, however, including predictors such as house and stock prices adds greatly to the usefulness of a composite index.

Given these findings, it may be worthwhile to include unit root methods as a tool in identifying the emerging risk of a financial crisis.

The rest of this paper is organized as follows. The particulars of the data are explained in section 2.1. Section 2.2 reviews the unit-root based early warning methods, sections 2.3 and 2.4 introduce the performance evaluation setup. Section 2.5 explains how our composite indices are formed. Sections 3.1 and 3.2 present the empirical results for single variables and composite indices, respectively. Section 4 concludes.

2. Empirical analysis

2.1 Data

[4] Phillips *et al.* (2011), for example, argue that standard unit root tests are inappropriate tools for detecting bubble behavior, because they fail to distinguish effectively between a stationary process and a periodically collapsing bubble model. In their view, the latter look more like data generated from a unit root, or even a stationary autoregression process, than a potentially explosive process.

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, Great Britain, 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, partly due to the fact that the functioning of financial markets was markedly different from the present system. To limit possible bias due to structural changes in banking, we only start our sample from 1980. In addition, robustness checks are reported for a short-sample starting from 2003 to capture only the most recent financial crises.

The variables include credit aggregates, debt-servicing costs, residential and commercial real estate prices, stock prices, and other macroeconomic variables. The dataset is based on a quarterly data series compiled by the ECB and shared within the macro-prudential analysis group (MPAG). Most of the time series are based on publicly available data from BIS, ECB and OECD. The exceptions are the ECB's debt-servicing costs and commercial real estate price data, which are not publicly available. We have also amended the dataset with higher frequency monthly observations of stock market from Bloomberg and house prices for the BIS.

As we can see from Table 1, we have a rich set of variables from which to choose. Acknowledging previous research (e.g. Peltonen *et al.*, 2014; Eidenberger *et al.*, 2014; Ferrari and Pirovano, 2015: and Schularick and Taylor, 2012) showing that some variables related to labor markets, government finance, and output growth are poor predictors of crises, we nevertheless include the entire dataset into our first round of analysis on the off chance we might be missing something. Only in subsequent analysis do we concentrate on the most relevant variables.

Evaluating the quality of the warning signals implies good data on financial crises. Thus, choosing the right period is critical in such analysis; wrong choices automatically invalidate the results.[5] To narrow our search, we use the crisis classification scheme of the ECB and ESRB (Detken *et al.*, 2015), which is based extensively on expert opinion from individual central banks. While differences remain, these datasets broadly agrees with the crises episodes used in the datasets of Laeven and Valencia (2012) and Reinhart and Rogoff (2010).

As an alternative to the defined crisis period, we use loan losses of banks (in relation to total lending) to illustrate stressed periods. The loan loss data was collected by Jokivuolle *et al.* (2015). These data are quite different from crisis indicator data; they exist in ratio form rather than as binary indicators.

2.2 Unit root tests for rational bubbles

This section provides background to the rolling window ADF test used in Taipalus (2006a) and the sup ADF (SADF) test of Phillips *et al.* (2013).

Taipalus (2006a, 2006b) estimates the familiar ADF test equation

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

[5] Recent evidence is provided in Ristolainen (2015).

for all possible subsamples of length W. At each point t, the window includes observations $[t - W, t]$. The parameter γ is tested for H0: mild unit root ($\gamma = 0$) against the alternative hypothesis of explosive behavior ($\gamma > 0$). A warning signal is issued at the point t if the null hypothesis is rejected. Because of this hypothesis specification, the right tail of the Dickey-Fuller distribution is used (i.e. null is rejected if the test statistic exceeds the 95% critical value). Note that this differs from the usual use of the ADF test, where H0 is I(1) and the alternative hypothesis is I(0) (i.e. the left tail of the Dick-Fuller distribution is used). A possible simplification here is to just compare the estimate of γ to a threshold value to extract the warning signals as in Taipalus and Virtanen (2016).[6]

There are three parameters to be selected: the length of the rolling window (W), the lag length of the ADF equation (p), and the significance level of the test (α). To make the tests comparable, we use fixed parameters throughout the data. We can see already beforehand that changing the parameters improves the performance of certain variables (e.g. real estate variables would perform better with longer window lengths in terms of the usability of the indicator, and each variable would perform somewhat better if α is optimized). We use a fixed window length of three years, a lag length of one, and fix the value of α to 0.05. With quarterly data, this amounts to only twelve observations in the window, which is considered too short for a normal (left-tailed) ADF test. However, based on our experience with real data and extensive Monte Carlo simulations, it seems that even short windows produce reasonably good results in right-tailed hypothesis testing.

Phillips *et al.* (2013) use the same empirical regression model (2). However, their backward sup ADF (BSADF) test estimates the ADF statistic repeatedly on a backward expanding sample sequence. The critical value of test, which differs from the normal ADF critical value, is compared to the sup value of this sequence.[7] The BSADF test also has three free parameters: the lag length p in Equation (4), the significance level α for the SADF test, and a minimum window length parameter w_0. To facilitate comparison of the two methods, we use the aforementioned parameter values p=1 and α=0.05 and set the minimum window size w_0 equal to the fixed window size w (3 years) used in the rolling ADF tests. Equating window sizes and lag lengths ensures that the estimation samples for the two methods are the same.

Another important point concerning both tests is that short-run explosive contractions in time series ("negative bubbles") are also identified as warning signals in the test statistics. In practice, this means that the methods issue warning signals from e.g. stock market crashes or contractions in house prices. In the context of early warning methods, however, such signals are not considered relevant as they typically show up after the crisis is underway. Thus, in our evaluation, we remove all warning signals that occur when the corresponding time series is decreasing in value.

---

[6] Extensive Monte Carlo testing gives (at least partial) justification for procedure. We could use even a simpler AR(1) equation and consider the where the AR parameter $\rho_1 \geq 1$. This is because the "t-statistic" is defined as the ratio of the estimated coefficient and its standard error, and the test statistic for zero corresponds to a p-value that is close to 0.958. Thus, when the estimated coefficient is zero or positive, the p-value will always be above 0.95. This also partly explains why the right-tailed test works so well with small window sizes. Since we are looking for test statistic values close to zero, variance does not matter much as the estimated coefficient will also be close to zero.

[7] See Phillips *et al.* (2013) for details on how critical values are calculated.

2.3 Performance evaluation

Our performance evaluation is based on the number of type I and type II errors. It consists of the policymaker's loss function and a "usefulness" measure. The weights of the loss function reflect the presumed preferences for the errors. This methodology draws upon the policy loss functions of Demirgüc-Kunt and Detragiache (2000) and Bussière and Fratzscher (2008), and the usefulness measure proposed by Alessi and Detken (2011) and later supplemented by Sarlin (2013).

The loss function of Alessi and Detken (2011) is defined as follows:

$$L_{AD}(\theta) = \theta T_2 + (1-\theta)T_1 = \theta \frac{C}{A+C} + (1-\theta)\frac{B}{B+D} \qquad (5)$$

where the right-hand side is a weighted average of the type I and type II error rates, T_1 and T_2, respectively.[8] *A* is the number of periods in which an indicator provides a correct signal (crisis starts within 1 to 3 years of issuing the signal), and *B* the number of periods in which a wrong signal is issued. To account for publication lags, all quarterly data are lagged by one quarter. *C* is the number of periods in which a signal is not generated during a defined period from the onset of the crisis (1 to 3 years). Finally, *D* denotes the number of periods in which a signal is correctly not provided. In other words, *A*=*TP*, number of true positives; *B*=*FP*, number of false positives; *C*=*FN*, number of false negatives; and *D*=*TN*, number of true negatives. θ is the parameter revealing the policymaker's relative risk aversion to 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.

Sarlin (2013) augments the policymaker loss function with the unconditional crisis probability such that

$$L_S(\mu) = \mu P T_2 + (1-\mu)(1-P)T_1 \qquad (6)$$

where $P = \frac{A+C}{A+B+C+D}$ is the unconditional crisis probability as estimated from the sample.

For either loss function, the relative usefulness statistic is defined as

$$U_r = \frac{\omega - L}{\omega} \qquad (7)$$

where for Alessi and Detken (2011) $\omega = \min(\theta, 1-\theta)$ and for Sarlin (2013) $\omega = \min(\mu P, 1-\mu P)$. The normalization parameter ω ensures that the maximum value of the relative usefulness is 1.

In our data, $P_1 = 0.1$ and we set $\mu = 0.9$. Incidentally, with these settings the two alternative measures of usefulness are the same when $\theta = 0.5$.

---

[8] In the formula, the order of T_1 and T_2 differs from some of the earlier literature. This is merely 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 thus set H_0, i.e. "no crisis within the next 3 years" so that false positive means a false alarm. A type II error (false negative) incorrectly retains a false null hypothesis. Thus, our case false negative means failure to detect a crisis.

In addition to reporting usefulness, we compare the performance of the rolling window ADF indicator with the traditional signaling indicator derived from the trend gap, where the trend is calculated with one-sided Hodrick-Prescott filter with a smoothing parameter of 400,000.[9] We obtain the ROC curve[10] of the relative trend gap using the 3-to-1-year pre-crisis window as the dependent variable. To illustrate dependence on the window length w and significance level α, we plot together the false positive and false negative rates of the unit root indicator with different window lengths varying from 12 to 48 observations and confidence levels varying from 0.8 to 0.99.

2.4 Lead-lag structure of signals and alternative performance measures

Different indicators are likely to signal risk to the financial system at different times. Drehmann *et al.* (2011) addresses this issue by recommending flexibility in forecast horizons. Early warning signals reveal the build-up of imbalances, but do not tell us when problems begin. Thus, if an alarmingly high value is observed at least once during a lengthy pre-crisis period, the crisis must be regarded as predicted.

We try to answer the question of how many time periods prior to the crisis that indicator signals are most relevant. This time period is calculated by first estimating a sequence of models that explain a dummy variable, a dummy which flags the starting period of each financial crisis, using a warning signal indicator lagged by n=1,2,…24 periods. The sought time period is the lag length of the model with highest maximum likelihood, given that the regression coefficient is positive and statistically significant at 5% significance level.

The manner in which consecutive signals are obtained could be informative. For instance, there could be two completely different patterns, i.e. warning signals arrive in a *steady* cumulative manner, or they are *sporadic*, with one warning signal rarely followed by another.

The usefulness measure with the 3-to-1-year pre-crisis window is conceptually appealing, but ignores the possibility of complex signaling patterns. To allow greater flexibility in the timing and pattern of signals, we construct an additional measure: the *success rate* of each variable to predict a forthcoming crisis within a window of five years before the crisis starts. Here, the criterion for a crisis to be classified as "predicted" is that the variable produces a warning signal for at least six consecutive quarters within the five-year pre-crisis window, allowing a break of at most one quarter within the alert. As the number of crises predicted does not take into account that an indicator signaling all the time would predict all the crisis, we construct another measure – number of crises falsely predicted. As above a pattern of six consecutive alarms would be considered a false alert. If there is no data on the variable during the pre-crisis window, the crisis is omitted from the calculation.

2.5 Composite indicators

As the unit root tests apply to single variables, it is of interest to consider whether aggregating information from the tests for multiple variables can provide additional predictive performance.

[9] This value, originally proposed by Borio (2012), is currently widely used in official contexts. See Gerdrup *et al.* (2013) and Repullo and Saurina (2011) for critical analysis and discussion.
[10] The receiver operating characteristic (ROC) curve is the graph connecting all possible pairs of type I and type II errors obtained by altering the signaling threshold.

One variable may indicate clearly some historical crises, while others perform better with other crises. To explore the issue, we compose composite indices by summing up the single variable signals for a set of variables. As the number of possible combinations of variables is daunting, we only consider a few example indices. Many alternatives are possible, of course.

The composite indices are formed by taking a weighted sum of the single variable warning signals $I_k \in \{0,1\}$:

$$C_N = w_1 I_1 + w_2 I_2 + \cdots + w_N I_N \tag{8}$$

where w_k are the weights. The composite index issues a signal when $C_N \geq C^*$ for a preset signaling threshold C^*. For simplicity, we consider only *equally weighted indices* parameterized by $w_k = 1/N$. In this case all the relevant signaling thresholds are given by $C^* = k/N$, where $k = 1,2,\dots,N$. Consequently, the composite index issues a warning signal if at least k of its components alert simultaneously. The composite indices can, of course, be used in a similar "cumulative" way as individual indices to see whether warning signals keep coming in a repeated manner.

3. Empirical results

3.1 Single variate indicators

To illustrate our approach, we offer a set of graphs for three predictor variables in Figures 1–3 (credit-to-GDP ratios, debt-service ratio, and real house prices). In line with the existing early-warning literature, we see that these predictor variable tend to be among the most relevant in signaling financial crises. The graphs show the development of the underlying variable, the signals from the unit roots, and the pre-crisis/crisis periods. The indicators all perform quite well in terms of signaling alarms well in advance of crisis onset. For example, the signals from debt-service ratio and credit-to-GDP gap (Figures 1–2) 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, which according to the crisis dataset did not experience systemic crises.[11] The graphics also indicate that the respective lead-lag relationships do not follow a uniform pattern and do not match perfectly with the ECB's pre-crisis or actual crisis measures. Yet again, we see the need to be flexible with the prediction window and to have alternative prediction targets such as loan losses (as discussed below).

Our first *usefulness* measures are for the the rolling window ADF test for the full set of variables using three different window lengths (Table 2). Usefulness values are typically best for the shortest window size (3 years). As shorter window is also conceptually more appealing from the early-warning point of view, we adopt this window size for the rest of our study.[12] The top performers in terms of relative usefulness are the various credit-to-GDP ratios, large credit aggregates, and debt-servicing ratios. As pointed out above, some variables do not perform well in terms of the usefulness criterion. This set of weak predictors, which we now drop

[11] The banking sectors in these countries suffered considerable losses during 2008–2012.
[12] The PSY method uses the respective window size as the minimum window size.

from the analysis, includes the level of GDP, level of interest rates, the unemployment rate, household disposable income, and the value of household mortgages. While stock prices and real estate price indicators only give mediocre or low performances in this comparison, we keep them for the remainder of the analysis. Referring back to Figure 3, the timing mismatch between the signal and the pre-crisis period seems to be the main reason real estate indicators attain relatively low usefulness in the ADF test.

The main results of the paper for both ADF test and PSY test are reported in Table 3. The sample are the 15 EU countries from 1980 to 2014. As expected, the indicators that perform very well are the same for both test methods: credit-to-GDP ratios, real credit stocks, and debt-servicing ratios. The correlation between the usefulness of the methods is quite high (0.7). Although, on average the usefulness measures for PSY (0.31) are 50% higher than for the basic ADF test (0.20), this is mainly an artefact of the choice of the significance level parameter α. The fixed α=0.05 for the PSY test balances false alarms and missed crises, while for the ADF tests the rate of missed crises (FNR) on average is more than double the false alarm rate (FPR). As for high(er) frequency data, despite higher number of observations for the unit root tests, we do find no significant improvement relative to corresponding quarterly data. However, having the data at our disposal earlier is an advantage in itself.

We now turn to the *number* of predicted crises and falsely predicted crises based on the metric that counts consecutive signals within five years of crisis onset (see Table 4). While this criterion is more lax than our usefulness measure, it still provides insight into the behavior of different variables. For example, only slightly more than half of the crises in our data were preceded by a clear bubble in the residential real estate market as evaluated from the price-to-rent variable. On the other hand, for over two-thirds of the crises in our data, there was a long period of explosive growth of the total credit-to-GDP ratio before the crisis. Recalling Paul Samuelson's famous observation that "the stock market has forecast nine of the last five recessions," we must concede the same problem arises in early warning models for banking crises. Here, the model performance is similar to earlier studies. The ratio of correct crisis prediction to false crisis prediction is at best 2.5 for the debt-to-disposable income ratio, i.e. our best indicator predicted seven of the past five crises. Similarly, the prediction-to-false-alarm ratios for debt-servicing ratios and real-estate price ratios rank relatively high on this metric, about 1–2 and close to 1, respectively. According to this metric, the credit-to-GDP ratio, which had the highest relative usefulness, predicted about *eleven* of the past five crises. The bottom line is that one should be careful when ranking the indicators according to relative usefulness (or any other metric) as the results can be quite sensitive.

The *lead-lag structure* of the alerts with respect to the starting period of the financial crisis also provides valuable policy insights and sheds light on the timing mismatch between the alerts and the prediction horizon stipulated by the relative usefulness measure. The last two columns in Table 5 shows the lag length calculated according to the maximum likelihood criteria as explained in section 2.4 for the ADF and PSY methods, respectively. Shorter lag length means there is a shorter time on average between indicator alerts and the onset of the banking crisis. The findings for the unit root method that the shortest lag length (one to two years) occurs with debt-servicing costs and equity prices comport the findings of Kalatie *et al.* (2015), using the same signaling method. We find that the longest lag lengths are with the house price-based indicators with lag lengths up to four and five years. The credit-based indicators fall somewhere within a two- to three-year lag. Hence, it appears that the credit-based indicators may benefit having near optimal lag length when compared to the one- to three-year windows used in the relative usefulness measure. The relative usefulness of equity prices and debt-servicing ratios increases if the prediction horizons are shorter, while the house price-based indicators benefit from a longer prediction horizon.

How do the unit root methods compare to conventional early warning methods? The *signaling method* is perhaps the most widely used simple method to extract signals from early warning indicators. The method alerts whenever an indicator moves beyond a threshold specified by the policymaker. The threshold value is optimized based on the history of financial crises and the policymaker's loss function.

In Table 6, we consider an out-of-sample (2003 to 2014) comparison of the unit root methods with the conventional signaling method. The unit root methods either use an optimized significance level parameter α based on 1980–1999 training data or our fixed α=0.05 setup. The signaling method is based on a threshold optimized with the 1980–1999 training data. With either fixed or optimized α, the relative usefulness is higher on average than with the conventional signaling method, but there is no strict dominance. With α fixed, the unit root methods perform similarly for the short sample as well as for the full sample, and the differences between ADF and PSY are the same as highlighted earlier. When α is optimized with the training data, some indicators perform better and others worse. The net effect, however, is that the relative usefulness values drop slightly for both methods. For ADF, the average relative usefulness falls from 0.22 to 0.20. For PSY, the drop is from 0.28 to 0.24. While optimizing the significance level, parameter α ex-ante based on training data produced on average no additional benefit out-of-sample compared to using the simple α=0.05 assumption. Ex post optimizing of the significance level would, of course, still be beneficial.

Contrary to most other prediction models, our unit root indicator does not crucially depend on the choice of the threshold values of the trend gaps, but rather the choice of lag lengths, window sizes, significance levels, etc. To investigate the sensitivity to different window lengths and the significance level parameter, Figure 4 presents a plot of the trade-off between shares of type I and type II errors for the ADF method.[13] Figure 4 also shows the ROC curve for the conventional signaling method. These results seem to follow a familiar pattern for most variables, i.e. the credit and debt-service ratios generally outperform other variables, but specific performance depends on the parameter choices for the model. As a rule, most parameter choices do not lead to wildly divergent outcomes. Indeed, some choices perform better than the parameter choices used to obtain our main results. A notable example is the real estate price-to-rent ratio, where a longer window length produces a much better result.

Finally, a look at the results with banks' loan losses as the crisis indicator concur with our earlier finding on the ability to predict banking system distress and financial crises. In all cases, the big peaks in the loans losses/lending ratios can be predicted with the early warning indicator (in Figure 5, the credit-to-GDP ratio is used in computing the indicator values). Thus, the results do not seem to be overly sensitive to the choice of crisis definition. The definition

[13] The respective curves would also be qualitatively similar for the PSY method.

of crisis periods is, of course, something that is important from the point of view of evaluation of ex post performance of different methods, but it is far more important to traditional prediction methods (logit models, neural networks) than to our unit root indicator, which does not necessarily require historical data for estimates or training (for additional discussion, see Drehmann and Juselius, 2013).

3.2 Composite indicators

Combining information from multiple indicators can bolster the performance of unit root methods. Recall from section 2.5 that composites are formed by counting simultaneous alerts from individual indicators. For example, either "at least one alerts" or "both alert" may be used as warning signal indicators. Given the vast number of possible composite indicators combinations, however, we restrict the number of components to three and always include our best-performing single indicator, total credit-to-GDP ratio, as a component in each composite index. Figure 6 depicts an example the composite indicator SUM 3.1.

We again first consider the case with a fixed confidence level parameter, α=0.05, and evaluation in our full sample (1980–2012). The results are reported in Table 7b. For the rolling window ADF method, the best two-component index is composed of credit-to-GDP ratio and household debt-servicing cost ratio (SUM 2.3 in Table 7b) where "at least one indicator alerts" is considered an alert. The respective relative usefulness (0.52) is 18% higher than for the credit-to-GDP ratio alone. This is also the optimal two-component index for the PSY method, although the relative usefulness with this index, 0.55, is only marginally better than with than 0.54 with total credit-to-GDP alone. For both methods, there are also other two-component composite indices that yield similar or better performance than the credit-to-GDP indicator alone, as well as combinations that reduce signaling quality.

For the rolling window ADF method, the best three-component index comprise the two components of the best two-component index and the debt-servicing ratio of non-financial corporations (SUM 3.5 in Table 7b) such that at least one indicator alerting is considered a signal. The respective relative usefulness is only marginally better (Ur=0.53) than with the two-component indicator. For the PSY method, there are two equally good three-component indices (Ur=0.59). In both cases, two indicators alerting is considered a signal. The first combines credit-to-GDP ratio with total credit to households and the debt-servicing ratio (SUM 3.6 in Table 7b). The second combines the credit-to-GDP ratio with total credit to households and the real stock price index (SUM 3.7 in Table 7b). Again, in both cases there are other composite indices with nearly as high relative usefulness.

Further analysis reveals that the composite indicators have quite stable performance in the two sub-periods. Composites that perform well during their 1980–1999 training, continue to perform relatively well in the 2003–2012 out-of-sample estimation (see Table 7c). However, due to random variation, the performance differences between the best-performing composites are so small that there generally is no single best composite index that would be the best in all sub-samples.

In any case, the results suggest that even in the simplest form, it is useful to consider the warning signals from several indicators at the same time. If a policymaker considers only a few indicators, more weight should go to 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 and stock price developments add positively to the aggregate useful information. While this approach reduces the number of missed alerts significantly, our results also indicate a substantial increase in the number of false alerts. This leads to better usefulness values, especially when the policymaker is more averse to missing a crisis than false alerts ($\theta > 0.5$).

4. Concluding remarks

This study has demonstrated that an early warning indicator or set of indicators based on unit root testing can help in predicting financial crises, assuming, of course, that the indicators are based on relevant time-series information and are computed using an appropriate set of parameter values e.g. windows length and number of lags. Although the choice of these values is determinative and creates a certain amount of specification uncertainty, unit root indicator approaches have several unquestionable advantages. They are easy to compute and flexible. They may be used with different time frequencies. There is no upper limit in principle for the size of the data in terms of the number of indicators. They also allows repeated tests and accumulation of information. For example, using weekly data, one can scrutinize the pattern of repeated warnings instead of considering the test procedure as one-shot experiment. Although we have a way to go in identifying the full advantages of this approach, the promise of unit root methods is indisputable.

As for the extensions, a compelling issue is how to combine the alerts from different variables to some kind of a risk measure. For instance, one could formulate an additive model, search for the optimal combination of variables and optimize the weights with which each variable is added to the risk measure. The model could also take in account the length of the alert, e.g. explosive growth of the credit-to-GDP ratio continuing over several years certainly predicts a higher risk of a crisis than a few isolated alerts. Naturally, performing such optimization using historical data means one is betting on the probability that future crises will unfold in the same way as past crises, an assumption that may or may not hold. The underlying irony of Reinhart & Rogoff's "this time is different" policymaker excuse, of course, is that historical regularities seem to be strikingly persistent.

References

Alessi L, Detken C. 2011. Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity. *European Journal of Political Economy* **27**(3), 520–533.

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.

Betz F, Oprică S, Peltonen T, Sarlin P. 2013. Predicting distress in European banks. ECB Working Paper 1597.

Borio C. 2012. The financial cycle and macroeconomics: What have we learnt? Working Paper 2012/395, Bank for International Settlements.

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

Bussière M, Fratzscher M. 2008. Low probability, high impact: Policy making and extreme events. *Journal of Policy Modeling* **30**(1), 111–121.

Campbell JY, Lo AW, McKinlay AC. 1997. *The Econometrics of Financial Markets*. Princeton University Press, NJ.

Campbell Y, Shiller R. 1988a. The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* **1**(3), 195–228.

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

Cochrane J. 1992. Explaining the variance of price-dividend ratios. *Review of Financial Studies* **5**(2), 243–280.

Corsi F, Sornette D. 2014. Follow the money: The monetary roots of bubbles and crashes. *International Review of Financial Analysis* **32**, 47–59.

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

Demirgüc-Kunt A, Detragiache E. 2000. Monitoring banking sector fragility: A multivariate logit approach. *World Bank Economic Review* **14**, 287–307.

Drehmann M, Borio M, Tsatsaronis K. 2011. Anchoring countercyclical capital buffers: The role of credit aggregates. *International Journal of Central Banking* **7**, 189–240.

Drehmann M, Juselius M. 2013. Evaluating early warning indicators of banking crises: Satisfying policy requirements. BIS Working Papers No. 421.

Eidenberger J, Neudorfer B, Sigmund M, Stein I. 2014. What predicts financial (in)stability? A Bayesian approach. Discussion Paper Deutsche Bundesbank No. 36/2014.

Elliot G. 1999. Efficient tests for a unit root when the initial observation is drawn from its unconditional distribution. International Economic Review **40**, 767–783.

Elliot G, Rothenberg TJ, Stock JH. 1996. Efficient tests for an autoregressive unit root. *Econometrica* **64**, 813–836.

Ferrari S, Pirovano M. 2015. Early warning indicators for banking crises: A conditional moments approach. MPRA Discussion Paper No. 62406. Available at https://mpra.ub.uni-muenchen.de/62406/.

Franses P. 2013. Are we in a bubble? A simple time-series-based diagnostic. Erasmus School of Economics, Econometric Institute Report 2013-12.

Gerdrup K, Bakke Kvinlog A, Schaanning E. 2013. Key indicators for a countercyclical capital buffer in Norway – Trends and uncertainty. Staff memo. Norges Bank.

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

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

Jorda O, Schularick M, Taylor A. 2015. Leveraged bubbles. NBER Working Paper.

Homm U, Breitung J. 2012. Testing for speculative bubbles in stock markets: A comparison of alternative methods. *Journal of Financial Econometrics* **10**(1), 198–231.

Kalatie S, Laakkonen H, Tölö E. 2015. Indicators used in setting the countercyclical capital buffer. Bank of Finland Discussion Paper, No. 8/2015.

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

Koustas Z, Serletis A. 2005. Rational bubbles or persistent deviations from market fundamentals? *Journal of Banking and Finance* **29**, 2523–2539.

Leybourne S. 1995. Testing for unit roots using forward and reverse Dickey-Fuller regression. *Oxford Bulletin of Economics and Statistics* **57**, 559–571.

Leybourne S, Kim T, Taylor A. 2006. Regression-based test for a change in persistence. *Oxford Bulletin of Economics and Statistics* **68**(5), 595–621.

MacKinnon JG. 1994. Approximate asymptotic distribution functions for unit-root and cointegration tests. *Journal of Business and Economic Statistics* **12**, 167–176.

Peltonen T, Piloiu A, Sarlin P. 2014. Tail-dependence measures to predict bank distress. Unpublished mimeo.

Phillips P, Wu Y, Yu J. 2011. Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? *International Economic Review* **52**, 201–226.

Phillips P, Shi P, Yu J. 2013. Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P500. Cowles Foundation Discussion Paper 1914.

Reinhart C, Rogoff K. 2009. The aftermath of financial crises. *American Economic Review* **99**(2), 466–472.

Repullo R, Saurina J. 2011. The countercyclical capital buffer of Basel III: A critical assessment. CEPR Discussion Paper No. 8304.

Ristolainen K. 2016. How sensitive are financial crises predictions to the timing of crises? University of Turku, unpublished mimeo.

Sarlin P. 2013. On policymakers' loss functions and the evaluation of early warning systems. ECB Working Paper No. 1509.

Schularick M, Taylor A. 2012. Credit booms gone bust: Monetary policy, leverage cycles and financial crises, 1870–2008. *American Economic Review* **102**(2), 1029–1061.

Taipalus K. 2006a. Bubbles in the Finnish and US equities markets. Bank of Finland Studies, Series E, No. 3.

Taipalus, K. 2006b. A global house price bubble? Evaluation based on a new rent-price approach. Bank of Finland Research Discussion Papers 29/2006.

Taipalus K. 2012. Detecting asset price bubbles with time-series methods. Bank of Finland Scientific Monographs E47: 2012.

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

Wilcox D. 1989. The sustainability of government deficits: Implications for present value borrowing constraints. *Journal of Money, Credit and Banking* **54**, 1837–1847.

Table 1. Names and definition of all tested variables

Variable	F(*)	L(**)	Content
Total credit-to-GDP	Q	x	Ratio of (nominal) total credit to the private non-financial sector to (nominal) GDP
Household debt service ratio	Q	x	Debt service to income ratio, households
Debt service ratio	Q	x	Debt service to income ratio, households and non-financial corporations
Total credit	Q	x	Total credit to private non-financial sector, in billion local currency (real)
Total NFC credit (nominal)	Q	x	Total credit to non-financial corporations, in billion local currency (nominal)
Total NFC credit (real)	Q	x	Total credit to non-financial corporations, in billion local currency (real)
Bank credit-to-GDP	Q	x	Ratio of (nominal) bank credit to the private non-financial sector to (nominal) GDP
Bank credit	Q	x	Bank credit to private non-financial sector, in billion local currency (real)
Total credit (nominal)	Q	x	Total credit to private non-financial sector, in billion local currency (nominal)
Commercial RE price (real)	Q	x	Commercial property price index (real, source: ECB)
NFC credit-to-GDP	Q	x	Ratio of (nominal) total credit to non-financial corporations to (nominal) GDP
Bank credit (nominal)	Q	x	Bank credit to private non-financial sector, in billion local currency (nominal)
NFC debt service ratio	Q	x	Debt service to income ratio, non-financial corporations
Gross domestic product (real)	Q	x	Gross domestic product, quarterly levels, in million local currency (real)
Commercial RE price (nominal)	Q	x	Commercial property price index (nominal, source: ECB)
Commercial RE price (nominal)	Q	x	Commercial property price index (nominal, source: ECB)
M3 (real)	Q	x	Monetary aggregate M3 (real)
M3 (nominal)	Q	x	Monetary aggregate M3 (nominal)
Residential RE price (OECD, real)	Q	x	Residential property price index (real, source: OECD)
Residential RE price (ECB, nominal)	Q	x	Residential property price index (nominal, source: ECB)
Nominal interest rate	Q		Three-month money market interest rates (nominal)
Consumer price index	Q	x	Consumer price index
Stock price (nominal)	Q		Stock price index (nominal)
Residential RE price-to-rent	Q	x	Residential real estate price to rent index. (Source: OECD)
Gross domestic product (nominal)	Q	x	Gross domestic product, quarterly levels, in million local currency (nominal)
Government debt to GDP	Q	x	General government consolidated gross debt
Stock price	Q		Stock price index (real)
Residential RE price (OECD, nominal)	Q	x	Nominal real estate price index (OECD)
Gross disposable income	Q	x	Gross disposable income of households in million local currency
Loans to income	Q	x	Ratio of household loans to gross disposable income
Residential RE price-to-income	Q	x	Residential real estate price to income index. (Source: OECD)
Total household credit (nominal)	Q	x	Total credit to households, in billion local currency (nominal)
Household credit-to-GDP	Q	x	Ratio of (nominal) total credit to households to (nominal) GDP
Total household credit (real)	Q	x	Total credit to households, in billion local currency (real)
Housing loans (real)	Q	x	Loans for house purchases, in million Euros, outstanding amounts (real)
Housing loans (nominal)	Q	x	Loans for house purchases, in million Euros, outstanding amounts (nominal)
Unemployment rate	Q	x	Unemployment rate
Government bond yield	Q		Long-term government bond yield (nominal)
Rents	Q	x	Actual rentals paid by tenants including other actual rentals, index
Residential RE price (nominal)	M		Residential real estate nominal price index (various sources)
Residential RE price (real)	M		Residential real estate real price index (various sources, only certain countries)
Stock price (nominal)	M		Stock price index (nominal, source: Bloomberg)

Table 2. Usefulness values of rolling window ADF test for all variables, window sizes

	Window length=12			Window length=24			Window length=36		
Variable	Ur	FPR	FNR	Ur	FPR	FNR	Ur	FPR	FNR
Total credit-to-GDP	0.43	0.19	0.38	0.29	0.22	0.49	0.39	0.23	0.38
Household debt-service ratio	0.39	0.17	0.44	0.34	0.13	0.52	0.28	0.14	0.57
Debt-service ratio	0.37	0.10	0.52	0.26	0.08	0.65	0.26	0.05	0.69
Total credit	0.40	0.34	0.26	0.17	0.39	0.44	0.17	0.40	0.43
Total NFC credit (nominal)	0.35	0.42	0.23	0.07	0.40	0.52	-0.19	0.38	0.81
Total NFC credit (real)	0.34	0.29	0.37	0.10	0.33	0.57	0.03	0.37	0.60
Bank credit-to-GDP	0.34	0.20	0.46	0.39	0.22	0.38	0.45	0.24	0.31
Bank credit	0.34	0.33	0.33	0.30	0.37	0.33	0.30	0.40	0.31
Total credit (nominal)	0.33	0.43	0.24	0.13	0.44	0.43	0.06	0.41	0.53
Commercial RE price (real)	0.27	0.16	0.57	0.17	0.12	0.70	0.17	0.10	0.73
NFC credit-to-GDP	0.26	0.17	0.57	0.14	0.16	0.70	0.07	0.18	0.75
Bank credit (nominal)	0.27	0.46	0.27	0.21	0.45	0.34	0.20	0.43	0.37
NFC debt-service ratio	0.25	0.13	0.62	0.17	0.09	0.74	0.13	0.10	0.78
Gross domestic product (real)	0.23	0.35	0.42	0.21	0.38	0.42	0.02	0.30	0.68
Commercial RE price (nominal)	0.23	0.26	0.51	0.20	0.20	0.59	0.18	0.20	0.63
Commercial RE price (nominal)	0.23	0.26	0.51	0.20	0.20	0.59	0.18	0.20	0.63
M3 (real)	0.19	0.29	0.52	0.23	0.32	0.44	0.23	0.32	0.45
M3 (nominal)	0.21	0.38	0.42	0.25	0.31	0.43	0.08	0.27	0.65
Residential RE price (OECD, real)	0.15	0.23	0.62	0.11	0.21	0.68	0.17	0.17	0.66
Residential RE price (ECB, nominal)	0.14	0.33	0.53	0.07	0.36	0.57	0.15	0.38	0.47
Nominal interest rate	0.14	0.03	0.82	0.00	0.01	0.99	0.00	0.01	0.99
Consumer price index	0.12	0.39	0.49	-0.5	0.35	0.70	-0.07	0.28	0.79
Stock price (nominal)	0.12	0.15	0.74	0.01	0.14	0.85	-0.02	0.16	0.87
Residential RE price-to-rent	0.13	0.19	0.68	0.10	0.17	0.73	0.13	0.14	0.72
Gross domestic product (nominal)	0.14	0.38	0.48	0.12	0.29	0.58	-0.07	0.25	0.81
Government debt to GDP	0.11	0.05	0.85	0.07	0.10	0.83	0.17	0.00	0.83
Stock price	0.10	0.11	0.79	-0.02	0.10	0.92	-0.01	0.08	0.93
Residential RE price (OECD, nominal)	0.10	0.34	0.56	0.05	0.35	0.60	0.19	0.31	0.49
Gross disposable income	0.07	0.25	0.68	0.11	0.32	0.57	-0.12	0.37	0.75
Loans to income	0.07	0.40	0.53	0.38	0.28	0.33			
Residential RE price-to-income	0.07	0.16	0.77	0.12	0.16	0.72	0.17	0.16	0.67
Total household credit (nominal)	0.05	0.48	0.47	0.06	0.48	0.47	0.11	0.44	0.46
Household credit-to-GDP	0.03	0.32	0.65	0.23	0.31	0.46	0.35	0.29	0.36
Total household credit (real)	0.01	0.42	0.56	0.18	0.45	0.37	0.27	0.41	0.31
Housing loans (real)	0.00	0.42	0.58	0.06	0.46	0.47	0.32	0.43	0.25
Housing loans (nominal)	0.00	0.43	0.58	0.14	0.44	0.42	0.18	0.44	0.38
Unemployment rate	-0.01	0.06	0.96	-0.01	0.04	0.96	-0.03	0.03	1.00
Government bond yield	-0.04	0.04	1.00	-0.01	0.01	1.00	0.00	0.00	1.00
Rents	-0.16	0.50	0.66	-0.06	0.45	0.61	0.26	0.20	0.54

Ur denotes the relative usefulness value, FN (FP) the proportion (%) of false negative (positive) predictions. The policymaker's preference parameter is θ=0.5. Data cover 1980 to 2012. A one-quarter publication lag is used for quarterly variables except for equity prices.

Table 3. Prediction probabilities and the usefulness values for fixed confidence level parameter. Data covers 1980 to 2012 and the confidence level parameter α is fixed at 0.05.

		Usefulness				False ratios (%)			
		θ=0.5		θ=0.6		ADF		PSY	
Variable	F	ADF	PSY	ADF	PSY	FPR	FNR	FPR	FNR
Bank credit-to-GDP	Q	0.34	0.54	0.11	0.46	18.0	4.6	26.3	1.7
Bank credit	Q	0.34	0.39	0.17	0.34	29.8	3.4	46.8	0.9
Total credit-to-GDP	Q	0.44	0.53	0.25	0.45	16.3	3.9	28.5	1.5
Household credit-to-GDP	Q	0.04	0.34	-0.29	0.23	27.3	8.7	38.6	2.9
Total credit	Q	0.40	0.36	0.26	0.33	29.9	2.7	51.8	0.6
Total household credit (real)	Q	0.02	0.31	-0.26	0.28	35.6	7.6	54.4	0.8
Loans to income	Q	0.01	0.18	-0.27	0.03	32.9	13.7	38.2	7.6
Debt service ratio	Q	0.37	0.34	0.10	0.07	9.3	5.4	9.4	5.6
Household debt service ratio	Q	0.41	0.44	0.20	0.25	14.0	6.9	15.3	6.1
NFC debt service ratio	Q	0.25	0.16	-0.06	-0.21	10.2	9.2	9.3	10.8
Residential RE price-to-income	Q	0.07	0.20	-0.32	-0.11	13.4	9.1	17.8	7.0
Residential RE price-to-rent	Q	0.14	0.22	-0.20	-0.05	16.1	8.5	20.3	6.8
Residential RE price (OECD, real)	Q	0.15	0.30	-0.16	0.10	20.5	7.5	24.6	5.1
Stock price (nominal)	Q	0.11	0.23	-0.27	-0.05	12.7	7.9	19.8	5.8
Stock price	Q	0.10	0.19	-0.30	-0.15	9.5	8.4	11.2	7.2
Residential RE price (ECB, nominal)	Q	0.12	0.26	-0.15	0.15	29.4	6.5	45.2	2.7
Residential RE price (OECD, nominal)	Q	0.10	0.19	-0.18	0.06	29.3	6.8	48.3	3.1
Residential RE price (nominal)	M	0.22	0.14	0.15	-0.02	52.6	2.5	42.6	6.7
Residential RE price (real)	M	0.12	0.19	-0.07	-0.05	39.0	7.8	26.2	9.7
Stock price (nominal)	M	0.30	0.18	0.12	-0.17	29.6	4.3	12.4	8.3

F denotes frequency of the time series. Window length (ADF) and minimum window length (PSY) is 12 quarters (or 36 months). θ is the policymaker's preference parameter for false positive rate vs. false negative rate. FP = False Positive, FN = False Negative. A one-quarter publication lag is used for quarterly variables, except for equity prices.

Table 4. Number of predicted crises and false alarms. Data covers 1980 to 2012 and the confidence level parameter α is fixed to 0.05.

Variable	F	Crises found ADF	Crises found PSY	False alarms ADF	False alarms PSY	N of crises
Bank credit-to-GDP	Q	13	17	20	22	17
Bank credit	Q	17	17	31	31	17
Total credit-to-GDP	Q	13	17	17	21	17
Household credit-to-GDP	Q	10	14	16	16	16
Total credit	Q	15	17	33	33	17
Total household credit (real)	Q	13	16	25	25	16
Loans to income	Q	5	5	3	2	6
Debt service ratio	Q	10	9	7	9	16
Household debt service ratio	Q	8	9	4	6	12
NFC debt service ratio	Q	6	4	5	5	11
Residential RE price-to-income	Q	7	10	6	10	15
Residential RE price-to-rent	Q	10	13	13	15	16
Residential RE price (OECD, real)	Q	13	13	16	16	17
Stock price (nominal)	Q	5	8	14	25	17
Stock price	Q	3	4	9	11	17
Residential RE price (ECB, nominal)	Q	10	12	17	18	13
Residential RE price (OECD, nominal)	Q	12	15	27	28	17
Residential RE price (nominal)	M	8	7	8	7	8
Residential RE price (real)	M	6	5	7	6	8
Stock price (nominal)	M	8	4	21	10	12

F denotes frequency of the time series. *Criteria for predicted crisis = an alert of six or more consecutive quarters during a 5-year pre-crisis window. Breaks of one in the alerts are allowed. Window length (ADF) and minimum window length (PSY) is 12 quarters (or 36 months). θ=0.5 is the policymaker's preference parameter for false positive rate vs false negative rate. A one-quarter publication lag is used for quarterly variables, except for equity prices.

Table 5. Alerting leads of both tests. The alerting lead (i.e. how many quarters before a crisis we are most likely to get a signal from this variable) is calculated separately for the rolling window ADF and PSY tests using the methodology explained in Section 2.4.

Variable	Alerting lead, quarters ADF	PSY
Bank credit-to-GDP	10	10
Bank credit	12	13
Total credit-to-GDP	8	5
Household credit-to-GDP	21	21
Total credit	8	5
Total household credit (real)	17	12
Loans to income	19	16
Debt service ratio	7	5
Household debt service ratio	6	9
NFC debt service ratio	5	4
Residential RE price-to-income	20	24
Residential RE price-to-rent	10	10
Residential RE price (OECD, real)	15	9
Stock price (nominal)	6	6
Stock price	6	10
Residential RE price (ECB, nominal)	16	16
Residential RE price (OECD, nominal)	10	24

Each data series has quarterly frequency. Data covers 1980 to 2012 and the confidence level parameter α is fixed at 0.05. Window length (ADF) and minimum window length (PSY) is 12 quarters (or 36 months). A one-quarter publication lag is used for quarterly variables, except for equity prices.

Table 6. Out-of-sample performance comparison. Threshold-optimized out-of-sample usefulness is based on evaluation period 2003–2012, where the confidence level parameter α is optimized based on training with 1980–1999 data. In case of fixed threshold, the confidence level parameter α is fixed at 0.05.

		Usefulness, opt threshold		Usefulness, fixed thresholc		Optimal critical value		Usefulness, benchmark	Usefulness ratio	
Variable	F	ADF	PSY	ADF	PSY	ADF	PSY	signaling	ADF	PSY
Bank credit-to-GDP	Q	0.28	0.45	0.35	0.43	-0.98	1.24	0.25	1.12	1.79
Bank credit	Q	0.36	0.46	0.35	0.29	-0.31	3.23	0.38	0.94	1.21
Total credit-to-GDP	Q	0.45	0.31	0.51	0.38	-1.08	-0.12	0.27	1.68	1.15
Household credit-to-GDP	Q	0.10	0.24	-0.12	0.26	-1.68	0.36	0.28	0.36	0.86
Total credit	Q	0.57	0.43	0.57	0.30	-0.26	3	0.22	2.58	1.96
Total household credit (real)	Q	0.17	0.20	0.08	0.19	-2.94	1.07	0.27	0.62	0.74
Loans to income	Q	-0.02	0.15	0.01	0.18	0	0	-0.03	n/a	n/a
Debt service ratio	Q	0.50	0.46	0.52	0.47	-1.23	0.25	0.28	1.79	1.64
Household debt service ratio	Q	0.39	0.39	0.46	0.50	-1.55	-0.79	0.32	1.20	1.23
NFC debt service ratio	Q	0.35	0.38	0.29	0.18	-2.03	-1.32	0.09	3.85	4.23
Residential RE price-to-income	Q	-0.05	0.09	-0.07	0.06	0.14	0.05	0.18	n/a	0.49
Residential RE price-to-rent	Q	0.00	0.10	0.02	0.17	0.46	1.41	0.03	n/a	3.35
Residential RE price (OECD, real)	Q	0.02	0.29	0.07	0.33	0.6	1.38	0.07	0.23	4.16
Stock price (nominal)	Q	0.04	0.50	0.23	0.34	1.13	-0.12	0.37	0.10	1.36
Stock price	Q	0.03	0.20	0.17	0.20	0.94	0.93	0.29	0.12	0.69
Residential RE price (ECB, nominal)	Q	0.11	0.13	0.13	0.22	-1.22	2.82	0.11	1.04	1.18
Residential RE price (OECD, nomina	Q	-0.01	0.03	0.15	0.28	1.77	3.32	0.43	n/a	0.07
Average		0.19	0.28	0.22	0.28	-0.48	0.98	0.22	1.20	1.63
Median		0.11	0.29	0.17	0.28	-0.31	0.93	0.27	1.04	1.22

F denotes frequency of the time series. Window length (ADF) and minimum window length (PSY) is 12 quarters. The usefulness ratio is the usefulness of the unit root method (ADF or PSY with optimized threshold) divided by the usefulness of the benchmark method. The benchmark signaling method is based on the one-sided HP-filtered trend gap (or relative trend gap for non-ratio variables) with smoothing parameter λ=400,000. The policymaker's preference parameter is θ=0.5. A one-quarter publication lag is used for quarterly variables, except for equity prices.

Table 7. Performance statistics for composite indicators. Panel a) shows the variables included in each composite. Panel b) presents the evaluation results for the full-sample with fixed confidence level parameter α=0.05. Panel c) presents the 2003-2012 out-of-sample usefulness results both for confidence level parameter α that is optimized based on 1980-1999 data, and for the case that α is fixed to 0.05.

a) Variables included in each composite

Composite	Variable 1	Variable 2	Variable 3
SUM 2.1	Total credit-to-GDP	Debt service ratio	
SUM 2.2	Total credit-to-GDP	NFC debt service ratio	
SUM 2.3	Total credit-to-GDP	Household debt service ratio	
SUM 2.4	Total credit-to-GDP	Total household credit (real)	
SUM 3.1	Total credit-to-GDP	NFC debt service ratio	Residential RE price-to-rent
SUM 3.2	Total credit-to-GDP	Bank credit-to-GDP	NFC debt service ratio
SUM 3.3	Total credit-to-GDP	Total credit	Residential RE price-to-rent
SUM 3.4	Total credit-to-GDP	Total credit	Total household credit (real)
SUM 3.5	Total credit-to-GDP	NFC debt service ratio	Household debt service ratio
SUM 3.6	Total credit-to-GDP	Total household credit (real)	Debt service ratio
SUM 3.7	Total credit-to-GDP	Total household credit (real)	Stock price

The first number in the composite name denotes the number of variables in the composite. All composites are tested with all relevant alerting thresholds.

b) Full sample 1980-2012, fixed parameters

	At least one alerts						At least two alert						At least three alert					
	ADF			PSY			ADF			PSY			ADF			PSY		
Variable	Ur	FPR	FNR	Ur	FPR	FNR	Ur	FPR	FNR	Ur	FPR	FNR	Ur	FPR	FNR	Ur	FPR	FNR
Credit-to-GDP	0.44	16.3	3.9	0.53	28.5	1.5												
SUM 2.1	0.47	20.4	3.0	0.52	30.5	1.4	0.33	5.0	6.3	0.35	7.2	5.7						
SUM 2.2	0.46	18.2	3.5	0.53	28.9	1.5	0.24	6.8	10.0	0.16	8.4	10.9						
SUM 2.3	0.52	20.4	2.5	0.55	30.8	1.1	0.34	6.1	9.5	0.45	10.7	6.9						
SUM 2.4	0.39	34.4	2.3	0.40	48.4	0.6	0.15	10.2	9.8	0.54	26.3	2.1						
SUM 3.1	0.48	26.8	2.3	0.50	35.3	1.1	0.30	8.8	7.4	0.33	15.2	6.1	0.05	1.6	14.1	0.11	3.9	12.9
SUM 3.2	0.44	26.4	2.7	0.51	36.3	0.9	0.44	11.1	4.4	0.56	20.3	2.2	0.16	4.2	11.7	0.18	5.8	11.0
SUM 3.3	0.38	39.1	1.9	0.34	54.7	0.6	0.49	16.3	3.3	0.52	32.1	1.3	0.13	4.8	10.2	0.28	12.3	7.3
SUM 3.4	0.38	42.4	1.5	0.30	58.8	0.5	0.40	22.5	3.5	0.47	41.6	0.7	0.16	9.4	9.8	0.54	26.0	2.2
SUM 3.5	0.53	22.0	2.3	0.54	31.1	1.1	0.38	9.8	7.7	0.44	14.6	5.9	0.22	3.1	11.8	0.18	4.3	12.3
SUM 3.6	0.39	37.7	1.9	0.39	49.1	0.6	0.39	11.3	5.0	0.59	22.7	1.6	0.13	2.6	11.1	0.34	7.2	7.6
SUM 3.7	0.36	38.1	2.2	0.36	51.7	0.6	0.31	11.6	5.8	0.59	24.7	1.4	0.01	1.4	13.9	0.27	2.6	10.0

Window length (ADF) and minimum window length (PSY) is 12 quarters. Policymaker's preference parameter θ=0.5 is used in the usefulness calculation. FP = False Positive, FN = False Negative. A one- quarter publication lag is used for quarterly data, except for stock market data.

c) Short sample 2003–2012, optimized parameters

	At least one alerts				At least two alert				All three alert			
	α=0.05		α optimized		α=0.05		α optimized		α=0.05		α optimized	
Variable	ADF	PSY	ADF	PSY	ADF	PSY	ADF	PSY	ADF	PSY	ADF	PSY
Credit-to-GDP	0.51	0.38	0.45	0.31								
SUM 2.1	0.52	0.32	0.39	0.28	0.50	0.52	0.56	0.49				
SUM 2.2	0.49	0.36	0.41	0.25	0.31	0.20	0.41	0.45				
SUM 2.3	0.57	0.35	0.33	0.22	0.41	0.54	0.53	0.49				
SUM 2.4	0.41	0.21	0.18	0.18	0.20	0.38	0.48	0.35				
SUM 3.1	0.41	0.32	0.32	0.20	0.39	0.28	0.46	0.44	0.05	0.15	0.07	0.19
SUM 3.2	0.43	0.25	0.29	0.19	0.54	0.52	0.43	0.40	0.18	0.22	0.38	0.57
SUM 3.3	0.37	0.25	0.30	0.27	0.62	0.39	0.57	0.46	0.13	0.22	0.16	0.12
SUM 3.4	0.41	0.18	0.17	0.16	0.57	0.32	0.47	0.39	0.21	0.39	0.59	0.42
SUM 3.5	0.55	0.34	0.30	0.17	0.48	0.53	0.48	0.39	0.25	0.20	0.43	0.54
SUM 3.6	0.40	0.19	0.13	0.16	0.60	0.42	0.46	0.38	0.14	0.48	0.58	0.46
SUM 3.7	0.42	0.22	0.22	0.19	0.34	0.38	0.50	0.37	0.06	0.27	0.04	0.25

Window length (ADF) and minimum window length (PSY) are 12 are quarters. Policymaker's preference parameter θ=0.5 is used for usefulness calculation. A one-quarter publication lag is used for quarterly data, except for stock market data.

Figure 1. Rolling window ADF test alerts for the debt-service ratio variable

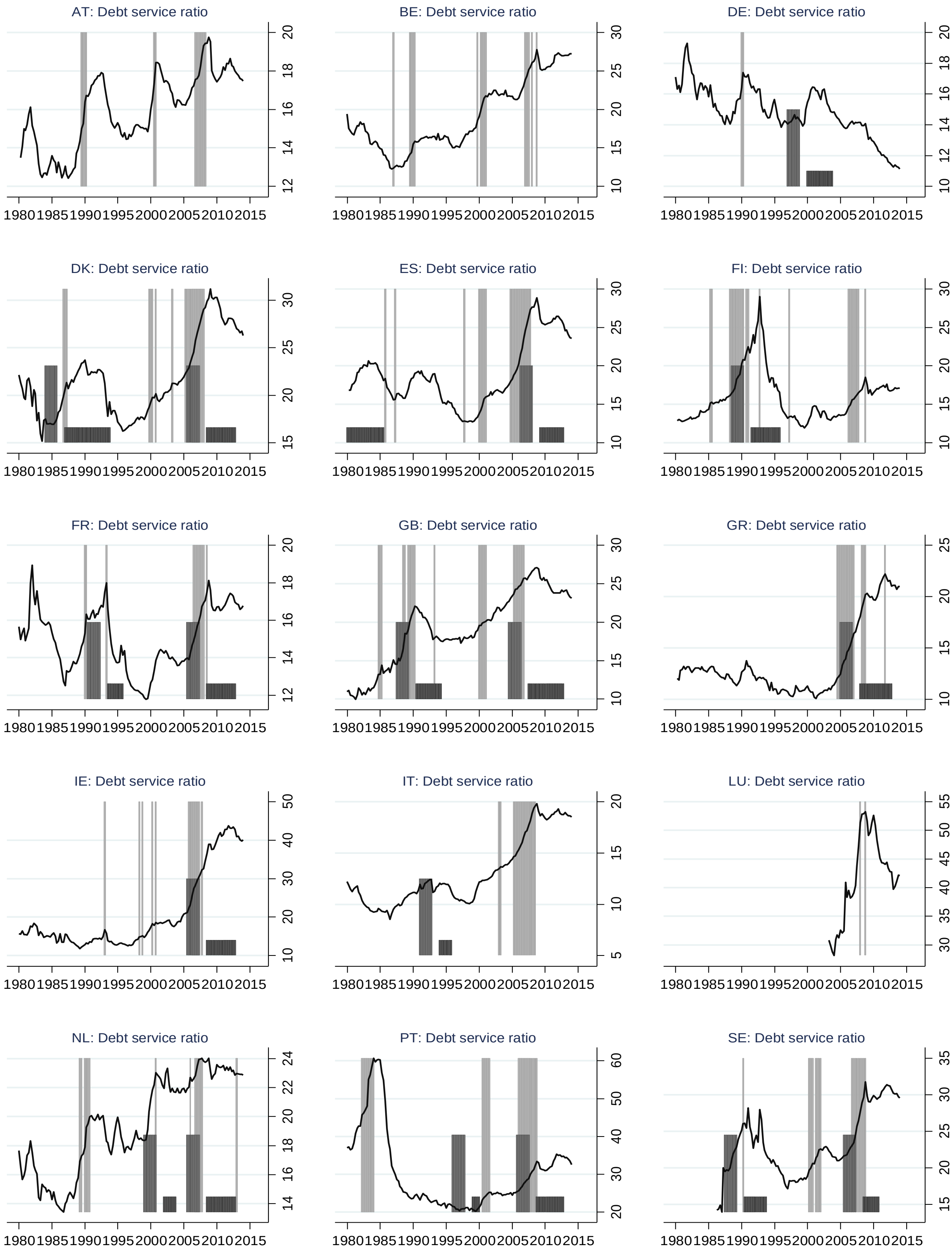


Figure 2. Rolling window ADF test alerts for the total credit-to-GDP ratio variable

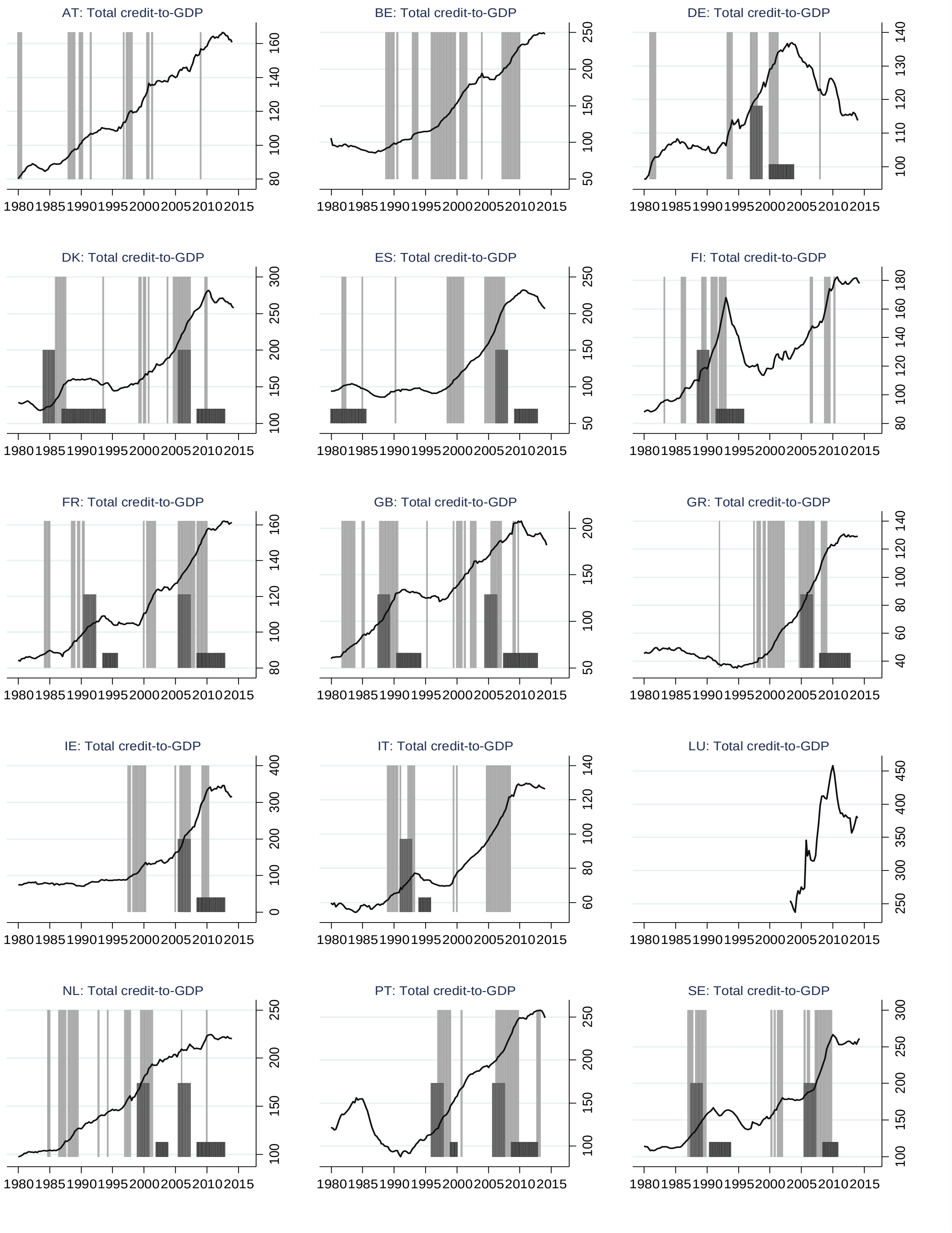


Figure 3. Rolling window ADF test alerts for the residential real estate price variable

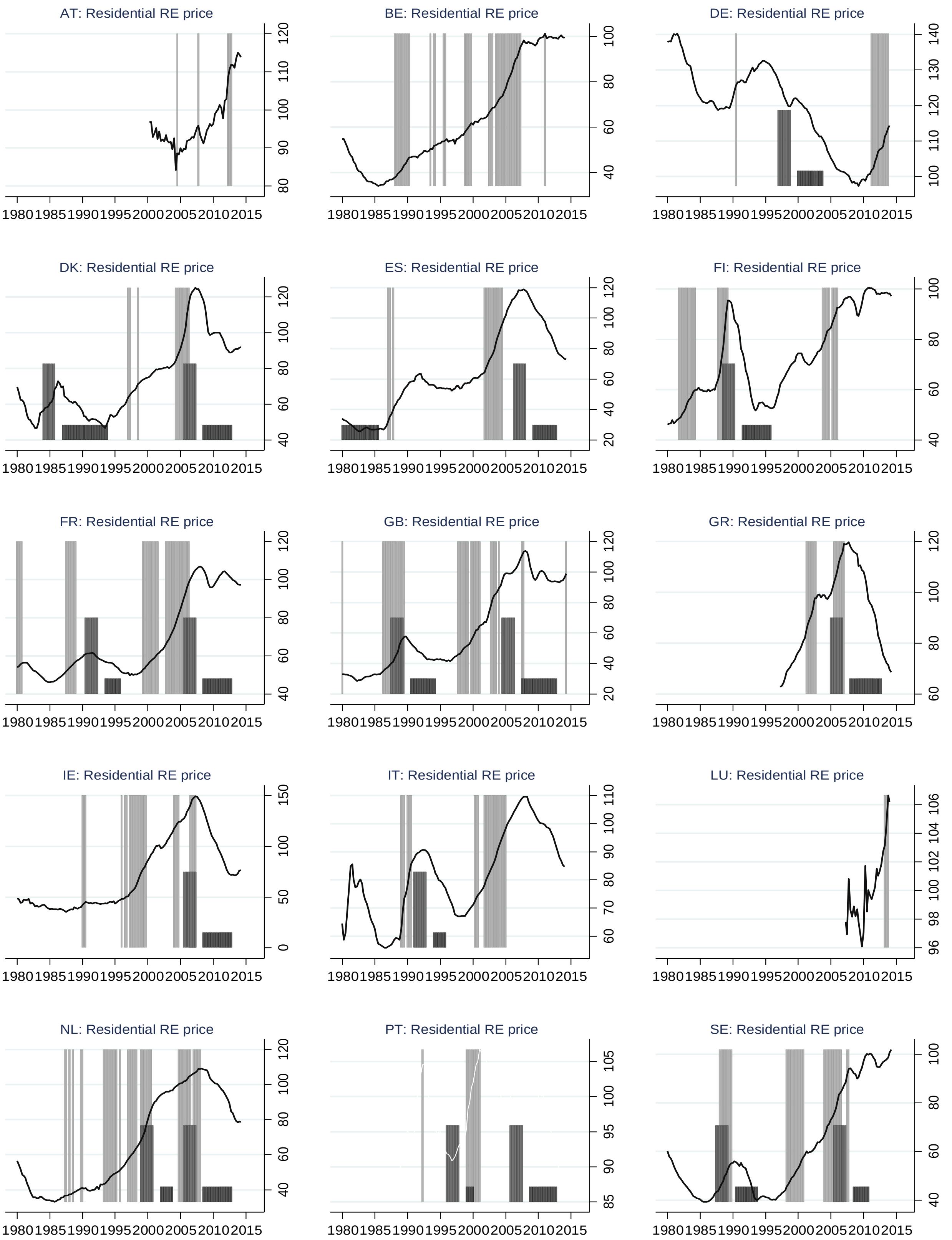
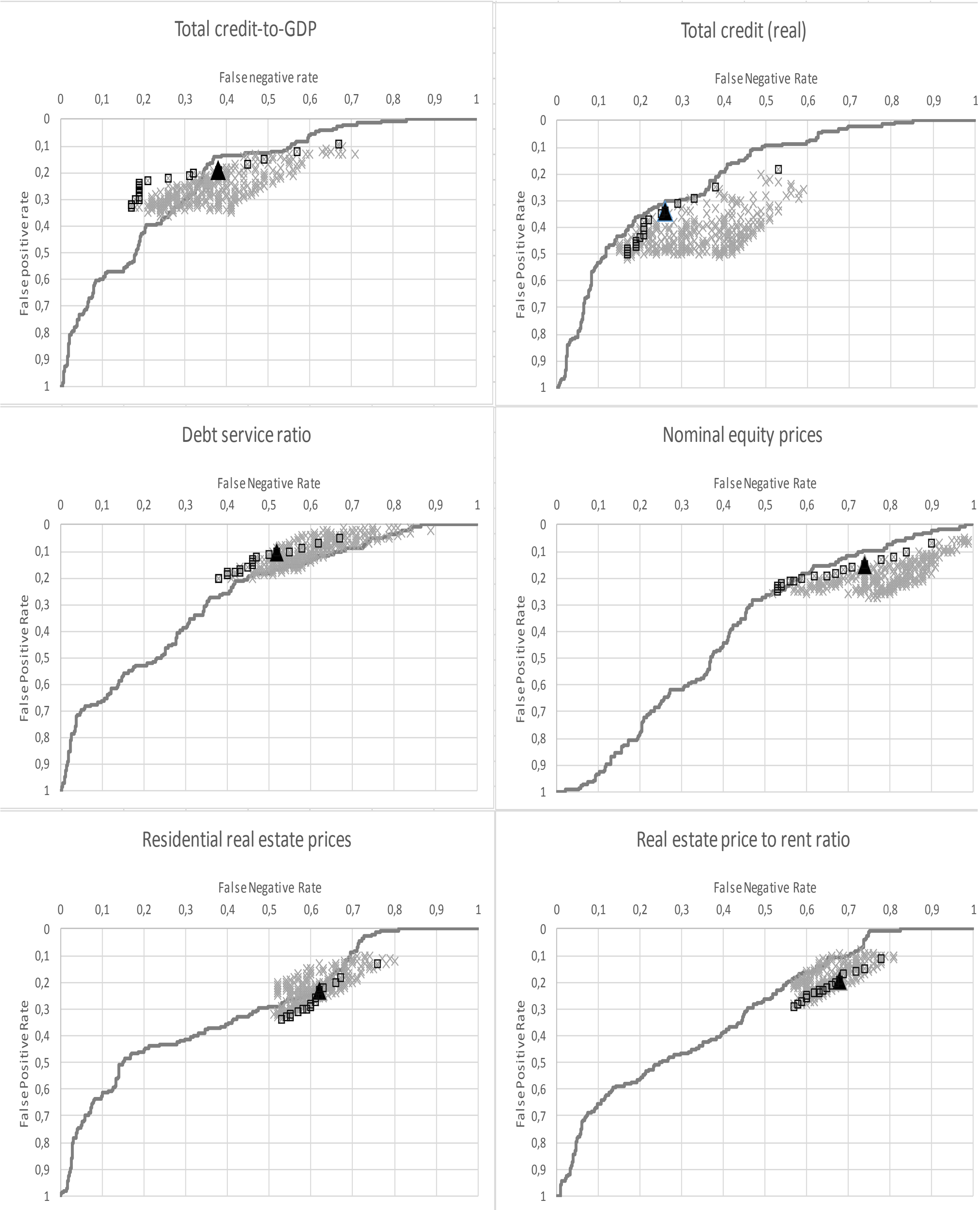
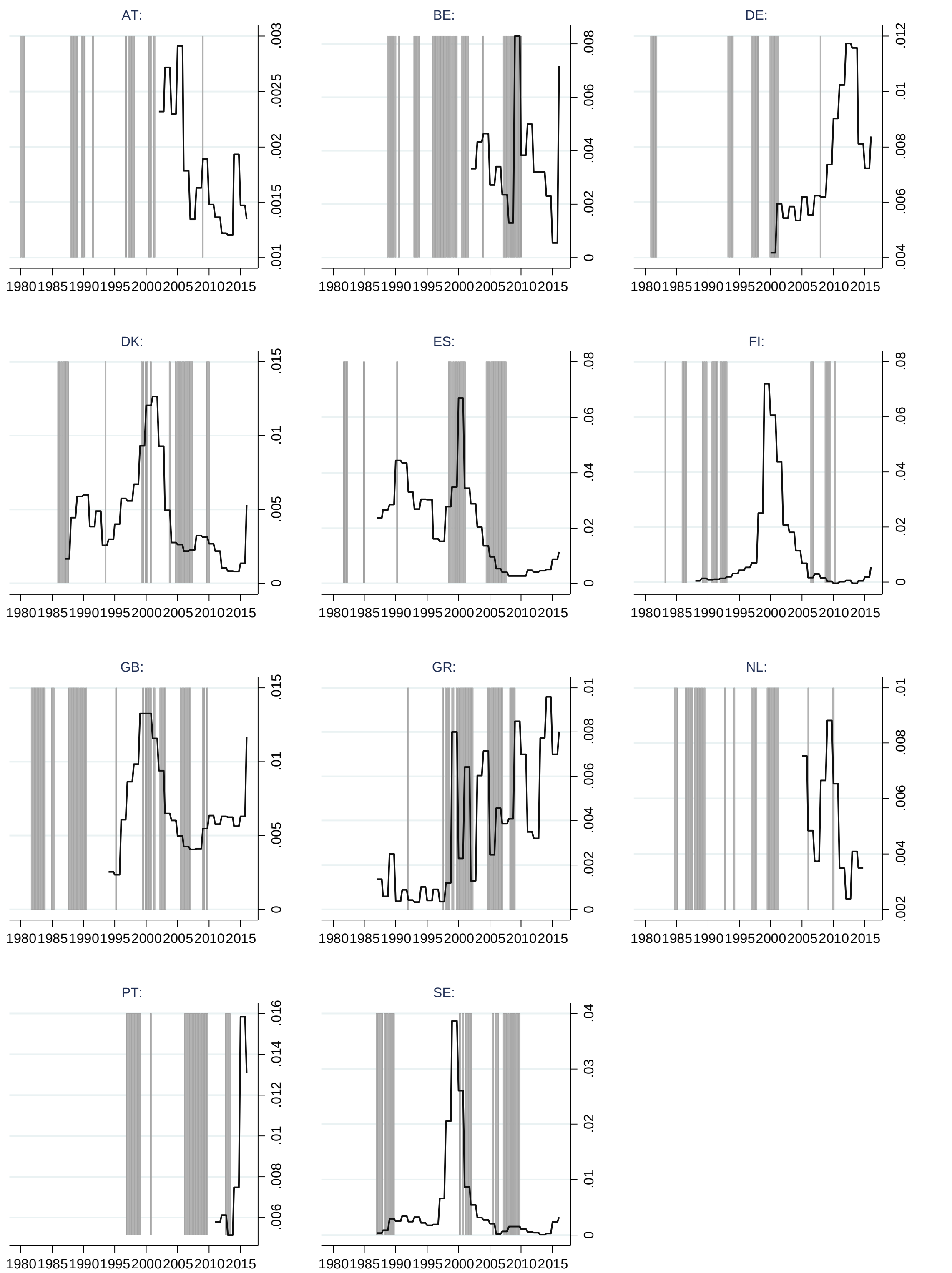


Figure 4. The ROC curve and impacts of different parameters



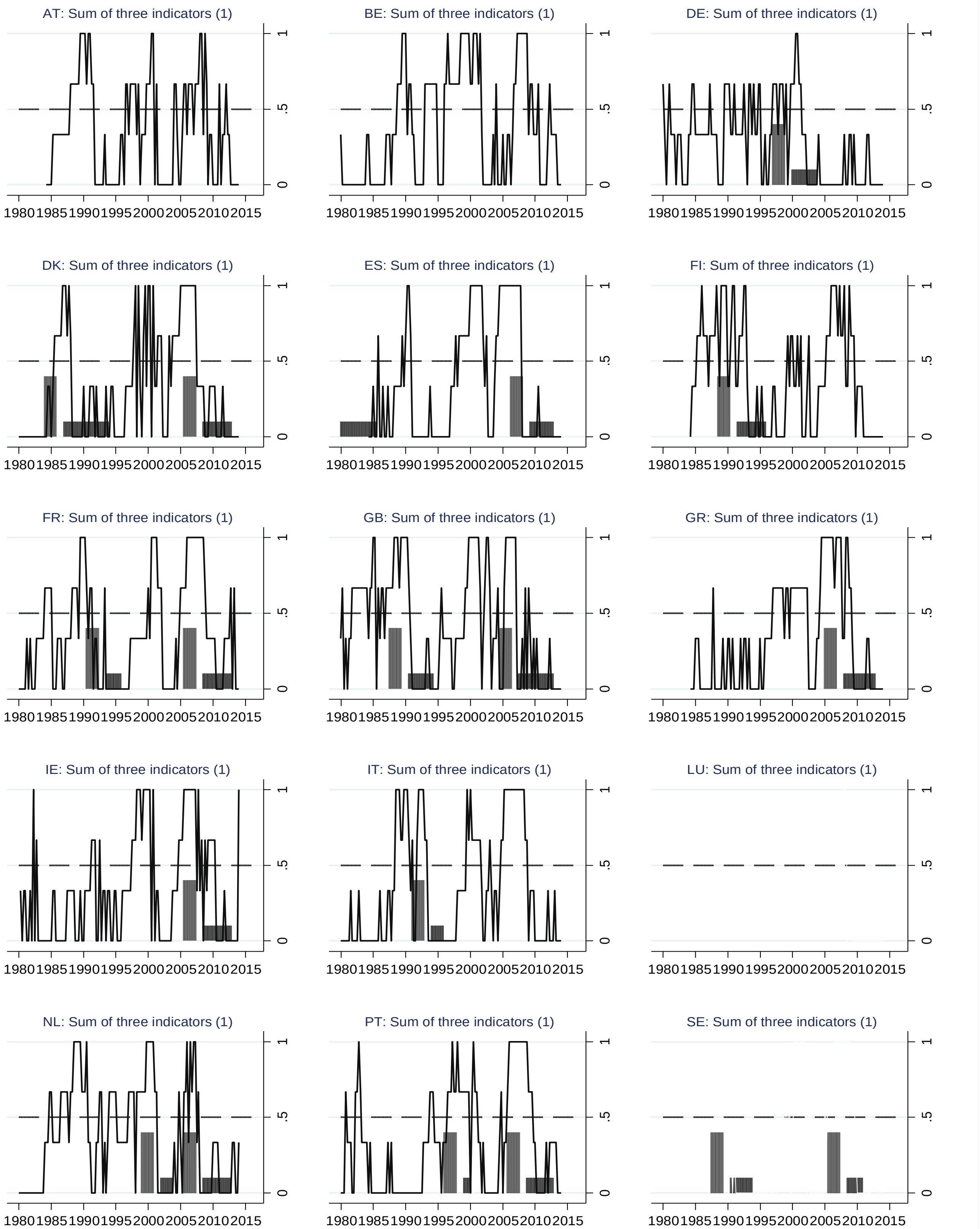
Solid line: The ROC curve of the relative trend gap. Grey boxes: FP and FN rates achieved with various window lengths and significance levels; Black boxes FP and FN rates achieved with a window length of 12 and various significance levels; black triangles: FP and FN rates achieved with a window length of 12 and significance level of 0.95.

Figure 5. Early warning of banks' loan losses



The loan loss data was not available for France, Ireland, and Luxembourg.

Figure 6. Performance of a sum of three indicators



Authors' biographies:

Eero Tölö is a senior researcher in the Financial Stability and Statistics Department, Bank of Finland.
Katja Taipalus is head of the Financial Stability and Statistics Department, Bank of Finland.
Matti Virén is Professor of Economic Policy, Department of Economics, Turku School of Economics, University of Turku.
Timo Virtanen is a Ph.D. candidate at the Turku School of Economics, University of Turku.

Author's addresses:

Eero Tölö, Bank of Finland, Snellmaninaukio, PO Box 160, FI-00101 Helsinki, Finland
Katja Taipalus, Bank of Finland, Snellmaninaukio, PO Box 160, FI-00101 Helsinki, Finland
Matti Virén, University of Turku, Turku School of Economics, FI-20500, Finland
Timo Virtanen, University of Turku, Turku School of Economics, FI-20500, Finland