

Taipalus, Katja

Working Paper

Signaling asset price bubbles with time-series methods

Bank of Finland Research Discussion Papers, No. 7/2012

Provided in Cooperation with:

Bank of Finland, Helsinki

Suggested Citation: Taipalus, Katja (2012) : Signaling asset price bubbles with time-series methods, Bank of Finland Research Discussion Papers, No. 7/2012, ISBN 978-952-462-790-0, Bank of Finland, Helsinki,
<https://nbn-resolving.de/urn:nbn:fi:bof-20140807591>

This Version is available at:

<https://hdl.handle.net/10419/212210>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.


Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Katja Taipalus

Signaling asset price bubbles with time-series methods



EUROJÄRJESTELMÄ
EUROSYSTEMET

Bank of Finland Research
Discussion Papers
7 • 2012

Signaling Asset Price Bubbles with Time-Series Methods

Bank of Finland Research
Discussion Papers 7/2012

Katja Taipalus
Financial Markets and Statistics Department

Abstract

This paper provides an early warning indicator for bubbles in financial markets. The indicator is based on traditional unit root tests, more precisely on the augmented Dickey-Fuller test and may be used in a repeated manner with rolling samples. The performance of the indicator is tested extensively via Monte Carlo simulations and comparisons of the results with the most powerful standard (stability) tests. The new indicator seems to be more robust and to have more power than the standard tests. In empirical application to US stock market data for 1871-2010, the new indicator signals most of the consensus bubbles and gives warning signals well ahead of the crash, in most cases as early as 12 months ahead. The indicator also signals most of the 'negative bubbles' before their turning points.

The author would like to thank Matti Viren, Esa Jokivuolle, Jouko Vilmunen, Pentti Saikkonen, Heikki Kauppi and Ari Hyytinen for their comments at various stages of this work. I would also like to thank Nina Björklund and Tarja Yrjölä for research assistance.

Keywords: asset prices, financial crises, bubble, indicator, unit-root

JEL classification numbers: G12, C15, G01

Introduction

Because bubbles can at worst have severe repercussions for the functioning of the financial system and the economy as a whole¹, there has emerged a growing literature on how one can spot bubbles. Though every bubble has its own features, there are some common symptoms that precede them all. One such is the emergence of overconfident expectations of emerging trends. Overconfidence in asset prices means overly positive expectations concerning the duration of rising prices, ie the existence of rational bubbles. Because overconfidence is hard to detect, a constantly diminishing dividend-price ratio can serve as a reference: if price expectations are rising, but higher dividends fail to materialize, the price rise is probably not based on fundamentals. In such case the price can be seen as a composite of fundamental value and a rational bubble component.

A bubble is hazardous for financial- and macro-stability, especially as it can amplify a credit boom by inflating collateral values and causing misallocation of economic resources. In order to cushion the negative macroeconomic effects of a bubble, one needs to detect it early - as soon as prices begin to rise. The warning alarm should be simple and easy to interpret. Unfortunately, traditional stability tests have several limitations due to which they are unable to achieve high accuracy in the case of a periodically collapsing process. One of the major difficulties in using traditional unit root tests is the I(1) dominance, which biases the test results.

This paper presents two modified uses of traditional unit root test parameters to construct two new early warning indicators. New statistical limits are found via extensive Monte Carlo simulations for use in interpreting the signals sent by ADF-regression coefficients and pure AR-regression coefficients. To overcome the problem of I(1) dominance, I make repeated use of tests based on rolling samples. Different window lengths are tested to find the fastest, yet most robust, length of subsample, by which to evaluate the dividend-yield process, which displays constant growth without mean reversion.

Though the novel methods presented here are fairly simple, the Monte Carlo simulations show that the modified indicators have more power than the old testing methods. One major advantage is their ability to react quickly to changes in the underlying data (at best, the bubble alarms were detected after 7 to 9 simulated unit-root observations), yet seldom giving false alarms. In addition, these indicators signals are easy to interpret. The most distinctive feature though is their ability to spot simulated unit root periods from the data, even based on

¹ For the effects, see Bean (2004), Herrera and Perry (2003), Mishkin (2001), Dupor and Conley (2004), von Goetz (2004), Mishkin and White (2003), Kindleberger (2000), Kent and Lowe (1997), Allen and Gale (2000), Filardo (2000), Goodhart (1993), Bernanke and Gertler (1999), Cecchetti et al (2000), Bryan et al (2002), Goodfriend (2003), Lansig (2003a), Mussa (2003), Gilchrist et al (2004), Lansig (2003a) and Lansig (2003b))

relatively modest sample sizes. Concerning the signaling power, both of the novel methods are able to correctly signal about 70-80% of the simulated unit root periods. As shown by the simulations, these percentages compare very favorably with the other tested methods commonly in use.

The real data applications give promising results. In the case of US stock market data from as early as 1871 and covering the period up to September 2010, both of the new indicators are able to spot the major consensus booms and busts from the stock market data. The signals are accurate and arrive sufficiently early (in most cases, as early as 12 months prior to the peak) to have afforded regulators enough time to act.

1. Unit root tests and persistence changes

In economic modelling there is a long tradition of using fixed-parameter autoregressive processes. In recent research growing evidence suggests that the parameters of autoregressive processes fitted to economic time series are not fixed over time, but instead display persistence changes², once or even more frequently. Evidence of persistence changes in the stationarity of processes has also had a big impact on the evolution of unit-root testing procedures.

The traditional tests in the unit root literature, the Dickey- Fuller (DF, 1979, 1981) test and its augmented version (ADF) have both been shown to have severe limitations, especially in the case of changing persistence³.

As a result a number of testing procedures have appeared in the academic literature that are intended to deal with processes displaying persistence changes. Such procedures have been suggested for example by Kim(2000), Kim et al (2002),Busetti and Taylor (2001) and Busetti and Taylor (2004), Leybourne, Kim and Taylor (2004) and Harvey et al (2006) and more recently by Shin et al (2010a , 2010b) as well as by Phillips et al (2011). The procedures in Kim (2000) as well as in Busetti and Taylor (2001) were based on LBI- type stationarity tests rather than traditional unit-root tests. Concerning the methodologies offered as improvements on the conventional DF-testing methodology, one of the first was the procedure presented by Elliot et al (1996) and Elliot (1999), where the methodological improvement was based on detrending: the series was to be detrended before running a DF regression. Detrending was used later, for example, by Taylor (2002). Pantula et al (1994) and Leybourne (1995) used a slightly different approach that relied on OLS detrending. Leybourne et al (2003) used the

² To name a few, Stock and Watson (1996), Garcia and Perron (1996), Kim(2000) and Busetti and Taylor (2004).

³ Busetti and Taylor (2004) showed that the traditional ADF test is not consistent in the case of changing persistence, as the test does not converge to minus infinity with sample size when applied to series containing persistence breaks. This feature is due to the I(1) part's dominance in test results. Similar discrimination problems have been documented in Leybourne-Kim and Taylor (2006).

traditional DF test as a starting point, but explored the power gains achieved by GLS-based detrending of the series. A good summary of those unit root tests, which have significantly more power than the traditional ADF and DF methods, can be found for example in Leybourne et al (2005).

Although the literature presents several methods for dealing with persistence changes, a number of challenges remain as regards their practical applications. One of the foremost challenges is to locate multiple starting and ending points of unit root periods from continuous data. This is an especially difficult problem because the times of occurrence are not known in advance.

A new approach to deal with this problem has already been proposed by Banerjee, Lumsdaine and Stock (BLS) (1990). They treated the break date as unknown a priori, and their statistics were defined on the basis of recursive, rolling and sequential tests. The parameters that formed the basis for the BLS (1992) test were the minimal forward (reverse) recursive unit root test parameters. Leybourne, Kim and Taylor (2006) have later showed that this method did somewhat over-reject to constant $I(0)$ series. The use of subsamples in unit root testing were later analysed further by Taylor (2005), who examined the power of rolling and recursive augmented Dickey-Fuller tests. According to the results, the power of the tests depended heavily on the length of the subsample (window) and the warm-up parameter. Concerning the accuracy of the unit root tests, Choi and Chung (1995) explored the effects of sampling frequencies on the power of traditional unit-root tests (PP, Phillips-Perron and ADF). They found that using high-frequency data significantly improved the finite sample power, for example, of the ADF test. Even more recently, Shin et al (2010a and b) developed two testing methods based on the ADF test, which deal with multiple collapsing episodes within samples using a generalised sup ADF test.

Taking into account the previous research on the subject, the approach of this study is to create a new test setting by using traditional versions of unit root tests in a modified way in order to create a warning signal for an emerging bubble. The aim here is to study whether these simple modified applications of unit root tests can be used as easily applicable indicators of periods of persistence change and therefore as tools for early warning of emerging bubbles. Their reliability is analyzed via Monte Carlo simulations.

The theory underlying the construction of the test for rational bubbles in stock prices has been presented eg in five papers: Campbell, Lo and McKinlay (1997), Campbell and Shiller (1988a and b), Craine (1993) and Koustas and Serletis (2005). The analysis focuses on using dividend-price information, and the rationalization is simple: dividend yields provide a compact measure of how stocks are valued vis-à-vis their fundamentals. Low dividend yields are seen as

evidence of overpriced stocks compared to their earning ability, represented by their dividends (or future dividends), and high dividend yields can be seen as evidence of underpriced stocks. Looking at the dividend yield time-series tells even more: constantly diminishing dividend-price ratios can accordingly be held as a sign of worsening overpricing, ie a bubble, because if prices are constantly rising, these rising expectations should at some point be realized as higher dividends. If price expectations keep rising, but higher dividends fail to materialize, the price rise is not due to fundamentals (ie earning ability). In other words, the price can be seen as a composite of fundamental value plus a rational bubble component, as described eg by Craine (1993): 'rational bubbles satisfy an equilibrium pricing restriction implying that agents expect them to grow fast enough to earn the expected rate of return. The explosive growth causes the stock's price to diverge from its fundamental value'. Luckily, it is easy to locate the point at which the construction of the dividend yield series changes to a unit root (or even explosive) series using time-the series methodology with slight modifications.

In their book, Campbell, Lo and McKinlay (1997) develop a present-value approximating relation so that the traditional asset pricing model can be written in a form in which the log dividend yield should follow be stationary process in normal situations but to have a unit root where there is a bubble in asset prices. The key equation for testing for unit root behavior in dividend yields (derived in detail in Campbell, Lo and McKinlay 1997) is

$$(1) \quad d_t - p_t = -\frac{k}{1-\rho} + E_t \left[\sum_{j=0}^{\infty} \rho^j [-\Delta d_{t+1+j} + r_{t+1+j}] \right].$$

This is the fundamental equation of this paper. Recalling what Craine (1993) and Koustas and Serletis (2005) have pointed out, 'if the dividend growth factor Δd_t and the log of stock returns r_t are stationary stochastic processes, then the log dividend yield, $d_t - p_t$, is a stationary stochastic process under the no-rational-bubble restriction', we conclude that having a unit root in the log dividend yield is consistent with the existence of rational bubbles in stock prices.

2. The new indicators

This study focuses on providing tools to locate boom-bust cycles in asset markets. In econometric terms, such cycles can best be analysed using changing-stationarity models that encompass changes from stationary process to unit root (or even to explosive process, as shown by Phillips et al 2011) and then back to stationary process.

Several problems arise when traditional unit-root tests are applied to a series that contains stationarity changes. As noted above, most of the traditional tests are unable to handle well persistence changes from I(0) to I(1) and back to I(0), as the tests suffer substantial losses of

power in the presence of changing persistence⁴. Another limitation, especially as regards timely warnings, is that the tests are usually applied to long sets of data. This could easily lead to misjudgment as to the true nature of the process, since any I(1) observations within the sample would dominate the rest of the sample.

2.1 Addressing the problem of stationarity change

The solution offered here for avoiding I(1) dominance is to use shorter and rolling samples. These samples would be fixed in length but would update and roll forward one step (observation) at a time, adding one observation to the end of the sample and dropping the first observation from the sample. This sampling procedure keeps the total sample size fixed. In case of a unit root, this procedure removes the unit root from the sample and therefore helps to avoid the I(1) dominance. The idea of using subsamples in unit root testing has just recently gained more attention in the academic literature. For example, Shi et al (2010a and b) and Phillips (2011b) use fixed windows in their SADF test, and moving windows in their GSADF and BSADF tests. But the idea of using subsamples and moving subsamples is not so new. Taylor (2005) examined the rolling and recursive augmented Dickey-Fuller tests and Taipalus (2006a and 2006b) analysed the use of ADF statistics to search for bubbles when using rolling windows to form subsamples. Compared to the present literature (eg Shi et al and Phillips et al) the subsample construction here differs in several aspects. Shi et al (2010a) use a forward recursive ADF test (SADF test) based on an expanding sample size, rather than rolling but fixed size sample, as in this study. In their latter research, Shi et al (2010b) and Phillips et al (2011b) use moving samples, but their samples do not move by one observation at a time going forward, ie they do not "roll". Here the innovation is to roll the sample forward one step at a time, giving each sample its own indication-value, which is then used to evaluate the signal.

Phillips et al (2011a) did use rolling regressions with 77 observations in the sample in their empirical application focused on locating bubbles in the Nasdaq stock index. Interestingly, they report that identification of a bubble appears to be robust over regression schemes, but the estimated collapse seems to be earlier dated in the rolling scheme. This result is in line with the argument presented here, that unit-root dominance occurs sooner in the sample in the rolling scheme. Another point in favor of using rolling samples is the sensitivity of the indicator, which is obvious if one looks at the results of Phillips et al: when they used forward recursive regressions the test ignored the 1987 bubble. When they used the rolling (albeit quite long) window they got a signal during the bubble of 1987. These results clearly argue in favor of using rolling windows to get greater accuracy in the timing of the received signals.

⁴ To name a few references related to this subject: Phillips- Xiao (1989), Stock (1994), Byrne- Perman (2007), Perron (1989,1990) , Banerjee, Lumsdaine and Stock (1990) , Lee and Strazicich (2003), Lumsdaine- Papell (1997), Kapetanios (2005), Saikkonen-Lutkepohl (2002), Lanne et al (2002) and Lanne et al(2003), Elliot et al (1996).

2.2 Construction of the new indicators

The basis for these new indicators is a novel and very simple use of traditional Dickey-Fuller and Augmented Dickey-Fuller tests. Though these tests were largely neglected for a time as regards the methodology for locating bubbles in asset price series, they have received more attention in the recent academic literature. Phillips et al (2011a) present techniques involving recursive implementation of a right-side unit root test and a sup test. These tests are based on ADF t-values. In Shi et al (2010a) a method called the SADF (forward recursive ADF test) is presented, the idea being to implement the right-tail ADF test repeatedly on a forward expanding sample sequence and make inferences based on corresponding ADF statistics. In their later research, Shi et al (2010b) however found that if there are multiple collapsing episodes in the sample, the sup ADF test may not be able to detect the existence of bubbles. This is why they presented a modified version, the generalized sup ADF test (GSADF), which tests for major innovations in the forward moving sample structure. In the generalized sup ADF test, Phillips et al repeatedly implement the right-tail ADF test, but they change the sample sequence by letting the starting point of the sample change over a feasible range and superimpose expanding sample sequences onto each starting point. By using this structure, Phillips et al were able to show, via simulations, a considerable increase in power compared to their earlier version, the sup ADF test. They also presented a detailed proof for the asymptotic distribution for the GSADF.

Even though these applications produced good results, even simpler methods may yield yet more accurate empirical results and be able to signal both positive and negative bubbles. The methods of Shi et al (2010a,b) and Phillips et al (2011a, b) do not seem to be able to locate the negative bubbles at all.

The basic ideas of the two novel indicators offered here are very simple and are rooted in the theory underlying the basic features of the traditional Dickey-Fuller and Augmented Dickey-Fuller tests. The new interpretations are as follows:

Proposition 1. *As is known, the Dickey-Fuller test scrutinizes the possible existence of a unit root in a simple AR(1) model, which (with a unit root) can be written as*

$$(2) \quad y_t = \theta y_{t-1} + e_t ,$$

where y_t is the variable, t denotes time and θ is the coefficient of interest. A unit root is present whenever θ equals 1. The Dickey-Fuller test is based on the first-difference version of this equation, which (in its simplest form, without drift or trend) can be written as

$$(3) \quad \Delta y_t = (\theta - 1)y_{t-1} + e_t$$

The DF tests the unit root hypothesis, $H(0)$, ie whether the coefficient $\gamma = 0$, where $\gamma = (\theta - 1)$. The test is based on the residual term, instead of the raw data. The innovation suggested here is simply to use the AR(1) equation and estimate it over rolling-data samples, such that for each period and each new sample a new value for of the coefficient θ is obtained. An alarm is triggered when the value of the AR parameter is at least 1.0, which signals the presence of a unit root in the sample and thus warns of a possible bubble.

Interpreting an AR value of at least 1.0 as a bubble warning seems justified. For example, Phillips et al (2011a) argue that if bubbles are present, it should be possible to detect explosive tendencies in the price data. As to the interpretation, unit-root or higher values can, in terms of autoregressive behavior, cause such behavior in asset prices as are observed in the markets when bubbles are present.

One further point should be made concerning the coefficient value 1.0 as the limit value. As is known, the least squares regression produces downward biased estimates of coefficients, so that in order to improve the indicator signals, critical values of even less than 1.0 should be used. Concerning modification of the interpretation of the ADF coefficient, the suggested modification follows:

Proposition 2. *The augmented Dickey-Fuller test is merely an extension of the original D-F test, to include lags in the autoregressive process. The testing procedure is the same, except that the ADF model now takes the form*

$$(4) \quad \Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta y_{t-i} + e_t ,$$

where α and βt are the deterministic components (constant and linear time trend), y is the described variable and e_t is the error term, expected to be identically and independently distributed. As both of the deterministic components are restricted to zero, the process becomes a pure random walk. In the case of ADF, the main interest focuses on the value of the coefficient γ . In the conventional form, the test is run to see whether this coefficient takes the value zero, consistent with the existence of a unit root; the alternative hypothesis is $\gamma < 0$. In the traditional testing procedure, the coefficient values per se are not examined; instead, they are used to calculate the test statistics (t -values), which are then compared with the critical values. This is also the way in which the more recent applications of right-side ADF tests are performed. The modification suggested here is extremely simple: instead of using the t -values, one would use the coefficient values as such. The coefficients would be interpreted as signaling a unit root whenever the value of γ is at least zero. The regression is run over each period separately, using a rolling window of subsamples of a fixed size. New updated indicator values are obtained for each period,

as the sample rolls forward by one step (observation) at a time⁵. The subsample over which the regression is run, the ϑ_t , is defined so that:

$$(5) \quad \vartheta_t = y_t \dots y_{[(t+\pi)-1]},$$

where π denotes window length (36,48,60). As time advances by one period (t to $t+1$), so do the starting and ending points of the sample. (The sample rolls forward in such a way until the end-period reaches final period T).

Concerning the recent test-methods presented by Shi et al (2010a,b) and Phillips et al (2011a, b), there are major underlying differences between these methods and this indicator. The first difference relates to the use of a different construction of subsamples and to the ability of this method to provide period-by-period updates of the indicator. Another important difference lies in the use of actual coefficients in building the indicators presented here, as opposed the Shi et al (2010a,b) and Phillips et al (2011a, b) who focus on ADF t-values. Another difference surfaces in actual data applications: Shi et al (2010a,b) and Phillips et al (2011a) use only (real) stock price data, whereas here the methods are applied to dividend/yield data.

2.3 Selecting the most informative window length

A key question relates to the choice of window length used in the rolling regressions. Concerning the properties and functionality of the indicator, the rolling windows should be wide enough to support the Gaussian approximation when defining the regression parameters. On the other hand, concerning the core features of the phenomenon under study, which this indicator is meant to capture (asset price booms and busts), the subsample data should not be too long, because booms often last only a few years. In order to provide an early warning indicator, able to provide reliable and timely signals, the rolling data windows are limited to lengths of 36 , 48 and 60 observations or eg 3, 4 or 5 years of monthly data. In several other studies in which subsamples have been used, the samples have been much longer and therefore also less amenable to the updating of information (here, unit root values) from the sample. This could be problematic, as reactions to such an indicator would be slow.

⁵ Because the subsamples are overlapping, we could encounter a problem of correlation. To determine the seriousness of the situation, I compared, in MC simulations using similar datasets, the behavior of critical values in cases where the samples were overlapping versus simple one-off tests. The critical values simulated were quite similar, suggesting that, even though correlation might distort the results slightly, the overall impact should be relatively small.

3. Numerical results

It is well documented, that the link between asset price bubbles and financial stability can lead to highly adverse outcomes. One of the newest challenges confronting central banks, financial supervisors and regulators is to minimize systemic crises and their costs to the macroeconomy. The scope of monitoring required in order to promote macro stability is indeed wide and so the required toolkit is also expansive. A major theme in the literature on early-warning tools is the need to develop an alarm system for a heightened probability of emerging bubbles in asset valuations. For meeting this challenge, a viable early warning tool should have several important qualifications. Foremost, it should have good statistical power: it should give as few erroneous signals as possible but should still spot the majority of bubble observations. The type I and type II errors should be well balanced. The method should also identify emerging bubbles early enough to enable regulators to react. Further, the method should be robust. This means that, if bubble observations keep appearing in the data period after period, the method should signal bubbles repeatedly. Moreover, the indicator should be able to signal unit roots even where the persistence of a stationary process is already close to 1 (ie 0.9). This feature is especially important from the practical viewpoint: during normal market periods, the AR(1) regression coefficients in financial series are usually already close to 1.

3.1 Framework for Monte Carlo simulations

Monte Carlo simulations are used to explore the power and accuracy of the indicators introduced in section 2. The aim is to search for optimal length of the rolling window used to develop the new indicator. In addition, we look at how well the indicators perform when the underlying stationary part itself has a long memory, as is usually the case for financial market series.

The data observations for the simulations were generated by a program in STATA. The creation of observations is based on an AR(1) process for which the initial value is generated by a random seed. The OLS regression is kept simple (no trend or constant, iid error terms). For each analysis a set of 1100 observations were replicated 5000 times, making the number observations in each run 1100×5000 , and in each series the first 100 observations were omitted to avoid initialization effects.

Each of the 1100 series of simulated observations includes two breaks: from stationary period to unit root process and then back to stationary process. The break always occurs around the middle of the sample, since the first observation including a unit root is always observation no. 500. The first observation of the unit root is tied to the last stationary observation (traditional AR process with lag 1) to avoid sudden breaks in the process, which could invalidate the test results.

The last observation including a unit root is dependent on the length of the simulated unit root, for which there are three options. The length of the simulated unit root process is 36, 48 or 60 observations, so that the last observation including a unit root is observation no. 536, 548 or 560. The simulated unit roots were compared to real data bubbles appearing in the Shiller data. The simulated processes seem to be slightly more volatile than the real data bubbles⁶. A more volatile series means that the changes in stationarity should not be easier to spot in the simulated series and the ability of the method to signal stationarity changes is not likely to be overestimated.

To provide timely warning signals to policymakers, the indicator should be fairly quick to flash warnings of developing misalignments. In the simulations, the AR- and ADF- based indicators should therefore start to launch warning signals shortly after the start (observation no. 500) of the simulated unit root period. This can be examined by looking at the shifts in the AR- and ADF- distributions. Shortly after observation no. 500, the coefficient distribution should start to shift towards 0 for the ADF and towards 1 for the AR. Concerning the length of the bubble and length of the rolling window, the shortest possible combination was chosen (36 periods for both). This is because if the signals emerge from short periods and narrow windows, they should work even better for longer periods and wider windows.

Clearly, the distributions in both cases show that the coefficient values start to change quite quickly to the right, towards the limits of 0 and 1 as the first unit-root observations are taken into the sample. After five observations, a definite change is already discernable in both distributions (see figures in Appendix I) and already after 15 unit-root observations, there is a highly visible shift in the distributions. During the final phase, when all unit root-observations are included in the sample, the coefficient-distribution has already clearly shifted to the right.

As was seen, the first simulated unit-root observations entering the rolling window clearly starts to shift the distribution. A similar situation obtains when the unit-root period ends and stationary observations begin to enter the sample, ie the distribution starts to shift back. After all the unit-root observations have left the sample, the distribution returns to the form and place where it started. This is visible in the figures in Appendix I. *This feature is a very essential piece of information concerning the clear advantage of using rolling samples in running the indicator regressions. In rolling samples the I(1) process does not continually dominate the samples; instead, the narrow rolling windows are quickly updated to bring new information into the samples, whether from I(0) to I(1) or the reverse.* An enlargement of the memory of the stationary part does not affect the results since the coefficient distributions continue to react rather quickly to the start and end of the simulated unit-root period. This is a positive sign for the reliability of the indicator. The impact of the memory length of the stationary part on the indicator's ability to react can be illustrated by

⁶ These bubbles are based on Shiller's data and seem to be relatively smooth for traditional stock market data. It is not clear whether the data were smoothened.

showing how the 5th percentile, the average, the 50th and the 95th percentiles of the coefficient distributions shift during the simulated bubble. This is illustrated also in the figures in Appendix 1. In connection with the shape of the distribution of ADF coefficients and changes a number of questions arise relating to the critical value. This becomes obvious where the sample includes only unit-root observations. In this case the majority of ADF-coefficient values should already be around zero (indicating the existence of a unit root). From the figures in Appendix 1 we see that the majority of observations are less than zero, rather than in the neighborhood. This suggests that, because of the distributional features, use of zero as the critical level might not be efficient, since it might result in too few alarms of unit-root processes. This may relate to the well known fact that least squares regression produces downward biased coefficient estimates in the first order autoregression⁷. One way to decide on an alternative critical value would be to use Monte Carlo simulation to find new critical limits for the coefficients. The table in appendix II reports the 5% upper tail critical values for different window lengths. All the distributions are based on 5000 replications of datasets of 1000 observations.

The underlying process memory would be expected to be rather long in the real world during "normal" times. It is therefore reasonable to choose critical values based on simulations in which the stationary process is 0.9. As all these 5% critical values lie around the value -0.05 (see appendix II) that is what we use here as the additional critical value for the warning signals, rather than zero.

3.2 The power and accuracy of the AR- and ADF- signals

The performance of the ADF- and AR- based indicators are evaluated on the basis of three ratios that measure indicator performance in terms of the type I and type II errors. The indicators' sensitivity to changes in stationary-part persistence is studied in terms of the loss of power as the coefficient of the stationary process is increased from 0.6 to 0.9. A good indicator would not lose power even if the stationary process is highly persistent. Based on the simulations, we examine the sensitivity of the indicators to the lengths of simulated unit root in the data and rolling window, which are the bases for the indicator coefficient regressions. *Because the bubbles may vary in length, it is important for an indicator to be able to give warnings even when the bubble is of relatively short duration.*

To evaluate the indicators, we studied the **number of false alarms given by the indicators, ie the number of signals of unit-roots triggered by the indicators in periods in which there were no unit-roots present.** The percentages of the false signals for the AR- and ADF- indicators are detailed in the tables in Appendix III. The total number of false alarms seems to be very small,

⁷ As mentioned earlier, in case of the AR-indicator this would therefore mean, that the correct level to launch warning signals could actually be slightly under 1 instead of being precisely 1 as interpreted here.

even where the stationary-period memory is long (0.9). The percentages of false alarms before and after a unit-root period can be analyzed separately. According to these results, the probability of false alarms increases slightly after a simulated unit-root, but remains very small. For example, when the simulated unit-root period length is 36 observations, the stationary period is simulated using relatively long memory (0.8) and the rolling subsample over which the ADF- and AR-regressions are counted is 36 observations. Here, the total number of false alarms is only 2782 in more 4.5 million observations, ie the false-alarm rate is just 0.06%.

As observed from the results shown in Appendix III, the AR coefficient clearly produces the smallest number of false alarms. The difference in numbers of false alarms as between the AR and ADF indicators is even greater where the persistence of the stationary process is greater. For the ADF method, there is a clear difference between critical values (0 or -0.05) in terms of false alarms: using zero as the critical value clearly results in fewer false alarms than does -0.05. Overall, the number of false alarms for both methods seems to depend on the stationarity of the "normal" period: as the stationary-period persistence increases (from 0.6 to 0.8 to 0.9), the probability of false alarms increases for both methods. Increasing the length of the rolling window reduces the probability of a false alarm, especially for AR: the longer the rolling window, the less probable is a false alarm. In terms of indicator-specific features, lengthening the unit-root period does not seem to have a great impact on the number of false alarms given by the AR coefficient, in contrast to the ADF coefficient, for which the number of false alarms increases as the simulated unit root period gets longer. This can be explained by unit-root characteristics and by examining false alarms separately for the periods before and after a simulated unit root period. We know that a unit root will dominate a sample. In this regard, the AR coefficient seems to be more robust: the false alarms are clearly more frequent after long simulated unit-root periods for the ADF coefficient versus the AR coefficient.

Another important means of assessing indicator performance is to look at **the total number of simulated unit roots that the indicators are able to spot in the data**. This assessment can be done in two ways. First, the performance can be evaluated according to the indicator's *ability to mark correctly individual unit-root observations* in the simulated data. For example, in the case of 36 simulated unit root observations in the data, perfect performance means the indicator signals every one of these observations as a unit-root. Another performance metric (perhaps more relevant for practical applications) is to analyze *how many of the simulated unit-root periods are spotted* and signaled by the indicators. Here, it is sufficient that just one of the 36 simulated unit-root observations is signaled during the simulated unit root period. If none of the 36 observations is signaled, the indicator would be judged to have missed the unit-root period. Even a single signal would indicate the spotting of a unit-root period from the data. The AR and ADF indicators' ability to signal correctly individual unit roots or their periods is presented in the tables in Appendix III.

In terms of correctly signaling single simulated unit-root observations, it is surprising that both methods perform the more accurately, the greater the persistence of the stationary process. The ADF coefficient produces quite different results depending on which critical value (0 or -0.05) is used. When the stationary process was simulated using the coefficient 0.6 and the rolling window as long as the simulated unit root period (36 observations), use of the zero critical value (ADF(0)) results in correct signals for only 5.32% of the single unit roots. In contrast, when -0.05 is the critical value, the ADF(-0,05) correctly signals 15.75% of the simulated single-unit-root observations. This is close to the level for the AR coefficient, which correspondingly correctly signals 16.20% of the single simulated unit roots. The performance of the indicators improves in most cases when either the rolling window or the simulated unit-root period is lengthened. For example, when the stationary period coefficient was set as 0.6 and the length of the rolling window was increased to 60 observations and the simulated unit-root period to 60 observations, the ADF (-0,05) was able to signal 23.39% of the single simulated unit-root observations correctly. This last result is not surprising since a longer unit-root period is always easier to extract from the data. If the previous example is changed so that the rolling window is reduced to 36 observations, other elements being the same, the ADF (-0,05) correctly signals 17.50% of the single-unit-root observations and the AR method 23.14%.

These indicators seem to retain power as the stationary-period simulation parameters get longer memories, from 0.6 to 0.8 to 0.9. When the stationary period was simulated using the coefficient 0.9, and the unit root period and the rolling window length were set at 36 observations, the ADF (0) was able to signal correctly 4.95%, whereas the ADF(-0,05) correctly identified 17.28% of the single-unit-root observations. The AR coefficient did even better, correctly identifying 20.76% of such observations.

When indicator performance was evaluated according to **ability to identify unit root periods instead of single observations**, the results were quite different. The percentages for each method are much higher when the focus shifts to finding periods instead of single observations. Concerning the ADF method, in a simulation framework where the stationary period was simulated using a shorter memory (0.6) and the unit root and rolling window were set at 36 observations, the ADF (0) signaled 37.54% of the simulated unit-root periods, whereas the ADF(-0,05) was able to indentify 59.70%. The rates were much higher than for the single observations. In the same setting, the AR method correctly signaled 55% of the unit-root periods.

Once again, the longer the simulated unit root period, the more easily it is identified from the data. In many cases as many as over 70% of the simulated unit-root periods were identified. For example, when the unit root period was set at 60 observations (other settings being the same), the ADF(-0,05) was able to correctly signal 79.04% of the simulated unit root periods. And in another

option, when the stationary periods memory was increased (to 0.9), the rolling window being constant at 36 observations and the unit root period set at 60 periods, the ADF(-0,05) was able to correctly signal 81.36% of the unit-root periods. Simulation results and rejection frequencies along with correct signals are detailed in the tables in Appendix III.

Summing up the core results, it seems that the optimal length of rolling window is quite short, as the shorter rolling windows clearly performed best, whether the method was the ADF or AR coefficient. Concerning the methods, the most robust and precise indicators seem to be the rolling ADF coefficient, with -0.05 as the critical value and the AR regression coefficient with 1.0 as the critical value. Even though the correct signals of individual unit-root periods remained at a rather modest level, they do not compare poorly to other such methods of signaling unit root observations from data. The most important finding though is the huge accuracy improvement as measured by the number of correct unit-root-period signals as opposed to finding individual unit-root observations. As both of the methods are able to signal up to 70-80% of the periods correctly, they are worthy of further study, especially as regards real data applications and the ability to produce early warning signals.

3.3 Are continuous-signal methods more accurate?

A consistent indicator would produce continuous warning signals during a simulated unit-root period. It is therefore important to investigate how the accuracy of the AR and ADF methods change if the bubble signal is given only after the unit-root indicator has identified five (or more) consecutive single observations as bubbles. The evaluation was done with more limited data, just to get an idea of how the criterion of continuous warning signals affects the results.

The analysis is based on MC simulations, where the total number of observations is limited to 100000 and the stationary period's simulation parameters vary from 0.6 to 0.9 (as previously, only one of the coefficients is used at a time). To evaluate the ADF coefficient's performance, the critical limit was set at -0.05 and for the AR coefficient the critical value was 1.

The simulation results are reassuring: It seems that the methods' sensitivity to changes in stationary-period coefficient is greatly reduced. The results and accuracies are much more alike as between the methods in the continuous-signals case than in the case of single observations: the coefficient (whether 0.6 or 0.9 during the stationary part of the regression) does not play such an important role in the case of multiple signals.

Concerning the number of false alarms, we note that the use of multiple (five or more) alarms further reduces the total number of false alarms. The total number of correctly signaled unit-root periods (meaning that at least one set of five continuous alarms is triggered during the simulated

unit root period) is still around the 35% level, but the total number of correctly signaled single observations (meaning that each of the sets of five observations during the simulated unit-root period was correctly signaled as a bubble) is relatively low, approximately 10%. This can be seen from tables in Appendix IV, which includes the core results of the multiple alarm test.

The results of the continuous alarms can be compared to those of single alarms. The main difference between the use of single signals and multiple continuous signals is that the multiple continuous signals would reduce the total amount of false alarms but would reduce the total number of signaled bubble periods. In addition, the use of multiple continuous alarms makes both methods (AR and ADF coefficients) more robust to changes towards 'normal' period parameter changes. Therefore *it is seems prudent to use both methods - single and multiple alarms - in evaluating developments in a time-series, since it is found that the robustness of the alarm signal is greatly increased when five or more continuous bubble signals are received.* An especially important point is the small number of false alarms in these cases.

4. The power of conventional unit root tests in the case of rolling windows

To get an idea of whether the AR and ADF indicators really provide an improvement to the already existing group of stationarity shift and unit root indicators, one might well compare the performance of the indicators to the conventional, and the so-far most powerful ones: unit root and stability tests.

Evaluation of performance is done by comparing the results of Monte Carlo simulations. The simulation setting provides a full replication of that of the previous section. First, each of the conventional tests was run using the same simulated series as the AR and ADF analyses. This is to avoid any differences resulting from newly simulated data. Each method was tested using similar lengths of rolling windows (36, 48 and 60) to define values for the test parameters. The dataset was again 1000*5000 observations for each test, since the first 100 observations were omitted to avoid initialization effects. The only exceptions were the rolling CUSUM and rolling variance ratio tests, where the total sample was limited to 100 000, due to the core features of these tests, which are much more data-intensive and time consuming compared to the conventional unit root tests. In the case of rolling variance ratio, the sample was even smaller, being limited to 10 000 observations, once again due to the time consuming features.

The simulated breaks in the data are similar to those in previous section: each of the simulated datasets includes two breaks - from stationary period to unit root process and back to stationary process. And as before, the break is always situated nearly in the middle of the sample. The chosen conventional tests were the R-test, MAX-test, CUSUM-test and the variance ratio test.

The R-test was presented by Leybourne, Kim and Taylor (LKT, 2006). Following Banerjee et al (1992), Leybourne, Kim and Taylor (2006) demonstrated that the forward recursions could be used for testing against persistence change from I(0) to I(1), and the reverse time series recursions can be used to test against change from I(1) to I(0). Forward and recursive tests, however, cannot adequately discriminate between change in persistence and constant I(0) behavior, as mentioned in Leybourne, Kim and Taylor (2006), this being the reason they proposed a new test statistic based on the ratio of forward to reverse stats (the R) for use in unit root testing. The clear advantage of the R-test is that it remains consistent over changes from I(0) to I(1) and vice versa.

The t-test values associated with the forward and reverse recursion coefficients are the $DF^f(\tau)$ for the forward-and the $DF^r(\tau)$ for the reversed series t-tests, τ being the true break fraction. As the precise date of the change in the series persistence is usually unknown, Leybourne, Kim and Taylor proposed using the minimum of the sequence of t-stats over a set of subsamples (subsamples were constructed through using various break fractions). These minimum values were denoted by $DF^{f \text{ inf}} \equiv \inf_{\tau \in \Lambda} DF^f(\tau)$ and $DF^{r \text{ inf}} \equiv \inf_{\tau \in \Lambda} DF^r(\tau)$. The proposed R- statistic is constructed as the ratio

$$(6) \quad R \equiv \left| \frac{DF^{f \text{ inf}}}{DF^{r \text{ inf}}} \right|$$

Use of the minimum over a sequence of changing subsamples is problematic in this case. One of the main innovations here is to use rolling windows to define the subsamples that are always fixed in length. This is why I chose to calculate the R- statistic over subsamples of the same length and the selecting the minimum value in each sample.

For the second conventional test, the MAX-test, the starting point in Leybournes (1995) was to explore whether the unit root tests would improve in power if the conventional Dickey Fuller and Augmented Dickey Fuller tests were run not only in forward recursions but also with the reverse realizations of the data. The changes in definition of the new test-parameter was quite modest, since the MAX-test is transformed into a maximum of the t-test values of forward and reversed recursions over the chosen data sample.

Formally, the MAX- test by Leybourne (1995) can be simply defined using the standard Dickey-Fuller (DF) procedure. The forward regression t- value is denoted by DF_f and the t-value for the reversed series is denoted by DF_r . The MAX- statistic then being expressed as

$$(7) \quad \text{MAX} = \max(DF_f, DF_r)$$

and for the ADF,

$$(8) \quad \text{MAX} = \max(\text{ADF}_f, \text{ADF}_r)$$

The other two chosen tests, the variance ratio test and CUSUM-test use a slightly different testing approach. The focus in the Lo and MacKinlay variance ratio test is to analyze the random walk patterns in a time series. To put it short, in a random walk series the variance of a sample is linearly related to the length of the sampling interval. When a time series is split into n equal parts, the variance of the whole finite time series should be n times the variance of the first part, assuming the random walk hypothesis. In other words, the variance ratio for all the n sample parts should equal one. If the variance ratio stays under one, the series is mean reverting, ie. the series has a short memory and must include some negative correlation. When the variance ratio is greater than one, the series is persistent, meaning that the series has a long memory and positive serial correlation. Existence of a unit root in the series therefore indicates that the random walk hypothesis holds.

The variance ratio test used here, can be described as follows: The test-values are defined by calculating the variance ratio by applying the Stata module *lomackinlay* to predefined subsample data. The major problem may be the likelihood of heteroskedasticity, as it is known that $z(1)$ statistics may not have the usual asymptotic properties in case when the variance of innovations is unstable.

$$(9) \text{ Variance ratio, homoskedastic errors: } z(1)=z_1(q) = \sqrt{nq} \overline{M}_r(q) * \left[\frac{2(2q-1)(q-1)}{3q} \right]^{-1/2}, N(0,1)$$

where the number of periods q , over the which innovation parameter's effects on the values of the variable are screened. If the process is stationary, the innovation parameter should not have permanent effects, ie it should converge towards 0. In addition, in previous equations, nq is the sample size (n being the multiple for sampling frequency) and $\overline{M}_r(q) =$ the dimensionless centered variance ratio.

The null hypothesis in this test will be RMH, which can be interpreted to mean that the underlying series has a unit root. Rejection of the hypothesis is accomplished by applying the critical values presented in Lo-MacKinlay (1989)⁸.

Finally, regarding the CUSUM-test, the purpose here is to provide a completely different approach to testing for the existence of breaks. CUSUM is an old method that has been used mainly as a

⁸ There are quite a few articles where the Lo-MacKinlay variance ratio test is applied to financial series (eg Whang-Kim (2003), Ayadi-Pyun (1994), Hoque-Kim-Pyun (2007), Ajayi-Karembers (1996)), but concerning unit root testing, the rolling application of the tests has been rare.

statistical process control tool, as originally designed by Paige (1954). The underlying idea in CUSUM is to detect persistent changes or shifts in the underlying process. In the traditional CUSUM analysis, there are three important values: the center line, which represents the target value, the upper control limit and the lower control limit. If the process is in control, it should stay between these two limits. Observations outside of the borders signal changes in the underlying process. Very large shifts result many observations outside the limits.

Li (2007) sees CUSUM as being among the most effective procedures for detecting small shifts in the mean process. For this reason it should be able to spot changing persistence also in series stationarity. CUSUM and its modified version's ability to signal changes in the time series persistence have been analyzed for example by Leybourne-Taylor and Kim (2006b), whose CUSUM-based tests have the big advantage of generally not spuriously over-rejecting a process that does not display a change in persistence. The CUSUM test used here is the traditional CUSUM test, with the innovation that it is calculated by using rolling subsamples of data. In traditional CUSUM-testing, an alarm means that one should return and 'nullify' the process before continuing. One advantage in using rolling tests instead is argued to be that the distortion caused by an alarm should be reduced due to the effect of overlapping 'clean' samples.

The formulation of the CUSUM test here follows closely Mellin (2009). The traditional CUSUM⁹ test can be defined formally as follows. Let there be n observations of variable x_i , $i = 1, \dots, n$ with the midvalue ω_0 . Then the cumulative sum over n observations can be defined as

$$(10) \quad C_i = \sum_{j=1}^i (x_j - \omega_0), i = 1, \dots, n.$$

The cumulative sum C can be defined so as to collect all the deviations exceeding the reference parameter. In the basic formulation, CUSUM is only able to trace the positive deviations. In a more advanced approach it is possible to track both positive and negative deviations. There, in addition to C^+ , which collects all the positive deviations, it is possible to define C^- such that it collects all the deviations falling under the reference parameter. Calculation of cumulative sum over a certain set of variables $X = (x_1, \dots, x_n)$, with the expectation that the variables in the sample are normally distributed with parameters $(X) = \omega_0$, $D(X) = \sigma$, can be written as follows:

starting values:

$$(11) \quad C_0^+ = 0 \text{ and } C_0^- = 0$$

for $i = 1, \dots, n$, the values are defined as:

$$C_i^+ = \max[0, x_i - (\omega_0 + K) + C_{i-1}^+]$$

$$C_i^- = \max[0, (\omega_0 + K) - x_i + C_{i-1}^-]$$

⁹ For further details on the change detection parameters and CUSUM-procedures, see eg Part I : Changes in the Scalar Parameter of an Independent Sequence, <http://citeseerx.ist.psu.edu>

where ω denotes the target-value and K the reference value usually chosen to be situated halfway between the target-value and the value toward which the change in the process is hopefully leading. In the CUSUM-measure, the C_i^+ and C_i^- are presented as lines for values $i = 1, \dots, n$, and they are expected to stay between the minimum and maximum control-borders. The limiting borders ie. the minimum and maximum control-borders for the cumulative sum-values, are dependent on the variance of the process and are defined in the basic model as the $+H$ and $-H$, borders such that $H=5\sigma$.

In this study the CUSUM is calculated by using the Stata-module `cusum6`, which calculates the recursive residuals from a time series regression in order to generate the CUSUM as well as the CUSUM squared tests of structural stability, which is more thoroughly presented in Brown-Durbin and Evans (1975). The approach by Brown-Durbin and Evans has been the basis for numerous pieces of academic research.

4.1 Monte Carlo simulations

4.1.1 The power of the rolling R- and DF^{inf} - tests

For the R- and DF^{inf} - tests the interpretation of results must be done carefully. Since the R- test statistic can be constructed only after defining the DF^{inf} test statistic, it is useful to analyze also the DF^{inf} test for the case of rolling windows, especially since Leybourne et al (2006) find that when one wants to test whether the series is characterized by unit root (ie $H(0)$ of constant $I(1)$ behavior against the alternative $I(0)$), they recommend using the DF^{inf} and $DF^{r inf}$ tests instead of the R- test.

The R- test and F^{inf} - test parameters are calculated by using rolling windows of lengths 36, 48 and 60. The critical values to evaluate the signals are again taken directly from Leybourne et al (2006), where the 5% critical level limits can be found for both tests for as small a sample size as 60. These critical values are used for estimates to evaluate the R- and DF^{inf} - test statistics, though once again there might be a problem with the use of rolling windows instead of static samples for which these limits were originally created. As Shi et al (2010b) showed, the asymptotic behavior depends largely on the subsample size, distributions of smaller samples being leptokurtic.

In the case of the R-test, the major interest is to find out whether the test rejects the null hypothesis of constant persistence against the alternative of a change in persistence. In the case of rejection, attention is drawn to the tail of rejection, which may indicate the direction of the change, from $I(0)$ to $I(1)$ or from $I(1)$ to $I(0)$.

In the DF^{inf} - tests the major interest is to examine how many times the rolling test correctly rejects the null hypothesis of $I(1)$ during the stationary period, either before or after the simulated

bubble and how many times it falsely rejects this hypotheses during the simulated unit root period. The power and accuracy of the DF^{inf} - test are reported in the tables in Appendix V.

Concerning the power of the test for different memory lengths in stationary period, the DF^{inf} performs clearly better when the stationary period is simulated with shorter persistence, 0.6. In this case it is able to correctly reject the unit root hypothesis in 52% of the cases, even with the a very short window (36). As the window gets longer, the percentage of correct rejections increases. When the window includes 60 observations and the stationary period persistence is still 0.6, the test makes correct rejections in 88% of the cases.

The correct signals seem to be sensitive to changes in stationary-period persistence. This can be seen by comparing the results in the tables in Appendix V. As the regression coefficient for the stationary period increases in size (to 0.8), the number of correct rejections of the unit root hypothesis decreases sharply. For the shortest window (36), the method correctly rejects the null hypothesis in just slightly over 20% of the cases, and although increasing the length of the window improves the results somewhat, the method correctly rejects in less than 40% of the cases. The DF^{inf} method seems to perform much better in environments of shorter memory.

On the other hand, during the unit-root simulation, false rejections of the unit root null hypothesis for single observations are quite rare when the persistence during the stationary period is high, 0.8. For the shortest window (36), false rejections occur only in slightly over 16% of the cases. Increasing the window length does not seem to improve the accuracy, as the number of false rejections increases as the window is extended. The method's sensitivity to changes in stationary-period coefficient becomes visible once again when the false rejections during simulated unit root periods are compared for the series where the stationary period was simulated with shorter (0.6) versus longer (0.8) memory (see tables in Appendix V). False rejections are quite numerous when the stationary-period coefficient is 0.6.

Finally, the unit-root periods (rather than single unit root observations) that were rejected falsely during simulated unit root periods are presented in the tables in Appendix V. From these tables we see that the test correctly identifies more periods when the stationary part has short persistence. When the stationary-period coefficient is 0.6, much fewer unit root periods are falsely rejected.

The R-test environment is slightly different. The null hypothesis here is a constant I(0)-period. Rejection of the hypothesis would be interpreted as a signal of persistence change in the process and, as mentioned in Leybourne et al (2006), when the rejection occurs in the upper tail, this suggests a change in persistence from I(0) to I(1). As the main interest in this study is to find methods that can reliably signal shifts from I(0) to I(1), the focus will be to examine how many upper tail rejections the R- test is able to produce for simulated unit roots when the data are run in rolling form. Critical values for the evaluation are from Leybourne et al (2006).

The core results are shown in tables in Appendix V. The R tests rarely gives false alarms of unit-roots during the stationary period, even when the stationary period is simulated by using higher persistence, ie when the coefficient is 0.8. On the other hand, the number of false alarms increases somewhat after the simulated unit root period, compared to the period before the simulated unit root.

Though it rarely gives false alarms, the method unfortunately is unable to signal unit-roots correctly. The test misses nearly all of the simulated bubbles. It seems that it is too rigid to react to relatively quick changes in process persistence. This can be seen from the results; the longer simulated unit roots are signaled much more often than the short ones. It also seems that the method finds breaks easier from the data where the underlying stationary process is already close to a unit root. This feature also confirms that the test is relatively rigid. If the method is very slow and rigid, this could mean that in short rolling samples and in short unit-root periods, the critical limits should be calibrated from a much narrower distribution. The results concerning the asymptotic behavior for different subsample sizes by Shi et al (2010b) support this interpretation.

Due to the problem of missing nearly all of the simulated unit-roots, I decided to experiment in order to find out whether the problem was the critical values. I simulated a new 5% critical value for the test and used it as a new critical level for the R-test, where the underlying data included one simulated unit-root period (48 observations), the rolling sample was also 48 observations, and the stationary period was created using of coefficient 0.8. After these changes, the results also changed: even though the number of false alarms increased, so did the number of signaled breakpoints. Though the method was able to spot only 2% of the single unit-root observations, after the changes in critical value it was able to signal approximately 10% of the unit root periods (compared to 0.6%). This clearly indicates that when the R- method is as an indicator in the case of short rolling samples, the critical values need to be calibrated and redefined.

4.1.2 The power of the MAX- test

In the case of MAX-test, the test setting is most similar to the conventional Dickey-Fuller test. The $H(0)$ hypothesis is $I(1)$, and the final test-statistic for evaluation purposes is the maximum of the forward and reverse realizations. The critical values for the hypothesis testing are from Leybourne (1995), which includes tables for 10%, 5% and 1% critical levels for sample sizes as small as 25 and 50. Of course, the use of critical limits from earlier research can be problematic especially since the samples here are rolling instead of static. Since no more appropriate limits were available, these must suffice as rough estimates. This is why the final results must be interpreted with caution.

Concerning the simulation results, the rejection of the null hypothesis $H(0) = I(1)$ is difficult here. Therefore, though the MAX-test is able to signal most of the simulated unit-root observations in the data, it is unable to reject the unit root hypothesis in many cases during stationary periods.

Another problem seems to be that the amount of false alarms increases sharply when the stationary period is simulated using higher persistence. When the stationary period was simulated with persistence of 0.8 instead of 0.6, the rejection of the null became even more difficult. The core results can be seen in tables in Appendix VI, where for a coefficient 0.8 the amount of false alarms reaches a fairly high level.

This rejection problem is nothing new. In case of the conventional Augmented Dickey-Fuller test with t-values, the results remind us in a sense of the MAX- results (ADF t-test results where the stationary period was simulated using the coefficient 0.6 are presented for reference in Appendix VII). There is, however, an important difference between the ADF t-test and the MAX test. The MAX-test seems to be much more accurate. This result is congruent with the results reported by Leybourne (1995).

An interesting feature of the MAX test in the case of rolling windows is that the amount of false alarms falls quite sharply when the window length is increased, ie more data are included in the sample (see tables in Appendix VI). This clearly means that the MAX-method works better with longer samples, but not so well with very modest sample sizes. Another core feature is that the simulated unit root clearly dominates the samples for a long time, even after the break from I(1) back to I(0). This feature explains why there are more false alarms after, versus before, the simulated unit-root period (see tables in Appendix VI).

When these features are compared with the rolling AR- and ADF- results, it clearly seems that, even though the ADF- and AR-methods do not correctly signal as many unit roots as the MAX-method, they are better as early warning indicators in two important respects. Firstly, they are more robust to persistence changes in the stationary period since they give far fewer false alarms, even when the stationary period is simulated using as high a persistence as 0.8. Secondly, the differences in accuracy between small- and larger sample results are modest. Therefore we endorse the use of rolling AR- and ADF-coefficients in the case of small sample size.

4.1.3 The power of the rolling CUSUM-test

For the rolling CUSUM-tests, the test-procedure is quite different than the other tests presented above. In all of the previous tests a single test-value was created. When the rolling CUSUM is constructed, this feature is impossible. The rolling sample of CUSUM-test consists of 36, 48 or 60 single data observations, but all of the sample observations are valued separately, since a structural-break alarm is set off if any single observations lies outside the upper or lower bound. One option is to create an indicator that takes the value one if any single observation in a sample breaches the upper or lower bounds. I decided that searching all observations separately would be more informative, since then it is possible to see how many single alarms are triggered in each window.

Concerning the interpretation of the results it must be kept in mind that the CUSUM test is primarily a stability test. It should trigger an alarm whenever the construction of the process changes. Therefore the major interest is to examine how this method reacts to the start and end of a simulated unit root period. In addition, we want to know how many false alarms of structural breaks it gives outside of the simulated unit root periods, ie during stationary periods. The results are shown in the tables in appendix VIII. In addition to these one can construct a graphical example of how the rolling CUSUM-values react to the simulated unit root where the stationary-period coefficient is 0.6, the bubble is 36 periods long and the rolling window consists of 36 observations. In graphs C1 to C4 in Appendix VIII, the first figure illustrates the situation when the sample does not include any unit root observations. Figure C2 shows how the situation changes as 5 unit root observations are included in the sample, figure C3 shows the situation after 25 observations, and figure C4 shows the reaction when all unit root observations are included in the rolling window. Not all of the unit root observations breach the bounds, as is the case where only the first 25 unit root observations are included at the beginning of the sample. In this case the rolling CUSUM does give alarms during period 490 to 525, as the lower bound is broken in several occasions.

As the simulation results reveal, also the rolling CUSUM- test rarely gives false alarms of structural breaks. This result is in line with that reported in Leybourne, Kim and Taylor (2006b). A bit surprisingly, the test also misses many of the simulated unit roots in the data. Where the stationary period is already quite persistent, the method signals breaks more often. This could have something to do with the construction of the upper and lower bounds, as in the case where the stationary period is already quite persistent, the boundary values become narrower and are therefore easier to overrun.

Also the shortness of the samples seems to entail problems for the use of the rolling CUSUM. Since the underlying idea in this method is to detect shifts in the mean process, the construction of the mean process seems problematic in the context of a very short sample, as it becomes hard to recognize differences (especially when the majority of observations are already unit roots). Though not presented here, each of the simulations was reported in graphic form. From these rolling graphs it was easy to see that the rolling CUSUM seemed to react to the end of the simulated bubble more often than to the beginning. Therefore, the boundary was more often overrun when the process changed from unit root back to stationary. Instead of using only rolling samples, I also tested the whole samples (of size 1000 observations) using the conventional CUSUM test. According to these less extensive simulations and graphical analysis, it seemed that relatively often the whole-sample-based CUSUM was able to signal also the start of the unit root. This feature merely affirms that in order to operate with rolling samples the sample should be relatively large.

4.1.4 The power of the rolling Variance Ratio-test

Compared to the other tests presented here the number of total repetitions in the rolling variance ratio test was much smaller due to the test's time-consuming characteristics. The total amount of observations in each simulation was limited to 10 000.

The variance ratio test seems to be sensitive to two things: first to the underlying "normal" stationary-period regression coefficient and secondly to the length of the rolling window. As the memory during the "normal" period gets closer to 1 (ie moves from 0.6 to 0.8 and then to 0.9 in separate simulations), the number of correctly signaled bubbles increases, but so does the total number of false alarms. In each case the number of false alarms seems to increase after the simulated unit-root period. Unfortunately, at best only 6.4% of the single unit-root observations are signaled correctly from the simulated data, and in this case the false alarms already amount to 6.10%. A positive feature is that the number of unit-root periods spotted from the data is a quite high 30% this time.

From the table in Appendix VIII we see that the rolling variance ratio test seems to signal simulated unit roots correctly with higher probability when the window length is relatively short. On the contrary, it seems to less often signal stationary periods falsely as unit roots when the length of rolling window is increased. To reduce the probability of false alarms would therefore require that the rolling window be as long as possible; but to increase the probability of signaling unit roots correctly shorter rolling windows would be preferred.

Due to the sensitivity of the test to parameter change, it seems that this method might work well for what it was originally designed: to measure changes in long data series using shorter data samples. These data samples are separately constructed and they do not include overlapping, in contrast to rolling samples. Unfortunately, as the method is sensitive, the overlapping samples do not work well, as they share the common features. This might be one reason why the variance ratio test does not seem well suited for rolling samples.

5. AR- and ADF- based signals as leading indicators Evidence from history

In this section I study the market developments in the US stock markets and use the AR and ADF methods presented in section 3.2 to evaluate whether these methods are able to signal those periods as bubbles, which in previous literature have been recognized as periods of booms and which have been followed by busts. The bubble periods were collected from following literary references: Raines and Leathers (2000), Kindleberger (2000), Mishkin and White (2003), Shiller (2000), Bordo (2003) and IMF (2003). These consensus periods are used for reference purposes when evaluating the timing of bubble signals by the ADF- and AR-methods. I next examine how well the bubble warnings given by the AR- and ADF- indicators fit to these peaks and troughs in

cycles in real market history. Special attention will be devoted whether the bubble warnings come early enough to have been of use as early warning indicators of future problems.

5.1 Are AR- and ADF- indicators able to spot US stock market booms and busts?

How well do the AR- and ADF- bubble warnings spot these historical experiences of stock market booms, crashes and price busts? I ran the tests using each of the three window lengths (36, 48 and 60), but I report here only the statistical record of rolling tests with 36 observations. Overall, both the rolling AR- and rolling ADF- coefficients are able to signal most of the previously mentioned major booms in stock prices. It is interesting that they also trigger warnings in severe downturns, as the prices of stock have been depreciating in value for years. This feature might be related to the appearance of a negative bubbles, where the stocks get undervalued compared to their fundamentals. Especially interesting is that the cycle seems to turn shortly after the "negative" bubble warnings.

If we focus first on warnings of positive rational bubbles, it seems that both methods have the potential to serve as a leading indicator. Interestingly, they both signal warnings during the same periods, though the AR-method gives more alarm signals. Concerning the major stock-market booms - in the 1920's, the late 1980's and the technology boom at the end of 1990's - they are all spotted years in advance of the final crash. This is a valuable feature, since then the regulators and policymakers would have had time to react to the overheating. The table in Appendix X gives the precise timing of alarms given by rolling AR-indicators and rolling ADF-indicators as well as the timing of the consensus bubbles, but the major contribution of these indicators in real data is best illustrated by figures. Figure A in Appendix X shows the location of the consensus bubbles compared to US stock market developments in 1871-1949, the second figure (figure B) shows the location of the consensus bubbles in 1950-2010.

Figures C,D, E and F show, how the alarm signals by the ADF- and AR- indications fit into these periods, which commonly are considered periods of price misalignment. In these four figures the consensus bubbles are marked by green bars and the AR- and ADF- signaled periods of bubbles are marked by red lines. When the red line reaches value 1, this is a bubble signal. A bubble alarm will continue as long as the indicator gets the value 1.

The key message of these four figures is that the sudden and strong periods of growth, where the slope of the rise in stock prices clearly changes, are signaled as bubbles in the stock markets. Interestingly, these periods match well with the consensus bubble information, as the bubble alarms precede the consensus peaks. At best, they precede the peaks by years. This feature actually gives these indicators new significance, since they could yield important information as leading indicators of possible financial instability.

The signals given via the simple methods presented here also differ from the results of earlier academic research in several important respects. Phillips et al (2011b) are able to identify a few of the biggest booms in the long historical data for the US stock price series, periods that correspond fairly closely to those reported earlier, for example in Taipalus (2006). The two methods presented in this study, however, are able to give alarm-signals for even more of the consensus bubbles. An additional feature is that only these methods are able to warn of negative bubbles. In several cases these methods are also able to signal emerging bubbles earlier than those identified by Phillips et al (2011b).

6. Conclusions

This research presents two easy-to-use indicators that, due to use of short rolling windows in subsample construction and repeated regressions, are able to accurately detect emerging bubbles from monthly stock market data. Though these tools have been applied only to stock prices, they are relatively simple to apply to other asset prices, eg housing prices. The preliminary tests with US housing markets have generated promising results.

A clear advantage of these methods is their ability to perform accurately even with relatively modest sample sizes, as was shown via Monte Carlo simulations. Small sample sizes enable the use of these methods in various real time-series, and the indicators are able to detect even relatively short booms in prices. In addition, to my knowledge, these methods are among the few that are able to signal also negative bubbles, ie overly-negative corrections in asset prices compared to their fundamentals. Signals of negative bubbles usually arrive just before the turning point is observed in prices.

As the ADF- and AR- methods were run with US stock market data, they were able to signal major booms from the data and in many cases as early as 12 months prior to the crash. In simulated data, the methods found up to 70-80% of the simulated unit root periods. Because these methods seem to provide accurate and timely warning signals of exuberant prices, their potential use would seem to be extensive. In central banks these indicators could provide valuable information for two different types of pre-emptive policy actions: promoting financial stability and achieving the goals of macro-stability. As the tools now available to regulators require considerable time to take effect, it is crucial to get warning signals early enough - at best just when bubbles are starting to emerge. Therefore, especially interesting is the possibility to use these indicators together with various other stability indicators, such as credit growth, as signaling devices for regulators, as when to start the 'lean against the wind' in order to restrain dangerous developments or prevent unsustainable trends.

Literature

- Ajayi, R. A. - Karemera, D. (1996): A variance ratio test of random walks in exchange rates: Evidence from Pacific Basin economies, *Pacific-Basin Finance Journal*, Vol.4, No.1, p.77-91.
- Ayadi, F. O. - Pyun, C. S. (1994): An Application of Variance Ratio Test to the Korean Securities Markets. *Journal of Banking and Finance*, 18, p.643-658.
- Banerjee, A. - Lumsdaine, R.L. - Stock, J.H. (1990): Recursive and sequential tests of the unit root and trend break hypothesis: theory and international evidence. NBER Working Paper No. 3510
- Banerjee, A. – Lumsdaine, R.L. – Stock, J.H. (1992): Recursive and sequential tests of the unit-root and trend-break hypothesis: Theory and International evidence. *Journal of Business & Economic Statistics* 10, 271-287.
- Bordo, M. (2003): Stock Markets Crashes, Productivity Boom Busts and recession: Some Historical Evidence. Rutgers University, background paper for the WEO April 2003 chapter II.
- Brown, R.L. - Durbin, J. - Evans, J.M. (1975): Techniques for Testing the Constancy of Regression Relationships over Time. *Journal of the Royal Statistical Society (series B)*, Vol 37 No 2 p.149-192.
- Busetti, F. – Taylor, A.M.R. (2004): Tests of Stationarity against a change in persistence. *Journal of Econometrics* 123, 33-66.
- Busetti, F. - Taylor, A.M.R. (2001): Tests of stationarity against a change in persistence. Discussion paper, University of Birmingham.
- Byrne, J.P.- Perman, R. (2007): Unit roots and structural breaks: a survey of the literature, in (B. Rao, ed.), *Cointegration for the Applied Economist*, (second edition), Palgrave-Macmillan.
- Campbell, Lo and McKinlay (1997): *The Econometrics of Financial Markets*. Princeton University Press, NJ.
- Campbell and Shiller (1988a): The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors. *Review of Financial Studies*, Vol.1, No.3, 195-228.
- Campbell and Shiller (1988b): Stock prices, Earnings and Expected Dividends. *Journal of Finance*, Vol.43, No.3, 661- 676.
- Cavaliere, G. – Taylor, A.M.R. (2008): Testing for a change in persistence of non-stationary volatility. *Journal of Econometrics*, 147, 84-98.

- Choi, I. – Chung, B.S. (1995): Sampling frequency and the power of tests for unit root: A simulation study. *Economic Letters*, 49, 131-136.
- Craine (1993): Rational Bubbles - A Test. *Journal of Economic Dynamics and Control* 17, 829-846.
- DeJong, D. - Nankervis, J.C. - Savin, N.E. - Whiteman, C.H. (1992): The power problems of unit root tests in time series with autoregressive errors. *Journal of Econometrics*, 53, 323-343.
- Dickey, D.A. - Fuller, W.A. (1979): Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427-431.
- Dickey, D.A. - Fuller, W.A. (1981): Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057-1072.
- 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. *International Economic Review* 40, 767-783.
- Garcia, R. – Perron, P. (1996): An analysis of the real interest rate under regime shifts. *Review of Economics and Statistics* 78, 111-125.
- Harvey, D.I. - Leybourne, S.J. - Taylor, A.M.R.(2006): Modified tests for a change in persistence. *Journal of Econometrics* 134, p. 441-469.
- Hoque, H.A.A.B.- Kim, J.H. - Pyun, C.S. (2007): A comparison of variance ratio test emerging stock markets. *International Review of Economics and Finance*, 16:1.
- IMF(2003): WEO, April 2003 - Chapter II: When bubbles burst
- Kapetanios, G. (2005): Unit root testing against the alternative hypothesis of up to m structural breaks. *Journal of Time Series Analysis*, Vol.26, pp.123-133.
- Kim, T-H. - Leybourne, S. – Newbold, P. (2002): Unit Root Tests With a Break in Innovation Variance. *Journal of Econometrics*, 109, 365-387.
- Kim, J.Y. (2000): Detection of change in persistence of a linear time series. *Journal of Econometrics* 95, 97-116.

Kindleberger, C.P. (2000): *Manias, Panics and Crashes - a History of Financial Crises*. John Wiley & Sons Inc. 4th Edition.

Kousta and Serletis (2005): Rational Bubbles or Persistent Deviations from Markets Fundamentals. *Journal of Banking and Finance*.

Lanne, M. - Saikkonen, P. - Lutkepohl, H. (2002): Comparison of unit root tests for time series with level shifts. *Journal of Time series analysis*, Vol. 23, pp.667-685.

Lanne, M. - Saikkonen, P. - Lutkepohl, H. (2003): Test procedures for unit root tests in time series with level shift at unknown time. *Oxford bulletin of Economics and Statistics*, Vol. 65, pp. 91-115.

Lee, J. - Strazicich, M. (2003): Minimum Lagrange Multiplier unit root tests with two structural breaks. *Review of Economics and Statistics*, Vol 81, pp.1082-1089.

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

Leybourne, S. J. - Newbold, P. (2002): On the size properties of Phillips-Perron tests. *Journal of Time Series Analysis*, Vol. 20, Issue 1, p.51-61.

Leybourne, S. J. – Taylor, A.M.R. (2006b): Persistence change tests and shifting stable autoregressions. *Economic Letters* 91 (2006) 44-49.

Leybourne, S.J. – Kim, T.-H. – Taylor A.M.R. (2004): Regression based tests for a change in persistence. Department of Economics, Discussion Paper 04-12, University of Birmingham.

Leybourne, S. - Kim, T.H. - Newbold, P. (2005): Examination of some more powerful modifications of the Dickey-Fuller test. *Journal of Time Series Analysis*, Vol26, No.3, p. 355-369.

Leybourne, S. - Kim, T.-H. - Smith, V. - Newbold, P.(2003): Tests for a change in persistence against the null of difference-stationarity. *Econometrics Journal* (2003), Vol. 6, 291-311.

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

Li, H. (2007): Multivariate extensions of CUSUM procedure. A dissertation submitted to Kent State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy, August 2007.

Lo, A. - MacKinlay, G. (1989): The Size and Power of the Variance Ratio Test in Finite Samples: A Monte Carlo Investigation, *Journal of Econometrics*, 40, 203 — 238.

Lumsdaine, R.L. - Papell, D.H. (1997): Multiple trend breaks and the unit root hypothesis. *Review of Economics and Statistics*. Vol.79, pp.212-218.

Mellin, I. (2009): Tilastollinen laadunvalvonta - muut tilastollisen prosessin valvonnan menetelmät. CUSUM- kortti ja EWMA- kortti, luentomateriaali.

Mishkin, F.S. - White, E.N. (2003): US stock market crashes and their aftermath: Implications for monetary policy. Article in the book 'Asset Price Bubbles-the Implications for Monetary, regulatory, and International Policies' edited by Hunter, Kaufman and Pomerleano.

Page, E.S. (1954): Continuous Inspection Scheme. *Biometrika* 41 (1 / 2), p. 100-115.

Pantula, S. G. - Farias-Gonzales, G. - Fuller, W.A. (1994): A comparison of unit root test criteria. *Journal of Business & Economic Statistics* 12, 449-459.

Perron, P. (1989): The great crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57, 1361-1401.

Perron, P. (1990): Testing for a unit-root in a time series with a changing mean. *Journal of Business and Economic Statistics* 8, 153-162.

Pesaran, H. - Timmermann, A. (2005): Real-Time Econometrics. *Econometric Theory*, 21, 212-231.

Phillips, P.C.B. - Xiao, Z. (1998): A primer for unit root testing. *Journal of Economic Surveys* 12, 423-469.

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

Phillips, P.C.B. - Shi, S.P. - Yu, J. (2011b): Testing for Multiple Bubbles, Singapore Management University, working paper No.09-2011.

- Rea - Marcis (1996): Mutual fund shareholder activity during US stock market cycles, Investment Company Institute Perspectives, Vol.2, No.2
- Saikkonen, P. - Lutkepohl, H. (2002): Testing for a unit root in a time-series with a level shift at unknown time. *Econometric Theory*, 18, 313-348.
- Sayginsoy, O. (2003): Powerful and serial correlation robust tests of the economic convergence hypothesis. Working Paper, Department of Economics, Cornell University.
- Shi, S.-P. - Phillips, P.C.B. - Yu, J. (2010a): The hypothesis and model specification in the right-tail unit root test. Working paper.
- Shi, S.-P. - Phillips, P.C.B. - Yu, J. (2010b): Testing for periodically Collapsing Bubbles: A Generalised Sup ADF Test, Working paper
- Shiller, R. J. - Perron, P. (1985): Testing the random walk hypothesis: Power versus frequency of observation. *Economic letters* 18, p.381-386
- Shiller, R. J. (2000): *Irrational Excuberance*. Princeton University Press.
- Schwert, G.W. (1989): Tests for unit Roots: A Monte Carlo Investigation. *Journal of Business & Economic Statistics*, Vol 7, No.2.
- Stock, J.H. (1994): Unit roots, structural breaks and trends. In *handbook of Econometrics*, vol.IV., 2739-2841.
- Stock, J.H.- Watson, M.W. (1996): Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* 14, 11-30.
- Taipalus, K. (2006a): Bubbles in the Finnish and US equities markets. *Bank of Finland Studies E*:35.
- Taipalus, K. (2006b): A global house price bubble? Evaluation based on a new rent-price approach. *Bank of Finland Research Discussion Papers* 29/2006.
- Taylor, A.M.R. (2002): Regression-based unit root tests with recursive mean adjustment for seasonal and nonseasonal time series. *Journal of Business and Economic Statistics*, 20(2), p.269-281.

Taylor, A.M.R. (2005): On the use of subsample unit-root tests to detect changes in persistence. *Journal of Time Series Analysis*, Vol.26, No.5, 759-778.

Whang, Y.-J. - Kim, J. (2003): A multiple Variance Ratio Test Using Subsampling. *Economic Letters*, 79, p.225-230.

APPENDIX I. DISTRIBUTIONS OF THE ADF- AND AR-COEFFICIENTS DURING NORMAL PERIOD AND DURING BUBBLE¹⁰

Figure 1a:ADF

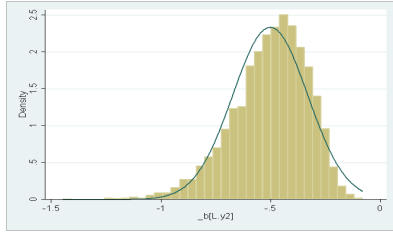


Figure 1b:ADF

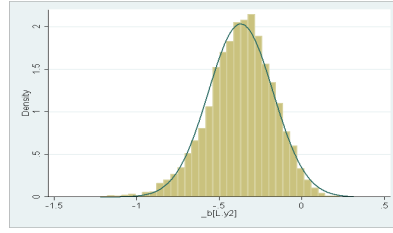


Figure 1c:ADF

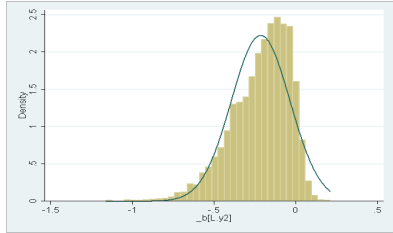


Figure 1d:ADF

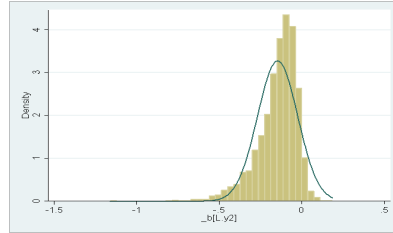


Figure 1e:ADF

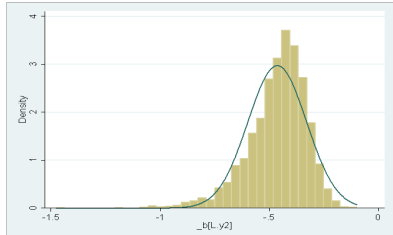


Figure 1f:AR

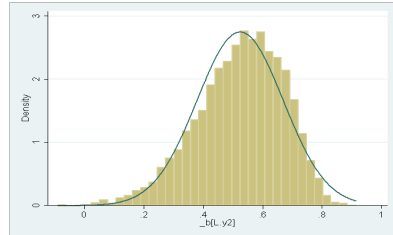


Figure 1g:AR

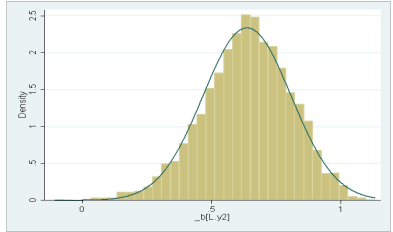


Figure 1h:AR

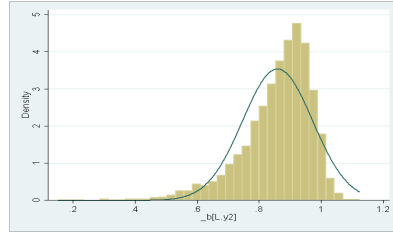


Figure 1i:AR

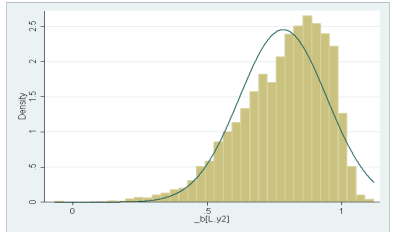
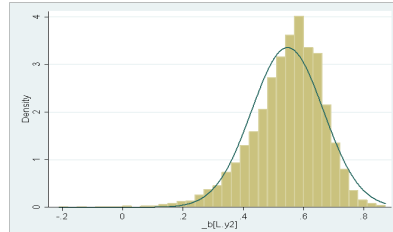


Figure 1j:AR



¹⁰ From figures 1a-1e it is easy to see how the ADF-coefficient distributions shift to the right and become more leptokurtic as more unit root observations enter the sample. In figure 1a there are no unit root observations, in figure 1b there are 5, in figure 1c the sample includes 15 unit root observations and in figure 1d all 36 unit-root observations are included in the sample. Figure 1e shows how coefficient values have reverted and distribution shifted back as all the 36 unit-root observations exit from the sample. Figures 1f-1j present the same outcomes for the AR- coefficients.

FIGURE 1k: Distribution of ADF coefficient values, the convergence of the average and the percentiles over simulated unit-root observations (stationary per.0.6)

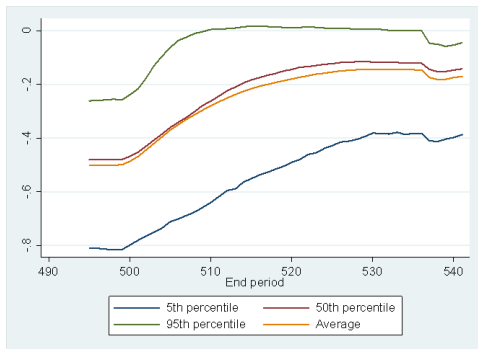


FIGURE 1l: Distribution of AR coefficient values, the convergence of the average and the percentiles over simulated unit-root observations (stationary per 0.8)

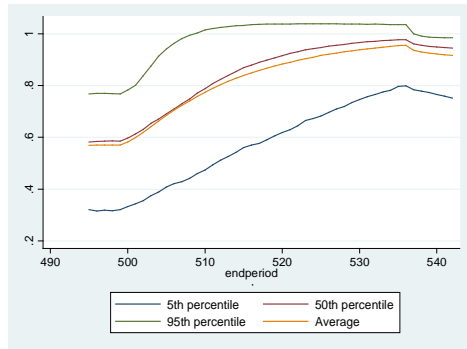


FIGURE 1m: Distribution of ADF coefficient values, the convergence of the average and the percentiles over simulated unit-root observations (stationary per. 0.8)

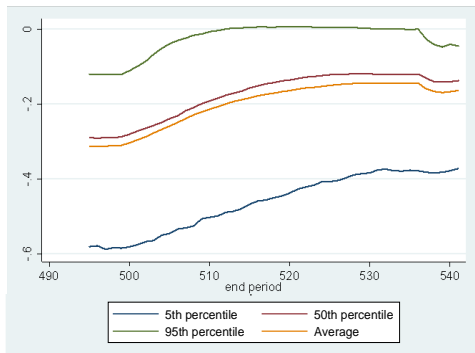
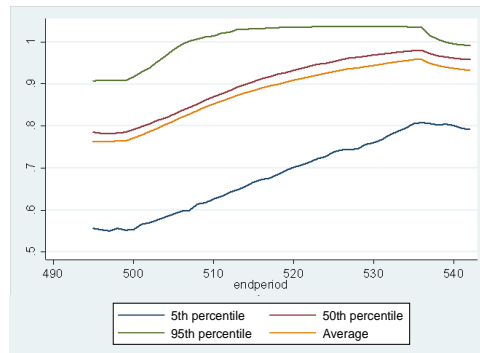


FIGURE 1n: Distribution of AR coefficient values, the convergence of the average and the percentiles over simulated unit-root observations (stationary per. 0.8)



APPENDIX II. CRITICAL VALUES

MC-Simulated Critical Values for the ADF-Coefficient

AR-coefficient in stationary period	Length of simulated unit root	Length of the rolling window	36	48	60
			0.6	36	-0,200
	48	-0,177	-0,150	-0,121	
	60	-0,152	-0,117	-0,093	
0.7	36	-0,153	-0,149	-0,142	
	48	-0,140	-0,128	-0,111	
	60	-0,127	-0,108	-0,888	
0.8	36	-0,104	-0,102	-0,100	
	48	-0,098	-0,094	-0,089	
	60	-0,092	-0,086	-0,077	
0.9	36	-0,051	-0,053	-0,053	
	48	-0,053	-0,051	-0,051	
	60	-0,048	-0,048	-0,047	

5% upper tail limits for ADF regression coefficients, T = 5000

APPENDIX III. REJECTION FREQUENCIES

FALSE ALARMS

%- of false alarms, where ADF coefficient 0 is used as a limit for the signal

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
			0.6	36	0,02%
	48	0,00%	0,01%	0,04%	
	60	0,00%	0,00%	0,01%	
0.8	36	0,13%	0,17%	0,22%	
	48	0,03%	0,07%	0,12%	
	60	0,01%	0,02%	0,07%	
0.9	36	0,82%	0,76%	0,80%	
	48	0,30%	0,35%	0,41%	
	60	0,12%	0,17%	0,22%	

%- of false alarms, where ADF coefficient -0.05 is used as a limit for the signal

0.6	36	0,31%	0,49%	0,65%
	48	0,41%	0,72%	0,98%
	60	0,58%	0,99%	1,98%
0.8	36	0,93%	1,10%	1,28%
	48	0,87%	1,18%	1,45%
	60	1,04%	1,47%	1,86%
0.9	36	4,50%	4,29%	4,42%
	48	3,85%	4,10%	4,27%
	60	3,80%	4,16%	4,41%

%- of false alarms, where AR coefficient 1.0 is used as a limit for the signal

0.6	36	0,01%	0,01%	0,00%
	48	0,00%	0,01%	0,00%
	60	0,00%	0,00%	0,00%
0.8	36	0,07%	0,06%	0,06%
	48	0,02%	0,04%	0,04%
	60	0,01%	0,01%	0,04%
0.9	36	0,75%	0,76%	0,74%
	48	0,29%	0,30%	0,30%
	60	0,13%	0,16%	0,17%

T = 5000

CORRECT SIGNALS, UNIT ROOT PERIODS

% of correctly signaled unit root periods, where ADF coefficient 0 is used as a limit for the signal

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36	37.54%	45.30%	53.72%	
	48	30.18%	36.74%	44.30%	
	60	24.82%	31.10%	37.42%	
0.8	36	35.00%	44.90%	22.86%	
	48	27.54%	35.56%	29.92%	
	60	22.44%	28.90%	36.20%	
0.9	36	38.68%	45.96%	53.64%	
	48	29.76%	35.86%	43.64%	
	60	23.82%	29.72%	35.78%	

% of correctly signaled unit root periods, where ADF coefficient -0.05 is used as a limit for the signal

0.6	36	59.70%	71.36%	79.04%	
	48	52.94%	65.90%	74.74%	
	60	48.18%	61.02%	70.36%	
0.8	36	60.48%	72.08%	53.04%	
	48	53.64%	66.00%	63.76%	
	60	49.38%	62.12%	72.02%	
0.9	36	64.30%	74.86%	81.36%	
	48	60.16%	71.46%	77.62%	
	60	56.30%	67.34%	74.46%	

% of correctly signaled unit root periods, where AR coefficient 1.0 is used as a limit for the signal

0.6	36	55.00%	67.90%	74.42%	
	48	41.88%	56.40%	63.22%	
	60	30.18%	44.54%	54.20%	
0.8	36	57.48%	68.34%	76.94%	
	48	43.54%	57.46%	64.76%	
	60	34.50%	47.62%	55.76%	
0.9	36	58.22%	59.74%	76.88%	
	48	46.26%	53.70%	66.20%	
	60	37.22%	49.44%	57.06%	

T = 5000

CORRECT SIGNALS

% of correctly signaled single unit root observations, where ADF coefficient 0 is used as a limit for the signal

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36	5.32%	5.32%	5.15%	5.13%
	48	4.89%	5.01%	4.89%	4.89%
	60	4.41%	4.71%	4.71%	4.69%
0.8	36	4.71%	4.71%	4.83%	2.26%
	48	4.24%	4.40%	3.35%	
	60	3.70%	4.15%	4.42%	
0.9	36	4.95%	4.95%	5.06%	5.03%
	48	4.77%	4.77%	4.60%	4.66%
	60	4.19%	4.28%	4.28%	4.29%

% of correctly signaled single unit root observations, where ADF coefficient -0.05 is used as a limit for the signal

0.6	36	15.74%	16.69%	17.50%	
	48	16.77%	19.23%	20.98%	
	60	15.91%	20.40%	23.39%	
0.8	36	15.80%	16.45%	16.45%	
	48	17.20%	19.86%	21.00%	
	60	16.82%	21.31%	24.22%	
0.9	36	17.28%	18.21%	18.59%	
	48	20.63%	22.04%	22.92%	
	60	21.64%	24.54%	26.46%	

% of correctly signaled single unit root observations, where AR coefficient 1.0 is used as a limit for the signal

0.6	36	16.20%	20.41%	23.14%	
	48	11.16%	15.47%	19.13%	
	60	7.81%	11.40%	15.34%	
0.8	36	17.36%	21.20%	23.75%	
	48	12.27%	16.58%	19.77%	
	60	8.89%	12.58%	16.12%	
0.9	36	20.76%	22.80%	24.83%	
	48	14.14%	18.08%	20.97%	
	60	10.66%	14.08%	17.29%	

T = 5000

APPENDIX IV. STATISTICAL POWER OF MULTIPLE PERIOD TEST

MC-simulation: Normal process 0.6, bubble length 36 observations, rolling window 36 observations				
Bubble signal received after 5 continuous signals				
ADF-method				
How many false alarms before bubble?				0,0 %
How many false alarms after bubble?				0,2 %
How many single 5-obs.bubble sets are signalled correctly?				9,1 %
Bubble period signalled; at least 1 (5-obs.) signal received during simulated bubble?				36,0 %
MC-simulation: Normal process 0.6, bubble length 36 observations, rolling window 36 observations				
Bubble signal received after 5 continuous signals				
AR-method				
How many false alarms before bubble?				0,0 %
How many false alarms after bubble?				0,0 %
How many single 5-obs.bubble sets are signalled correctly?				10,1 %
Bubble period signalled; at least 1 (5-obs.) signal received during simulated bubble?				34,2 %
MC-simulation: Normal process 0.9, bubble length 36 observations, rolling window 36 observations				
Bubble signal received after 5 continuous signals				
ADF-method				
How many false alarms before bubble?				0,0 %
How many false alarms after bubble?				1,9 %
How many single 5-obs.bubble sets are signalled correctly?				9,7 %
Bubble period signalled; at least 1 (5-obs.) signal received during simulated bubble?				37,6 %

APPENDIX V. STATISTICAL POWER OF ROLLING R- and DF^{inf} -TESTS

DF^{inf} -RESULTS

%- of total correct rejections of $H(0)=I(1)$ during stationary period

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root			
			36	48	60
0.6	36		52,04%	52,15%	52,42%
	48		73,21%	72,84%	72,99%
	60		86,74%	87,45%	88,21%
0.8	36		21,83%	22,06%	22,19%
	48		29,79%	29,88%	30,12%
	60		39,25%	39,56%	39,33%

%- of correct rejections of $H(0)=I(1)$ during stationary period, before and after simulated unit root

before					
0.6	36		55,80%	55,72%	55,85%
	48		81,07%	80,91%	81,04%
	60		97,33%	97,48%	97,61%
after					
0.6	36		49,60%	49,79%	50,10%
	48		68,28%	67,69%	67,73%
	60		80,33%	81,26%	82,28%

T = 5000

%- of false rejections of $H(0)=I(1)$ during simulated unit root in case of single unit root observations

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		39,40%	33,35%	29,02%
	48		63,18%	57,14%	51,17%
	60		83,15%	79,82%	66,19%
0.8	36		16,83%	15,22%	13,83%
	48		24,83%	22,01%	20,68%
	60		33,87%	31,90%	30,16%

%- of false rejections of $H(0)=I(1)$ during simulated unit root in case of unit root periods

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		29,90%	41,36%	38,02%
	48		25,12%	25,50%	23,20%
	60		11,42%	10,68%	10,08%
0.8	36		72,76%	70,18%	68,24%
	48		66,26%	65,68%	62,74%
	60		57,28%	56,30%	55,04%

T = 5000

R-RESULTS

%- of total false alarms during stationary period (i.e. erroneous signal of a break)

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		0,28%	0,57%	0,75%
	48		0,46%	0,64%	0,79%
	60		0,46%	0,70%	0,85%
0.8	36		0,56%	0,29%	0,37%
	48		0,28%	0,35%	0,42%
	60		0,29%	0,40%	0,50%

%- of total false alarms during stationary period (i.e. erroneous signal of a break), before and after the simulated unit root period

before					
0.6	36		0,00%	0,00%	0,00%
	48		0,00%	0,00%	0,00%
	60		0,00%	0,00%	0,00%
after					
0.6	36		0,46%	0,94%	1,26%
	48		0,75%	1,05%	1,31%
	60		0,75%	1,14%	1,39%

%- of correctly signaled single unit root observations

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		0,00%	0,00%	0,00%
	48		0,00%	0,00%	0,00%
	60		0,00%	0,00%	0,00%
0.8	36		0,00%	0,00%	0,00%
	48		0,00%	0,00%	0,00%
	60		0,00%	0,00%	0,00%

%- of correctly signaled unit root periods

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		0,00%	1,34%	1,82%
	48		0,00%	0,00%	0,20%
	60		0,00%	0,00%	0,00%
0.8	36		1,35%	2,94%	12,82%
	48		0,28%	0,60%	0,78%
	60		0,10%	0,14%	0,24%

T = 5000

APPENDIX VI. STATISTICAL POWER OF ROLLING MAX-TEST

%- of total false alarms, where unit root signal was received during stationary period

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	48	60
0.6	36		34,68%	34,70%	34,72%
	48		16,77%	17,01%	17,21%
	60		8,64%	9,26%	9,88%
0.8	36		73,16%	73,12%	73,13%
	48		60,65%	60,58%	61,05%
	60		47,10%	47,12%	47,73%

%- of false alarms before and after simulated unit root

before					
0.6	36		34,30%	34,24%	34,16%
	48		14,15%	14,16%	14,20%
	60		4,29%	4,37%	4,43%
after					
0.6	36		34,93%	35,01%	35,10%
	48		18,40%	18,83%	19,18%
	60		11,26%	12,28%	13,31%

%- of correctly signaled single unit root observations

0.6	36		81,23%	84,40%	85,92%
	48		69,80%	75,66%	79,03%
	60		58,69%	66,31%	54,61%
0.8	36		89,94%	90,76%	87,72%
	48		83,21%	85,68%	87,42%
	60		75,38%	78,67%	82,00%

%- of correctly signaled unit root periods

0.6	36		99,63%	99,92%	99,98%
	48		98,10%	99,58%	99,90%
	60		93,98%	98,12%	99,80%
0.8	36		99,86%	99,96%	99,98%
	48		99,10%	99,82%	99,94%
	60		97,70%	99,18%	99,60%

T = 5000

APPENDIX VII. STATISTICAL POWER AND ACCURACY OF CONVENTIONAL ADF t-TEST IN CASE OF ROLLING SAMPLES

POWER OF ROLLING ADF t-TEST

%- of total false alarms ($H(0)=I(1)$ is not rejected during stationary period)

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	% of total false alarms		
			36	48	60
0.6	36		65,89%	65,79%	65,72%
	48		45,97%	45,90%	45,88%
	60		28,11%	28,18%	28,15%
% of false alarms before and after simulated unit root					
before					
0.6	36		64,90%	64,84%	64,78%
	48		43,63%	43,61%	43,57%
	60		24,16%	24,14%	23,98%
after					
0.6	36		66,53%	66,42%	66,35%
	48		47,43%	47,35%	47,36%
	60		30,51%	30,67%	30,79%
% of correctly signalled unit root observations, single observations or periods					
single					
0.6	36		93,98%	94,73%	94,86%
	48		87,42%	89,97%	90,89%
	60		78,17%	83,02%	85,42%
periods					
0.6	36		99,90%	99,98%	99,98%
	48		99,62%	99,92%	99,96%
	60		98,84%	99,76%	99,94%

T = 5000

APPENDIX VII. STATISTICAL POWER OF ROLLING CUSUM-TEST

From figures C1-C4 one sees how rolling CUSUM-values change as more unit root observations enter the sample.. In figure C1 there are no unit root observations in the sample, in figure C2 there are 5, in figure C3 the sample includes 25 unit root observations and signals of structural breaks are received as the lower bound is breached several times. In figure C4 all 36 unit-root observations are included in the sample. As all observations are unit roots, there are no signals of structural breaks.

Figure C1.

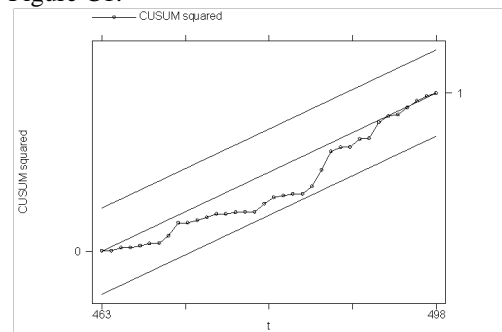


Figure C2.

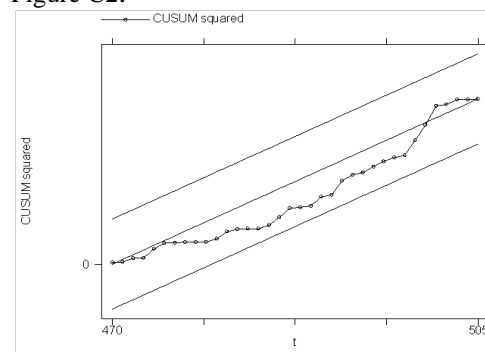


Figure C3.

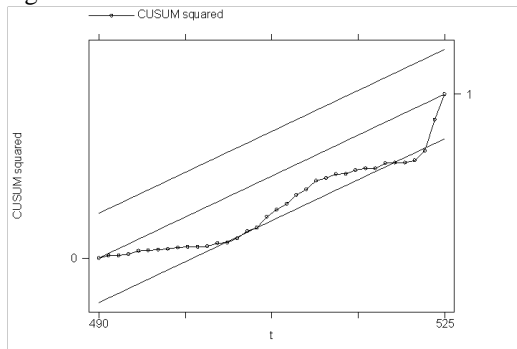
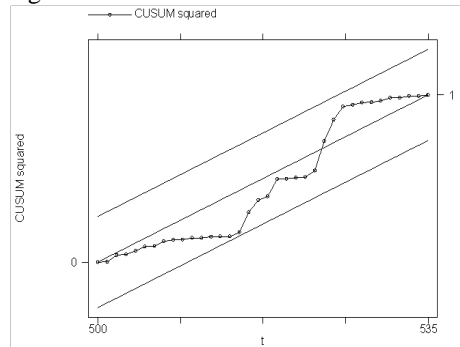


Figure C4.



POWER OF ROLLING CUSUM-TEST

%- of total false alarms for CUSUM-test

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root			
			36	48	60
0.6	36		1,06%	1,03%	1,06%
	48		1,58%	1,65%	1,72%
	60		2,28%	2,37%	2,47%
0.8	36		1,75%	1,72%	1,05%
	48		2,81%	1,57%	1,54%
	60		4,14%	2,40%	2,32%

%- of false alarms before and after simulated unit root

before					
0.6	36		1,00%	0,90%	0,92%
	48		1,29%	1,36%	1,37%
	60		1,71%	1,88%	1,87%
after					
0.6	36		1,10%	1,11%	1,15%
	48		1,27%	1,32%	1,40%
	60		1,81%	1,92%	2,05%

%- of correctly signalled unit root observations

0.6	36		1,33%	1,93%	2,19%
	48		1,77%	1,66%	2,30%
	60		2,50%	1,77%	2,37%
0.8	36		3,97%	3,50%	1,94%
	48		3,79%	2,97%	1,95%
	60		4,42%	2,50%	1,81%

APPENDIX VII. STATISTICAL POWER OF ROLLING VARIANCE RATIO TEST

%- of total false alarms for Variance Ratio-test

AR-coefficient in stationary period	Length of the rolling window	Length of simulated unit root	36	60
0.6	36		0,74%	1,01%
	48		0,47%	0,65%
	60		0,33%	0,57%
0.8	36		3,68%	--%
	48		2,65%	--%
	60		2,02%	--%
0.9	36		6,10%	--%
	48		4,93%	--%
	60		4,44%	--%

%- of total false alarms for Variance Ratio-test, before and after simulated bubble

before				
0.6	36		0,68%	0,60%
	48		0,32%	0,24%
	60		0,34%	0,15%
after				
0.6	36		0,78%	1,31%
	48		0,57%	0,96%
	60		0,40%	0,89%

%- of correctly signalled single unit root observations

0.6	36		2,97%	5,07%
	48		0,78%	2,80%
	60		0,17%	1,85%
0.8	36		5,83%	--%
	48		3,75%	--%
	60		3,36%	--%
0.9	36		6,44%	--%
	48		4,86%	--%
	60		4,56%	--%

%- of correctly signalled unit root periods

0.6	36		23,00%	42,00%
	48		9,00%	23,00%
	60		3,00%	19,00%
0.8	36		31,00%	--%
	48		19,00%	--%
	60		14,00%	--%
0.9	36		30,00%	--%
	48		23,00%	--%

APPENDIX X. STOCK MARKET BOOMS, CRASHES AND RECESSIONS IN THE US 1850-2010

Table: Precise bubble-warnings given by the 3years ADF- and AR- indicators

US Stock Market data			
Sub-sample length 36			
AR-signals bubble	ADF-signals bubble	Major cause of boom and bust*	Identifies boom or bust
12.1876,02-06.1877,12.1877-01.1878	-----	Railroad boom and following bust	bust
04.1878-11.1879	-----		
11.1885-04.1887	08.1886-11.1886	Railroad boom and following bust	bust
07.-09.1893	07.-08.1893	Silver agitation	bust
12.1893-01.1894	-----	Silver agitation	bust
02.1899-04.1899	-----		
08.1900-01.1900	-----	Boom before rich mans's panic	boom
01.1903-02.1904	01.1903-03.1904	Rich mans's panic	bust
12.1905-01.1906	-----	World financial crises	bust
08.1907-03.1908	08.1907-02.1908	World financial crises	
12.1915	-----		
05.1917-05.1918	07.1917-12.1917	War	bust
11.1925-01.1926	-----	Roaring 20's	boom
05.1928,09.1928-09.1929	11.1928,01.1929-03.1929,07.1929-09.1929	Roaring 20's, Crash in Oct 1929	boom
09.1931-01.1932,04.1932-06.1932	12.1931,04.1932-06.1932	30's recession	bust following 1929 crash
01.1937-06.1938	-----	Tight monetary policy	bust
05.1943-10.1943	-----		
01.1946-07.1946	02.1946,04.-07.1946	Post war slump, prices peak in July	bust
02.-03.1948	-----	Post war slump	
09.1954-03.1956	12.1954-04.1956	Strong market rise 1953-1955	
01.1959,04.1959-08.1959	-----		
08.1966-10.1966	08.1966-10.1966	September-October -25.2%	
05.1970-08.1970	05.1970-07.1970	Penn-Central, Bretton Woods	bust
04.1974-01.1975	04.1974-01.1975	Oil Shock	bust
02.1978-04.1978	03.1978	Oil Shock	bust
11.1978	-----	Oil Shock	
03.1986-09.1987	04.1986-08.1986,01.1987-04.1987,08.1987	Sharp rise in stock prices 1984-1987, crash in Oct.1987	boom
07.1995-03.1997,05.1997-10.1997	09.1995,11.1995-06.1996,09.1996-03.1997,05.1997-10.1997	Information technology boom and crash in 2000	boom
02.1998-07.1998	03.1998-07.1998	Russian default and LTCM	
08.2002-11.2002,02.2003	07.2002-10.2002,02.2003	Information technology bust	bust
03.2008	03.2008	Leverage-bubble in the US housing markets and securitisation	bust
07.2008-04.2009	07.2008-03.2009		

* For reference: IMF (2003), Shiller (2000), Raines-Leathers (2000), Mishkin-White(2003) & Bordo (2003)

FIGURE A. TIMING OF CONSENSUS BUBBLES IN US STOCK MARKETS 1871-1949

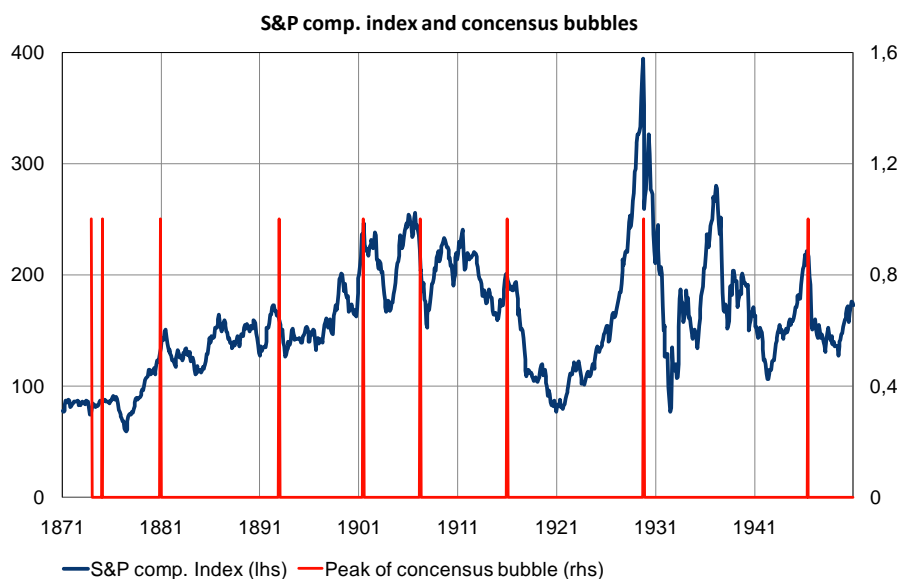


FIGURE B. TIMING OF CONSENSUS BUBBLES IN US STOCK MARKETS 1950-2010

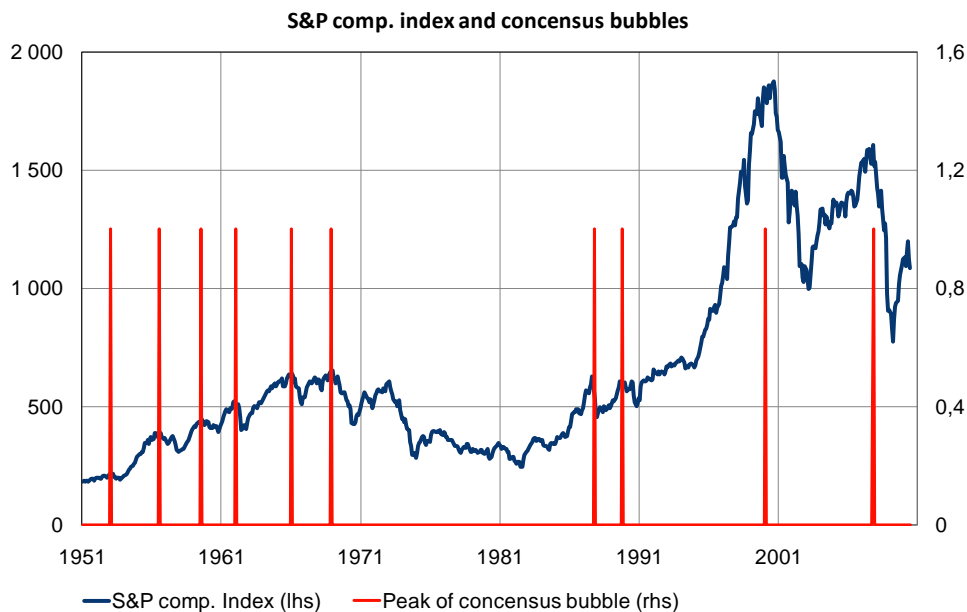


FIGURE C. S&P COMPOSITE INDEX, ADF- BUBBLE WARNINGS AND TIMING OF CONSENSUS BUBBLES 1871-1950

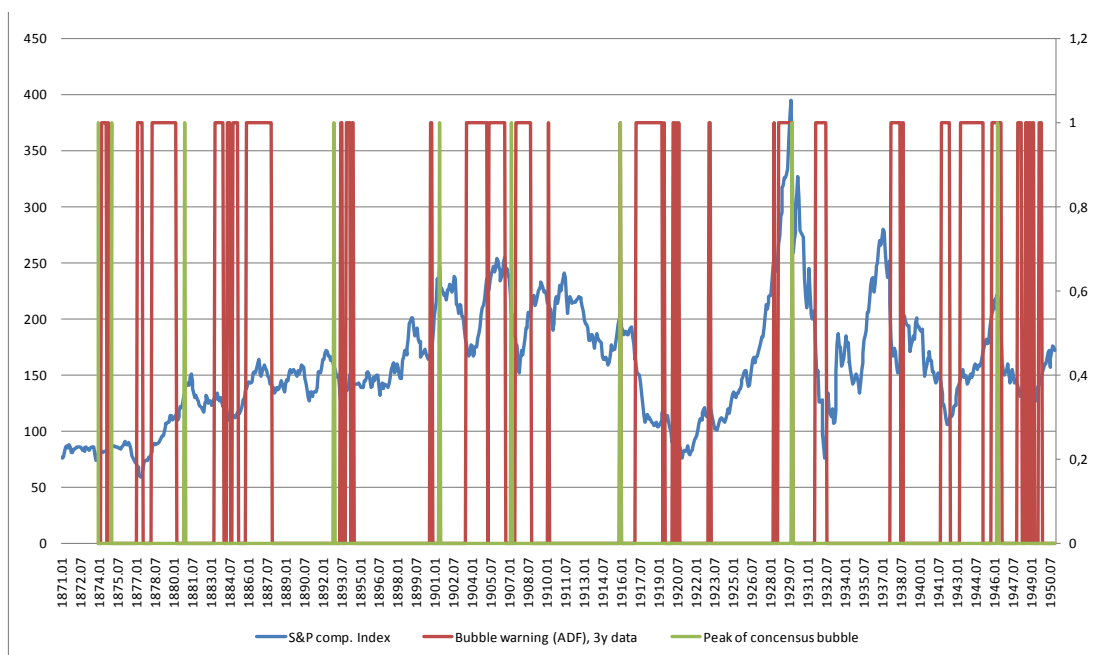


FIGURE D. S&P COMPOSITE INDEX, ADF- BUBBLE WARNINGS AND TIMING OF CONSENSUS

BUBBLES1951-2010

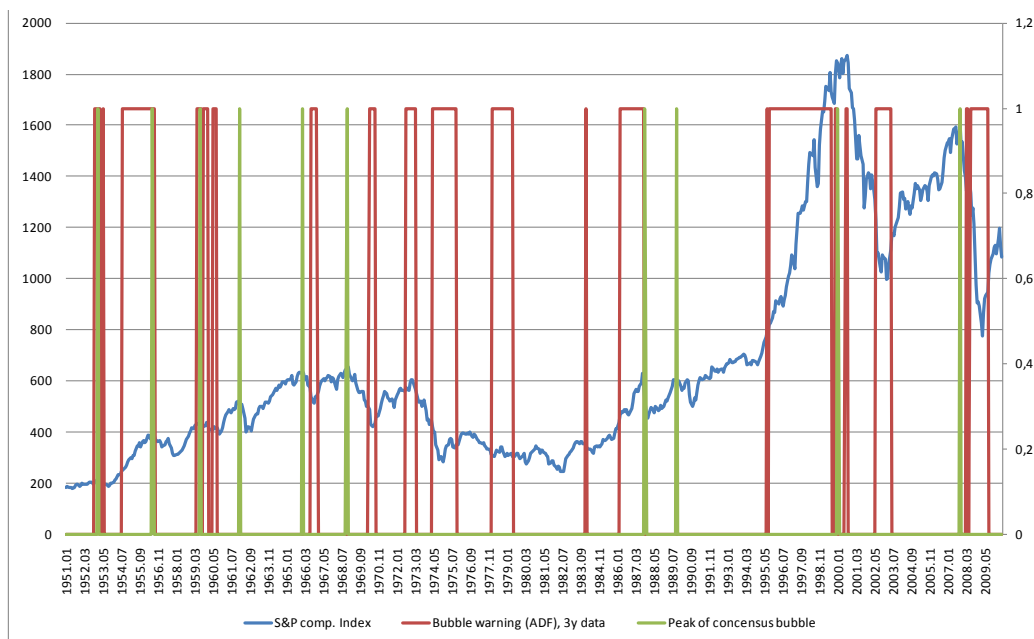


FIGURE E. S&P COMPOSITE INDEX, AR- BUBBLE WARNINGS AND TIMING OF CONSENSUS

BUBBLES 1871-1950

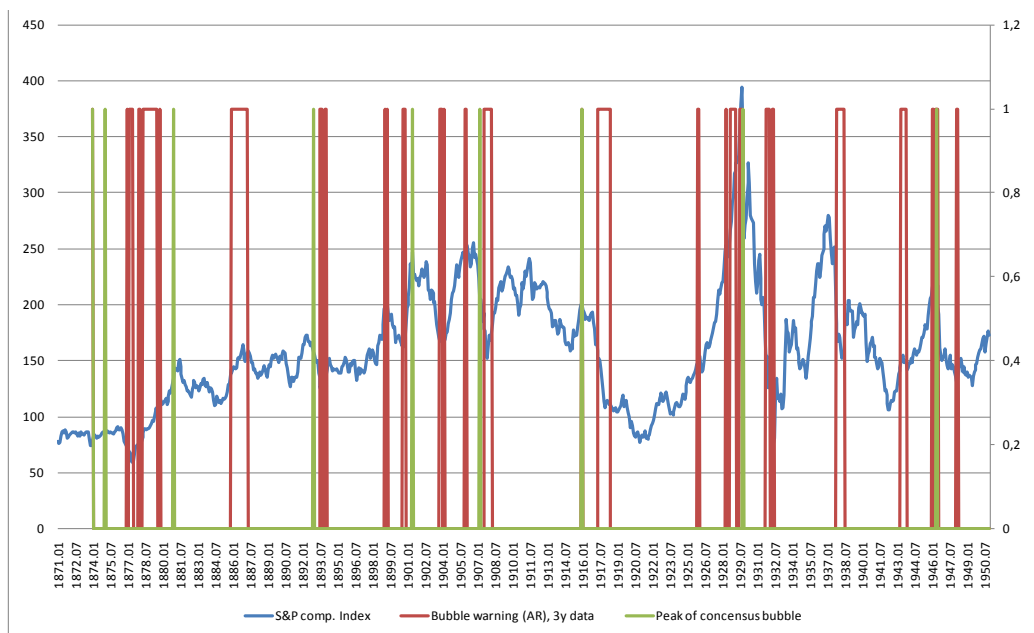
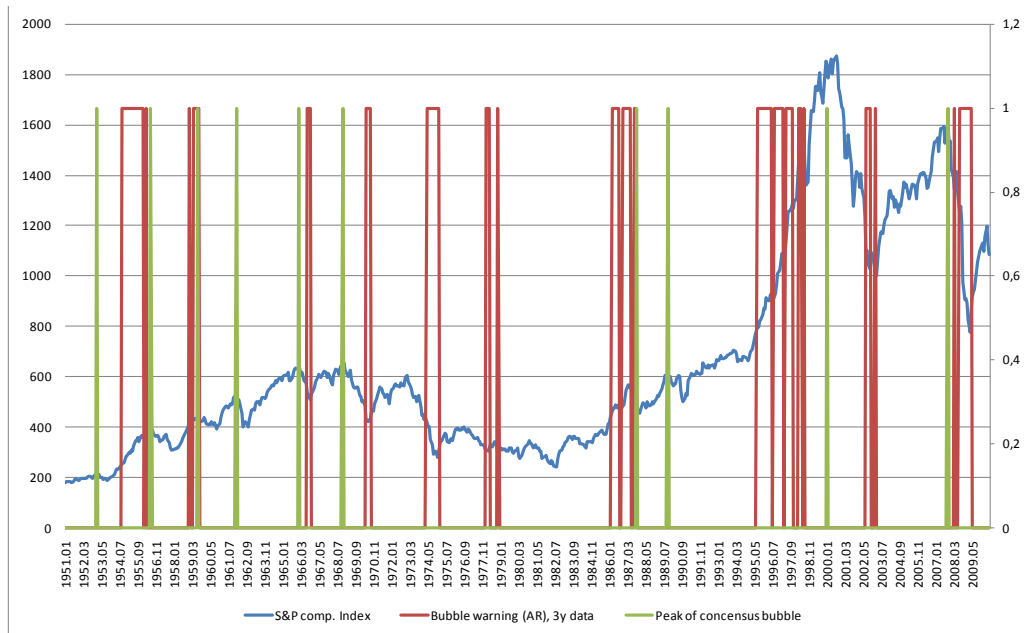


FIGURE F. S&P COMPOSITE INDEX, AR- BUBBLE WARNINGS AND TIMING OF CONSENSUS BUBBLES 1951-2010



**BANK OF FINLAND RESEARCH
DISCUSSION PAPERS**

ISSN 1456-6184, online

- 1/2012 Maria Teresa Punzi **Housing Market and Current Account Imbalances in the International Economy.** 2012. 27 p. ISBN 978-952-462-784-9, online.
- 2/2012 Luisa Lambertini – Caterina Mendicino – Maria Teresa Punzi **Expectations-Driven Cycles in the Housing Market.** 2012. 61 p. ISBN 978-952-462-785-6, online.
- 3/2012 George A. Waters **Quantity Rationing of Credit.** 2012. 26 p. ISBN 978-952-462-786-3, online.
- 4/2012 Karlo Kauko **Why is equity capital expensive for opaque banks?** 2012. 26 p. ISBN 978-952-462-787-0, online.
- 5/2012 Kaushik Mitra – George W. Evans – Seppo Honkapohja **Fiscal Policy and Learning.** 2012. 32 p. ISBN 978-952-462-788-7, online.
- 6/2012 Ian W. Marsh – Wolf Wagner **Why is Price Discovery in Credit Default Swap Markets News-Specific?** 2012. 41 p. ISBN 978-952-462-789-4, online.
- 7/2012 Katja Taipalus **Signaling Asset Price Bubbles with Time-Series Methods.** 2012. 47 p. ISBN 978-952-462-790-0, online.

