Improving EWIs for banking crises - satisfying policy requirements[1]

Mathias Drehmann and Mikael Juselius

Bank for International Settlements

August 2012

Draft please do not circulate

Abstract:

We outline the statistical properties that an ideal EWI for banking crises should have from a policy perspective. Specifically, indicators should be precise, have the correct timing and issue stable signals. We then relate these properties both to the objective function of an often employed estimation procedure, the signals extraction method, as well as to a standard set of indicators variables. We identify potential weaknesses with respect to both aspects and discuss remedies for each. In particular, we suggest using a more general objective function which allows for a more careful assessment of the optimal timing provided by the signal. We also introduce and discuss a new indicator variable, the debt service ratio, which displays more stable long-run dynamics than previous indicator variables. And we analyze combinations of individual variables. We find that these remedies improve EWIs to satisfy policy requirements.

JEL classification:

Keywords: EWIs, ROC

[1] The views expressed in the paper are those of the authors and do not necessarily represent the views of the BIS. Mathias Drehmann: Bank for International Settlements, Centralbahnplatz 2, CH-4002 Basel, Switzerland, mathias.drehmann@bis.org. Mikael Juselius: Bank for International Settlements, Centralbahnplatz 2, CH-4002 Basel, Switzerland, mikael.juselius@bis.org.

1. Introduction

Macroprudential policy can only succeed if emerging financial vulnerabilities are detected early enough for preventive action to be taken. How this can be done has receive much attention recently, with early warning indicators (EWIs) playing a prominent role.[2] More often than not, though, the forecasting performance of EWIs and other financial stability models is evaluated predominantly based on various statistical criteria without much regard for their policy relevance. The objective of this paper is to bridge this gap by explicitly relating the choice of statistical procedures for constructing and evaluating EWIs to various policy requirements. In doing so, we propose a number of novel extensions to existing techniques and EWIs.

We proceed in a straightforward manner: we focus on EWIs for systemic crises as these lead to substantial output losses (see e.g. …) and have been the main target for macroprudential policies. We begin by outlining the ideal properties that such indicators should have from a policy perspective. Specifically, these properties concern the relative costs of type I and type II classification errors, as well as the timing and consistency of the signals. We then draw the implications of each of these properties for the choice of estimation procedure and EWI.

Ideally, policymakers would know the cost and benefits of macroprudential intervention, which in turn would determine their relative aversion of type I versus type II errors. In a world where costs of macroprudential interventions are low, but benefits high, a policymaker will have a high aversion for failing to predict a crisis – a type I error – whilst placing less emphasis on type II errors. Indeed, if no reliable EWIs are available, this type of policymaker may prefer structural measures, such as high capital requirements, to safeguard against financial turmoil. However, in such cases, even relatively imprecise EWIs may help reduce reliance on the structural measures, thereby lowering costs of intervention. Of course, the opposite holds if benefits of interventions are low relative to costs, so that policymakers a high aversion against type II errors.

The costs and benefits of macroprudential interventions are, however, unknown. Given reasonable ranges for unobservable policy parameters, Drehmann (2012) finds in a simulation study that the scope for policymakers' preferences is surprisingly wide. Even the extreme cases, when policymakers essentially care only about type I or type II errors fit into this range for signals which are not perfect. This finding has direct implications for the choice of evaluation criteria. In particular, these criteria have to be neutral with respect to the policymakers' objective function. For this reason we apply a new technique – the receiver operating characteristic (ROC) curve – to assess the performance of potential EWIs.

The ROC curve summarizes all potential type I – type II trade-offs that a prediction model for binary variables can generate. Clearly, a model with fewer classification errors is better, and this is reflected in a larger area under the ROC curve (AUROC). As such, AUROC can be used as a summary measure to judge the forecast performance of any binary prediction model. As it has convenient statistical properties, significance tests or hypothesis tests can be easily implemented. In contrast to other methods for evaluating binary predictors, such as minimum noise-to-signal ratios or log probability scores, this method does assume an underlying loss function. Hence, evaluations based on the ROC curve are robust to different policy preferences and thus is an ideal approach in a situation where these are unknown.

---

[2] For a recent survey see e.g. Bisias et al (2012).

Despite its heavy use in other sciences like medicine, engineering or meteorology, the AUROC curve has not seen many applications in economics. Recent exceptions include Berge and Jorda (2011) who evaluate business cycle indicators, Taylor and Jorda (2010) who study different investment strategies, and Jorda et al (2011) who evaluate one crisis model. We expand on this literature by looking more closely at a large range of potential EWIs, as well as providing confidence intervals and hypothesis tests.

The second issue that we raise concerns the timing and consistency of the EWI. On the one hand, it needs to signal impeding crises sufficiently early. Because of long lags between the times that a signal is observed, policy action is taken, and the impact is felt in the economy (see e.g. Basel Committee (2010)), the signal should be received at least one and a half to two years before a crisis in order to be effective. On the other hand, EWIs should not signal crises too early, more than five years ahead, say. The reason is that political pressure, which may weaken the impact of macroprudential interventions, has a tendency to build-up as booms progress (eg Caruana (2010)). For the same reason, EWIs need to persistently issue signals as not to provide conflicting signals whether to intervene or not.

We investigate the temporal performance of established EWIs (see e.g…) in much more systematic way than has hitherto been conducted in the literature. In particular, we calculate the AUROC separately for each date in the forecast interval and compare the resulting time profiles. It turns out that the indicator variables, which issue the most consistent signals, are also the ones that are both smooth and highly persistent – ie each display (near) double unit-root behaviour.[3]

The finding of near I(2) dynamics with respect to the signal variables has implications for the choice of estimation procedure. In particular, the statistical properties of standard regression based models for binary choice are still to a large extent unknown under this degree of persistence.[4] Hence, estimation based on logit or probit type models seem less appealing for the present application. Moreover, these models are estimated to maximize a specific likelihood function that, to the extent to which it is subject to misspecification, can perform arbitrarily bad at specific points of the policymaker's loss function (Elliott and Lieli (2010)). For these reasons we adopt a non-parametric signal extraction approach in the spirit of Kaminsky and Reinhart (1999). This does not impact on the applicability of the ROC curve, as it can be used for all binary classifiers, whether they are derived from a regression model or a non-parametric approach.

Summarizing, the policy perspective that we have adopted here seems to call for an approach to constructing and evaluating EWIs which: (i) uses non-parametric estimation, (ii) evaluates forecast performance based on AUROC and (iii) explicitly takes timing and consistency into account.

We apply our approach to assess the performance of selected EWIs on a sample of 27 economies, covering quarterly time-series starting in the early 1980's. A total of 25 crises are included in the sample. We focus, in particular, on two indicators which have been found to work well in the past: the credit-to-GDP gap and the property-price gap. These gaps are defined in terms of the deviations between variables and thier long-term trends, and are taken to capture excessive leverage and asset price booms. We also assess the

---

[3] Auto-regressive processes which contain double unit-roots in their characteristic polynomial, typically display smooth patterns. We do not attach any structural meaning to such roots here. Instead we view them as a convenient way of characterizing the relevant order of persistency within a given sample.

[4] Park and Phillips (2000) develop an asymptotic theory for binary choice models where the conditioning variables are allowed to be I(1).

performance of a new indicator variable – the debt service ratio (DSR). The DSR is defined as the proportion of interest payments and mandatory repayments of principal to income. It can be interpreted as capturing incipient liquidity constraints of private sector borrowers and it has been recently introduced by Drehmann and Juselius (2012). Finally, we include GDP-growth as a comparator variable.

We find that both the DSR and the credit-to-GDP gap signifincatly dominate the other EWIs. In particular, both of these variables yield AUROC values which are very close to unity, i.e. they are close to being perfect forecasts. However, they differ more substantially with respect to their temporal profiles. The predictive ability and precision of the DSR is extremely high in the last 4 quarters leading up to a crisis, but they deteriorate steadily as the horizon increases. In contrast, the credit-to-GDP gap performs consistently well, even over horizons as long as 5 years before crises events. The other two EWIs have AUROC values which are significantly lower than those of the DSR and the credit-to-GDP gap over most horizons. Moreover, they are less reliable over time. For instance, the performance of the property price gap peaks 2-3 years before a crisis and deteriorates rapidly as the horizon shortens. The performance if real output growth is altogether unreliable.

The remaining paper is organized as follows…

2. Constructing EWIs based on policy requirements

Good forecasts often require a clear definition of the objective, as models can have many different uses (Lawerence et al (2000)). For example, scholars may be interested in testing whether certain theories predict out of sample, financial traders may want to take positions based on crises probabilities, and supervisors may need EWIs to rally support for prompt corrective actions. In this section, we discuss such objectives from the perspective of marcoprudential policy, which is concerned with system-wide financial risks that could have serious negative consequences for the real economy (IMF, FSB and BIS, 2009). Specifically, we limit attention to EWIs of baking crises, which can guide the build-up of buffers in "good times", so that these can be released to absorb losses in "bad times".[5]

The macroprudential policy objective has several implications for the choice of empirical strategy. First, it provides the framework to analyse costs benefits of interventions and thus helps to determine the policymaker's relative aversion for type I and type II errors. Second, it has implications for the desired temporal properties of the EWIs. In this section, we discuss both these issues in turn.

2.1 Costs of crises vs. costs of regulation

To make matters concrete, we analyze a vey simple economy looking only at the build-up phase for buffers. There are two states of the world: there is currently a boom which will inevitably lead to a crisis in the next period (C=1) or there is none (C=0). Policy makers can either impose buffers (B=1) or not (B=0).

[5] Policy discussions centre around two different objectives for macroprudential policies: increasing the resilience of the financial system or smoothing the financial cycle (see e.g. CGFS, 2010 or IMF..)

Table 1 summarizes the utility of policy makers ($U_{B,C}$) depending on the state of the world and the decision by policy makers. Starting with the world in which there is no crisis in the next period: if policy makers do not require buffers ($U_{B=0,C=0}$) welfare losses are zero, if buffers are implemented ($U_{B=1,C=0}$) the economy faces the costs of regulation. However, if there is a crisis and no buffers are imposed ($U_{B=1,C=0}$) the economy has to bear the full cost of a crisis. Finally, if a crisis materializes and buffers are imposed ($U_{B=1,C=0}$) , the economy has to pay the costs of regulation but the costs of crises are reduced by a factor α.

Table 1

	No crises next period	**Crises next period**
No buffers	$U_{B=0,C=0}$ *=0*	$U_{B=0,C=1}$ *=-Cost of crisis*
Impose buffers	$U_{B=1,C=0}$ *=-Cost of regulation*	$U_{B=1,C=1}$ *= -(1-α)Cost of crisis - cost of regulation*

α is between zero and one and summarizes the effectiveness of the macroprudential tool

Policy makers observe an indicator variable $S \in R$, which issues a noisy signal about the state of the economy. In particular, we assume that the true positive *(TP)* and false positive rate *(FP)* of the indicator variable are related to the threshold $\theta \in R$ with

TP(θ)= prob(S> θ| C=1)

FP(θ)= prob(S > θ| C=0).

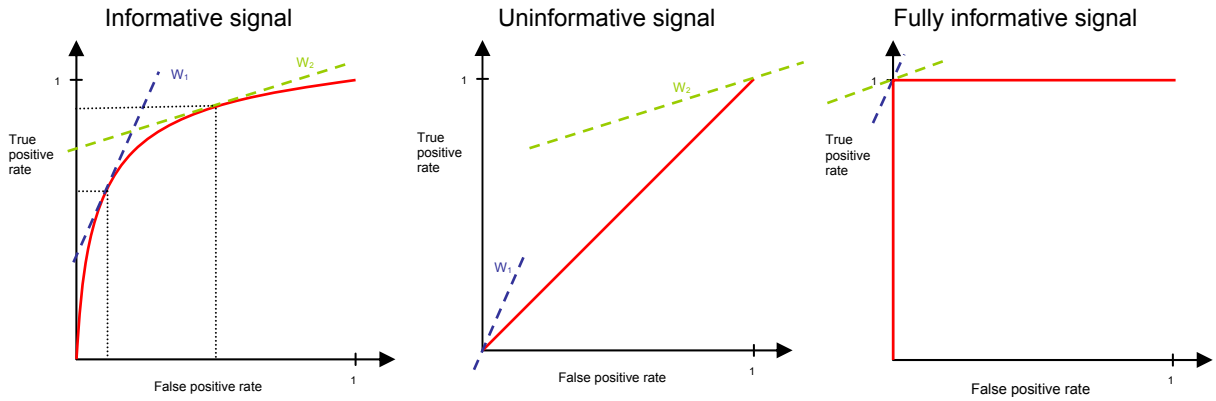
For any given value of θ, the trade-offs between true and false positives are contained in the upper half above a 45° line within a unit square.[6] Obviously, TP→0 and FP→0 as $\theta \to \infty$, and TP→1 and FP→1 as $\theta \to -\infty$. For uninformative indicators, these trade-offs will move along the 45° line as θ goes from minus to plus infinity, whereas the trade-offs will move closer to the upper boundary of the unit square for highly informative ones.

The trade-offs between FP and TP are summarized by the receiver operating characteristic (ROC). The ROC is a mapping from FP to TP and has been extensively in other sciences like engineering or medicine, but has only recently been introduced into economics (see e.g. Berger and Jorda (2011)). Plotting the ROC curve provides a quick visual summary of the quality of a signal. Graph 1 shows hypothetical ROC curves for a fully informative signal, a completely uninformative signal and an intermediate case.

[6] It case the trade-offs are in the lower half, a simple transformation of the indicator variable will ensure that they are in the upper half.

Graph 1

Signal quality and welfare



The decision problem for the policy maker is therefore to determine the threshold θ so to minimize the welfare costs of crises and regulation. Given a particular signal, the utility of the policy maker can be written as:

$$U(\theta)= p(crisis)^* [TP(\theta) * U_{B=1,C=1}+(1-TP(\theta))*U_{B=0,C=1}] + (1-p(crisis))*[FP(\theta) * U_{B=1,C=0}+ (1-FP(\theta))*U_{B=0,C=0}]$$

where p(crisis) is the (unconditional) probability of a crisis.

Given the ROC curve, the policy maker's decision problem can be recast as a maximisation problem with respect to FP, as θ determines FP. And for a specific FP, ROC tells the policy maker TP (see Graph 1), ie formally TP=ROC(FP). Thus

$$U(FP)= p(crisis)^* [ROC(FP) * U_{B=1,C=1}+(1-ROC(FP))*U_{B=0,C=1}] + (1-p(crisis))*[FP * U_{B=1,C=0}+ (1-FP)*U_{B=0,C=0}]$$

Maximizing the utility with respect to FP leads to the following condition

$$\frac{\partial ROC}{\partial FP} = \frac{1-p(crisis)}{p(crisis)} * \frac{U_{B=0,C=0}-U_{B=1,C=0}}{U_{B=1,C=1}-U_{B=0,C=1}} = \frac{1-p(crisis)}{p(crisis)} \frac{Cost\ of\ regulation}{\alpha * Cost\ of\ crises\ - Cost\ of\ regulation}$$

Thus, the optimal trade-off between true and false positive rate is at the point where the slope of the ROC curve equals the marginal rate of substitution between expected benefits from taking action if there is a crisis versus the expected costs of imposing the buffer in case there is no crisis.

If the expression on the right hand side of (2) is larger than unity – the relative cost of policy is high - the policy maker will care more about false positives than he will about true positives. Put differently, the policy maker will worry more about making type II errors than type I. The converse happens when the relative cost of policy is low. However, in the extreme case where the indicator is completely uninformative, $\partial ROC / \partial FP = 1$, implying that the policy maker will always take action if *Cost of regulation* < α * *p(crisis) *Cost of crisis*, and vice versa. Similarly, the ROC curve moves through the point (0, 1) for a fully informative indicator, implying that there is a *θ* such that *TP(θ) = 1* and *FP(θ) = 0* and the correct policy action will always be taken.

The decision problem of the policy maker can also be analyzed graphically. In FP\TP space the utility of the policy maker is a flat line as shown in Graph 1. The utility is higher, the higher the true positive rate with the same FP, ie shifting to the north-west in Graph 1 increases utility. The slope of the utility function depends on the probability of crises and the expected costs and benefits of interventions.

As can be seen from the graph, for an uninformative signal policy makers either always intervene (TP=1) or never (TP=0).[7] They will do the former, if the slope of the utility function is flatter than the ROC curve of the uninformative signal. Hence, for an uninformative signal policy makers will always intervene if the expected benefits of regulation outweigh the costs that have to be paid with certainty. Intervening in every period is equivalent to setting minimum standards, such as minimum capital requirements.

In the stylised model, policy makers can often improve utility by conditioning policy actions on an informative signal. If the signal is fully informative, it is optimal to act whenever the signal indicates this (right hand panel, Graph 1). For the more realistic case, policy makers want to act state contingent (middle panel, Graph 1),

To determine the optimal threshold, *θ*, policymakers have to know costs and benefits of intervention. Unfortunately, such information is rarely available. Drehmann (2012), therefore implements a simulation study showing that the scope for policymakers' preferences is surprisingly wide for realistic values of the costs of crisis, the costs of regulation, their benefits and the likelihood of crisis. Even the extreme cases, when policymakers essentially care only about type I or type II errors fit into this range for signals which are not perfect.

However, even without any knowledge of these utilities, it is still possible to evaluate the performance of the indicator. The idea is to generate a summary measure of the indicators classification ability by calculating the area under the ROC curve which is given by

$$AUROC = \int_0^1 ROC(FP)dFP$$

This area (AUROC) is increasing with the predictive power of the indicator across all possible thresholds θ and obviously lies between 0.5 for uninformative predictors, and 1 for and prefect predictors. AUROC can be estimated parametrically or non-parametrically and has convenient large sample properties so that hypothesis testing, for example whether AUROC is significantly different from 0.5, can easily be implemented (eg Jorda and Taylor (2011)).

Given its independence from policymakers preferences, we use AUROC as the measure to evaluate the forecast performance of EWIs in this paper.

---

[7] Technically, there is also a degenerate case where policy makers are indifferent between intervention and not and the slope of the utility curve is 1.

2.2 Timing, persistency, and consistency

Next, we discuss the appropriate timing of an ideal EWI. This issue is more complex from a policy perspective than from a purely statistical point of view. First, EWIs need to signal impeding crises early enough. For monetary policy, the typical lead-lag relationship between changes in the interest rates and inflation is around one to two years (CITATION). While the lead-lag relationships for macroprudential policies are hardly researched[8], they are likely to be at least as long. For instance, banks have one year to comply with increased capital requirements under the countercyclical capital buffer framework of Basel III (Basel Committee (2010)). In addition, data are reported with lags[9] and policymakers generally do act immediately on data developments, but observe trends for some time before they change policies (e.g. Bernanke (2004)).

Second, ideal EWIs should not signal crises too early. Booms are popular as money is made, output is growing and more and more households may get a foot on the property market ladder. Macroprudential policies responding to risks building-up in the background inevitable draw criticism. Combined with the argument that "this time it's different", this can undermine the effectiveness of macroprudential measures if they are introduced too early (e.g. Caruana (2010)). Taken together, these two requirements suggest that an ideal EWI should signal crises early enough, but not too early. For our empirical analysis, we argue that good EWIs issue a crisis signal at least 1 year before a crisis. Judging what it "too early" is more difficult. But to be conservative, we use at most a 5 year horizon.

An important additional requirement which has largely been overlooked in the literature concerns the persistency of EWI signals. As already discussed, in practice policymakers tend to observe data developments for some time before they gradually change policy instruments (e.g. Bernanke (2004)). In the context of monetary policy it also theoretically optimal for policymakers to be more cautious and respond less to new information, the noisier this information is (e.g. Orphanides (2003)).10 Rephrasing this for EWIs means that an indicator which fluctuates frequently between crises and no-crises signals is worse than an indicator delivering persistent signals.

It seems evident that the persistence of the signals is directly tied to the persistence of the underlying conditioning variables. This is, for instance, in line with Park and Phillips (2000), who find, in the context of regression based binary choice models, that policy is likely to manifest streams of little or intensive intervention when the explanatory variables are difference stationary. This suggests that variables which in past studies have been found to be useful for macroprudential policy may also be ones which display a high degree of persistence.[11] And this is what we find for all the variable analysed here (see Section 3.3) While this type of persistence may be benign from the policymakers' perspective, it can nevertheless have implications for both correct inference and the choice of estimator.

The preceding discussion on timings related issues has several direct implications for our preferred approach. First, we assume that a signal is correct if it forecasts a crisis in an

---

[8] Reference a few papers

[9] Typical reporting lags for an initial release of series discussed in this paper are: interest rates (0), equity prices (0), property prices (?) …..

[10] Gradual changes in interest rates are also theoretically optimal for other reasons….

[11] This conjecture, is consistent with the evidence in Borio et al. (2012), for instance, which suggests that the financial cycle is much longer than the conventional business cycle

interval over the next five years. This also addresses the uncertainty of dating crisis correctly and the difficulty of predicting the actual timing when a crisis will materialize. Moreover, we do not take any signals in the two years after the beginning of a crisis into account, as binary EWIs become biased if the immediate post-crisis period is included in the analysis (Brussiere and Fratzscher (2006)).12 This also has sound economic rationale: it makes no sense to predict another crisis immediately after one has materialized.

Second, we evaluate each indicator in terms of its ability to issue persistent signals. We do this by calculating the AUROC for each date in the forecast interval separately…

Third, we adopt a non-parametric signal extraction approach in the spirit of Kaminsky and Reinhart (1999) which may be more robust under double unit root variables than standard regression based models for binary choice. Also, the latter models are estimated to maximize a specific likelihood function that, to the extent to which it is subject to misspecification, can perform arbitrarily bad at specific points of the policymaker's loss function (Elliott and Lieli (2010)).

Perspective: interpretability and logic of the EWI signal

From a policy perspective, an ideal EWI does not only need to fulfil these statistical criteria, but has to be backed-up by a coherent, analytical framework which policymakers understand. Policymakers never rely purely on statistical tools, even for macroeconomic forecasting where the theoretical and empirical literature is far more advance. Instead they analyse a range of models and indicators and supplement it by judgement.[13] Findings in the literature show that practitioners value the sensibility of forecasts more than accuracy (Huss, 1987) and adjust forecasts if they lack justifiable explanations (Önkal-Atay et al (2009)). For an empirical strategy to find EWIs, this implies that simply data driven indicators, for example derived by a general to specific approach, are not suitable for policy purposes.

Ideally, the analytical framework would be based on one or several well established theoretical models, which are however not yet available for financial stability purposes. Even the most advanced models cannot account for crises, the main event we are chiefly interested in (e.g. Gertler…). Instead the more appropriate analytical framework in our view is in the tradition of Kindleberger (2000) and Minsky (1982), which see financial crises as the result of mutually reinforcing processes between the financial and real sides of the economy. In this view, financial imbalances are driven by, but also feed, an unsustainable economic expansion, which manifests itself in unusually rapid growth of credit and asset prices. As the economy grows, cash flows, incomes and asset prices rise, risk appetite increases and external funding constraints weaken. This, in turn, facilitates risk-taking. The financial system typically does not build up sufficient capital and liquidity buffers during benign economic conditions, when it is easier and cheaper to do so, in order to deal with more challenging times. At some point, imbalances have to unwind, potentially causing a crisis, characterised by large losses, liquidity squeezes and possibly a credit crunch.

---

[12] Cecchetti et al (2009) find that crises last nearly three years on average. In our sample, the minimum time between two crises in one country is five years. Thus by assuming that crises last only two years we bias our noise-to-signal ratio upwards, as only type 2 errors are issued during the quarters immediately following the end of the second year after an episode.

[13] See Lawrence et al (2006) for a survey on judgemental forecasting and its importance in practice.

3. Analysing potential EWIs

In this section, we construct and test a range of potential early warning indicators which fit the discussed policy requirements. As a first step, we focus on single variable indicators since using these to anchor potential policy actions has the advantage of being more transparent (see Drehmann et al (2011)). As will be shown, single indicators already provide very good guidance, leaving limited scope for incremental improvement through the use of multivariate approaches.[14]

Drehmann et al (2011) cover a wide range of potential indicators which could be used to anchor countercyclical capital buffers as one particular macroprudential tool. They consider macroeconomic variables, indicators of banking sector conditions and market indicators. Rather than repeating this exercise, we look at a small set of selected indicators capturing different aspects of the financial cycle.

In line with a Minsky (1982), Drehmann et al (2011) find that the best performing EWIs for banking crises are measures of credit and asset price booms. In particular credit developments are important, in line with the literature (Grouchinas and Obstfeld (2012), Reinhart and Rogoff (2008), Jorda et al (2011)). The single best indicator is the credit-to-GDP gap (see next section on data discussion). This is also the variable, which should act as the starting point of discussions about the level of countercyclical capital buffer charges according to the Basel Committee (2010).

More recently, Drehmann and Juselius propose the debt service ratio (DSR) as a useful early warning indicator. The DSR is defined as the proportion of interest payments and mandatory repayments of principal to income. It can be interpreted as capturing incipient liquidity constraints of private sector borrowers.

The second class of useful EWIs found by Drehmann are indicators of asset price booms, and here in particular the property price gap. They also assess real GDP growth, which performs poorly as an EWI. However, as a macroeconomic benchmark we include it in our analysis as well.

3.1 Data

We analyse quarterly time-series data from 27 countries. The sample starts in 1980 for most countries, and at the earliest available date for the rest. Table X in the Annex (to be completed) provides detailed definitions and sample coverage of the data.

The performance of the anchor variables for the countercyclical capital buffer is assessed against an indicator of banking crises. Admittedly, the dating of banking crises is not uncontroversial (eg Boyd et al 2009). We broadly follow the dating of crises in Laeven and Valencia (2008, 2010), but ignore such crises which solely involve, as well as were brought on by, banking sector losses on foreign loans.[15]

As binary EWIs become biased if the immediate post-crisis period is included in the analysis (Brussiere and Fratzscher (2006)), we do not take any signals in the two years after the

[14] Borio and Lowe (2002) and Borio and Drehmann (2009a) show that combinations of variables have somewhat better signalling properties for systemic financial distress than single indicators.

[15] Crises of the latter type were identified through use of information generously provided to us by national central banks. This results in XX crises being excluded from the sample. A full list of the remaining crisis dates is given in the Annex (Table X).

beginning of a crisis into account.[16] This also has sound economic rationale: it makes no sense to predict another crisis immediately after one has materialized.

Macroeconomic variables are collected from national authorities, the IMF international financial statistics and the BIS database. Property prices are based on BIS statistics and combine residential and commercial real-estate. They are available only for a smaller sample and shorter time period.

To measure credit in the economy, the literature generally relies on series from the monetary statistics measuring the credit provided by domestic banks to the private-non financial sector, such as the IMF-IFS. For several countries, this excludes important sources of credit to the economy, such as bond markets or cross-border loans. In the US, for example, bank credit accounts on average only for less than 40% of total credit in the last 10 years. And even more in more bank based systems such as France, this fraction amounts to only 70%. Drawing on a new data-base we therefore use measures of total credit to the private non-government sector based on a flow-of-funds concept as much as possible.

To derive the credit-to-GDP and the property price gap we measure the trend with a one-sided Hodrick-Prescott filter.[17] The backward-looking filter is run recursively for each period and the gap calculated as the difference between the actual value of the variable and the value of the trend at that point. Thus, a property price trend calculated in, say, 1988q1 only takes account of information up to 1988q1, and the GDP trend in 2008q4 takes account of all information up to 2008q4. This is an important practical constraint, as policymakers have to take decisions in real time and rely on data that are available at that point.[18] Before using any trend, we require at least eight years of information.[19]

The calculation of the Hodrick-Prescott filter involves a key smoothing parameter λ. Following Hodrick and Prescott (1981) it has become standard to set the smoothing parameter λ to 1600 for quarterly data. Ravn and Uhlig (2002) show that for series of other frequencies (daily, annual etc) it is optimal to set λ equal to 1600 multiplied by the fourth power of the observation frequency ratio. We set lambda for *all* the gaps to 400,000, implying that financial cycles are four times longer than standard business cycles. This seems appropriate, as crises occur on average once in 20 to 25 years in our sample. Drehmann et al (2011) explore different lambdas but find that lambda equal to 400.000 works best for early warning indicators. Equally, we could have used a time trend such as Grouchinas and Obstfeld (2012)).

Debt service ratios (DSR) are taken from Drehmann and Juselius (2012). Even though levels are surprisingly similar across countries and time despite different levels of financial

---

[16] Cecchetti et al (2009) find that crises last nearly three years on average. In our sample, the minimum time between two crises in one country is five years. Thus by assuming that crises last only two years we bias our noise-to-signal ratio upwards, as only type 2 errors are issued during the quarters immediately following the end of the second year after an episode.

[17] For property price gap, the difference between the actual data and the trend at each point in time is normalized by the trend in that period. For the credit-to-GDP gap, we simply take the difference between the actual data and the trend at each point in time.

[18] To test real time EWI properties, we would ideally use only the vintage of data policymakers had available at a particular moment. However, these data are not available for the large set of countries and the historical period we need to cover to establish robust indicators. It has also been shown that data revisions, at least for credit and GDP series in the US, are not of first order importance for our type of analysis (see Edge and Meisenzahl (2011)).

[19] Ideally, ten years of data would be better (eg, Borio and Lowe (2002)). But given that data are limited for some series in some countries, we chose a 8-year window to ensure sufficient observations.

development, some country differences persist, due to different rates of homeownership or different industrial structures. To account for this, we subtract 15-year rolling averages from the DSRs for our analysis.

3.2 Persistence of candidate variables

To assess the persistence of the potential indicator variables analysed in this paper, we estimate AR(k) processes for the levels and first differences of each variable and apply standard unit root test, as well as calculate the sum of autoregressive coefficients (see for instance, CITATION).[20] Table (to-be-added) reports the results. As can be seen from the table, the first differences of all of the key variables are highly persistent, sometimes statistically indistinguishable from unit roots processes. This implies that the levels of these variables display double unit-root behaviour.

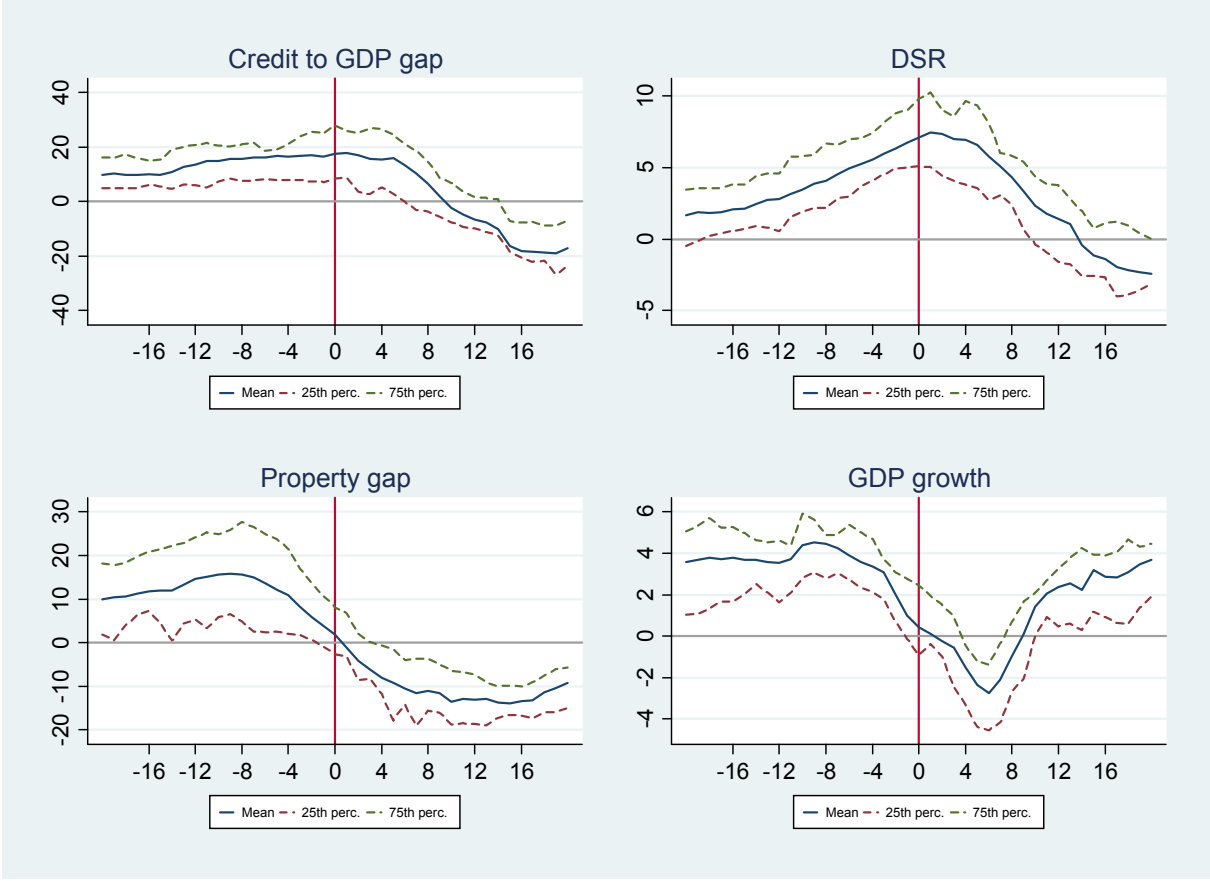
3.3 The behaviour of candidate variables around systemic crises

Before conducting our statistical tests, we look at the time profile for all four indicator variables around systemic banking crises. Graph 2 summarises the behaviour of the variables during a window of 20 quarters before and after the onset of a crisis (time 0 in the graphs). For each variable, we show the median (solid line) as well as the 25th and 75th percentiles (dashed lines) of the distribution across episodes. For the DSR, the credit-to-GDP gap and the property price gap, a value of zero corresponds to the average conditions outside the 40 quarter window, whereas this value is round 3% for real GDP growth.

The graph shows that the DSR, the credit-to-GDP gap and the property price gap are very high in the run-up to crises, albeit with different time profiles. The median DSR starts from a relatively low base and triples during the four years before a crisis, at which point it peaks. The credit-to-GDP gap, on the other hand, is already very high three to five years ahead of a crisis and rises much more slowly. The property price gap on the other hand, reaches its peak around 2 years before crises, after which it starts to decrease reaching nearly zero when crises hit on average.

[20] It is well known that the sum of autoregressive coefficients has a downward bias in small samples (CITATION). Hence, the actual persistence may be even larger than what is reported here.

Graph 2: Indicator variables around crises[1]



1 The horizontal axis depicts plus\minus 20 quarters around a crisis, which is indicated by the vertical line. The historical dispersion of the relevant variable is taken at the specific quarter across crisis episodes in the sample.

Real GDP growth shows a markedly different time-profile. It is around 4% four years prior to a crisis. It then starts to decline, with a slowdown gathering momentum in the year leading up to the crisis. Once the crisis materialises, GDP growth turns negative. After around two years, on average, the economy returns to its pre-crisis growth rate, suggesting that this growth rate is not particularly unusual. Interestingly, the 75th percentile shows that many crises are not preceded by any slowdown in output.[21]

For early warning purposes, Graph 2 suggests that the first three indicators should be useful, but that the DSR may perform better over shorter horizons and the credit-to-GDP gap as well as property price gap over longer ones. At best, a drop in GDP may issue a signal in the imminent quarters before a crisis.

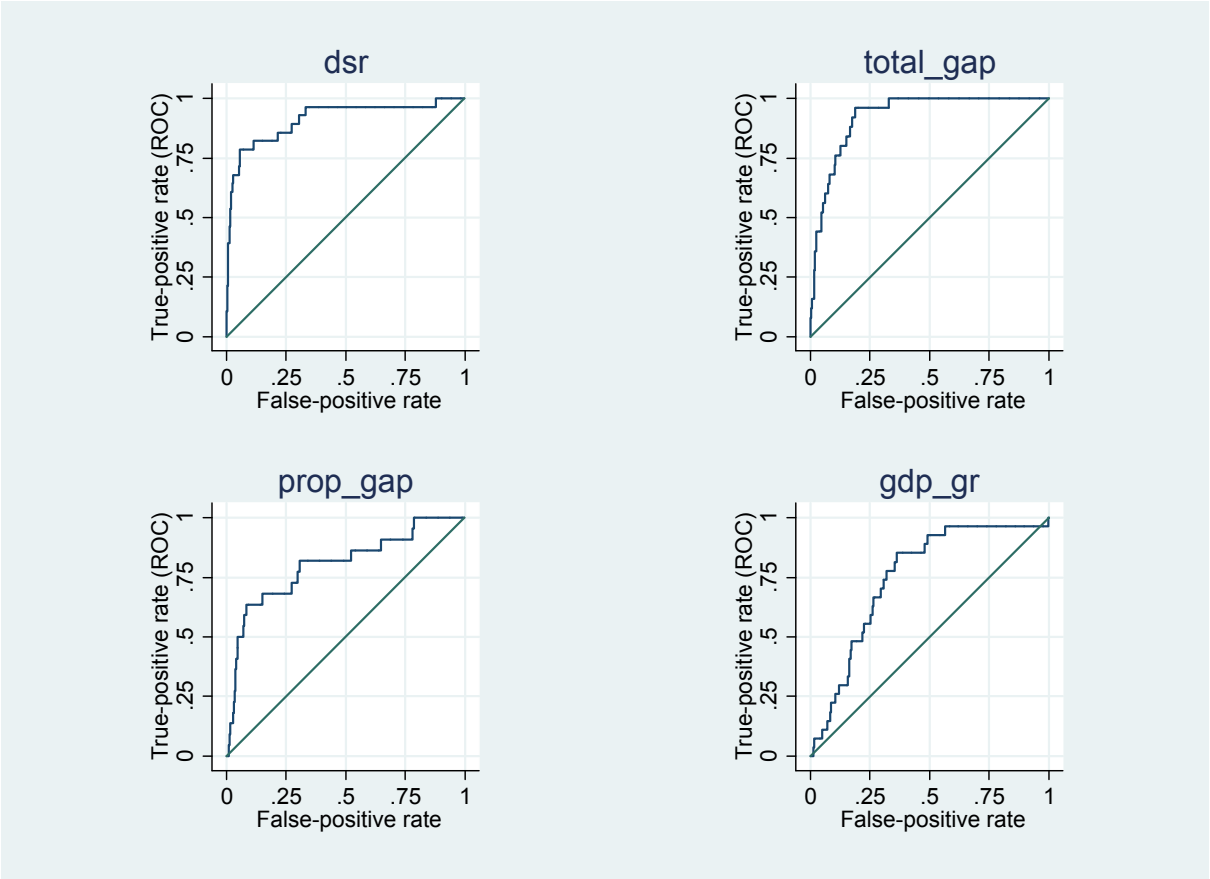
21 This provides a clear indication that models linking fragilities in the banking sector to weak macroeconomic fundamentals, such as macro stress tests, do not capture the dynamics of many crises (Alfaro and Drehmann (2009)).

3.4 ROC curves for 5 year flexible forecast horizons

Whilst the build-up of vulnerabilities may be detectable, the precise timing when crises materialize is not. If they were, market participants would anticipate this and developments would unravel (see e.g. Brunnermeier and Abreu (2003)). Kaminsky and Reinhart (1999) therefor use a three year horizon. The underlying idea of their approach is simple: a particular indicator will give a signal if it breaches a predefined threshold. A signal is correct if a crisis occurs at any point within the following three years. Otherwise, it is a false positive. In the first section, we follow this idea, but extend it to a five year horizon. In the second section we disentangle the forecast horizon into individual quarters to get some idea about the timing and persistency of signals.

Graph 3 shows the empirical ROC curves for all four indicators. It plots the false-positive rate versus the true-positive rate for all possible thresholds an indicator can breach. Looking at the graphs suggests that the DSR and the credit-to-GDP gap are the best performing indicators. The property and real GDP growth on the other hand provide less reliable signals.

Graph 3: ROC curves for indicator variables



This is confirmed by looking at Table 2, which shows the area under the ROC curve (AUROC) for the four indicator variables for two samples. The first includes all available data, the second is a homogenous data set considering only the periods when all variables are available.

AUROC is surprisingly high for all indicators. For the DSR and credit-to-GDP gap, AUROC is higher than 90%. The upper 95% confidence interval suggests that these are even close to 1, the value of the perfect indicator. The AUROC for the property price gap is somewhat lower with 80%. And depending on the sample, even GDP growth seems a good predictor as

the area under the curve is somewhere between 75% and 80% and statistically significantly different from zero. The latter AUROCs are also statistically significantly different from the AUROC of DSR.

All these areas are also high in relative terms. Jorda (2011) cites other studies showing that a widely used prostate-specific antigen (PSA) blood test has an AUROC of around 80% and that the S&P 500 has an AUROC of 0.86 for detecting in current time whether the economy is in recession or not.

Table 2: AUROC for indicators: 5 year flexible forecast horizon[1]

				95% confidence bands		
Sample	**Indicator**	**AUROC**	**std**	**low**	**high**	**obs**
All	DSR	0.91	0.035	0.83	0.97	2546
	Credit to GDP gap	0.93	0.016	0.89	0.96	2392
	Property price gap	0.80	0.054	0.69	0.91	2024
	GDP growth	0.75	0.042	0.66	0.82	2559
Homogenous	DSR	0.93	0.024	0.88	0.98	2010
	Credit to GDP gap	0.92	0.016	0.89	0.95	2010
	Property price gap	0.80*	0.057	0.69	0.91	2010
	GDP growth	0.79***	0.033	0.73	0.85	2010

[1] */**/** significantly different to AUROC of DSR

3.5 ROC curves for fixed forecast horizons

Much of the exceptional forecast performance shown in Table 2 is likely to be driven by the fact that we use a five year flexible forecast horizon, which considers a signal to be correct if a crisis occurs at any point within the next five years. In this sub-section, we therefore derive AUROC for each quarter individually within this period. Such an approach provides clear information about the time-profile and the persistency of each indicator variable.[22]

By doing so, we do not want, however, to implicitly suggest that the revealed average time pattern should be used to anchor policy very specifically, by for example suggesting that policy makers should for the first 4 quarters follow indicator X, then for 2 quarters indicator Y and so on. Future crisis will certainly not play out like this. Rather, our aim is to document

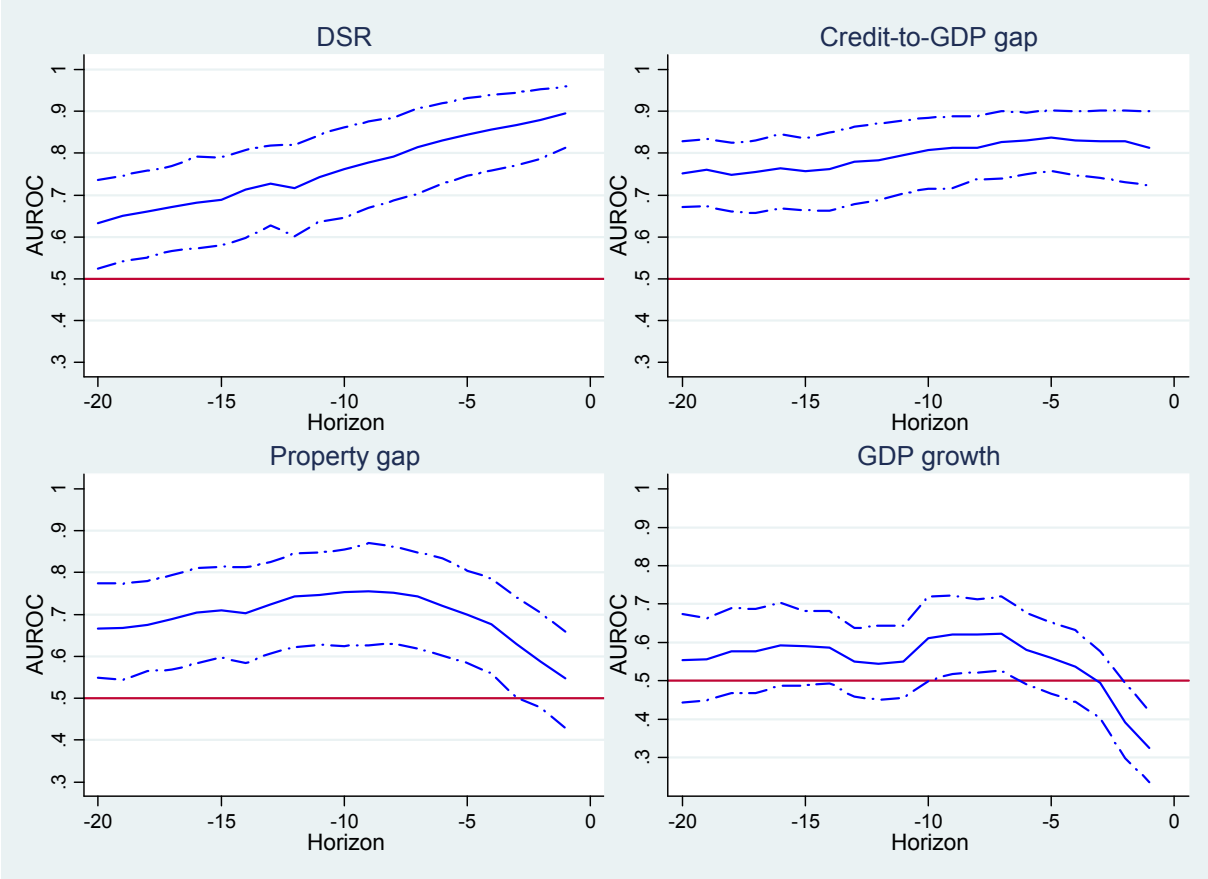
[22] Technically, when analysing a particular quarter, say quarter 5 before crisis, we only take account of signals issued in this quarter whiles ignoring all signals issues in the other quarters over the 5 year forecast horizon.

broad patterns, which help policy makers in determining the appropriate stance for macroprudential policies.[23]

Graph 4 highlights that the time-profile of the different indicator variables differs substantially. As anticipated from Graph 2, DSR's early warning properties are especially strong in the years preceding crises. One quarter before crisis, AUROC is close to 90% and not statistically different from the AUROC of the DSR using a full 5 year flexible forecast horizon. AUROC then drops continuously over the forecast horizon. However, the signalling quality of DSR remains very strong until year 5, when at the end it becomes insignificant.

The credit-to-GDP gap shows a markedly different pattern. Across the 5 years, AUROC fluctuates around 80%. It reaches its maximum of 84% in year 2 and the minimum of 75% in year 5.

Graph 4: AUROC over time[1]



[1] Solid line: AUROC, dotted lines: 95% confidence intervals. Horizon: quarters before a crisis. In percent.

[23] Would our results be used to guide policy decisions they would also be subject to the usual Lucas or Goodhart critiques. As Drehmann et al (2011) argue, though, the leading EWI properties would disappear by definition, if EWIs are well specified, their use would force banks to build up buffers to withstand the bust. Moreover, if, in addition, the scheme acted as a brake on risk taking during the boom, the bust would be less likely in the first place. However, the loss of predictive content per se would be no reason to abandon the scheme.

Graph 4 also indicates that the forecast performance of GDP growth was essentially an artefact of the flexible forecast horizon. Looking at individual quarters, its AUROC is essentially statistically insignificant, except in two quarter before crises. Here AUROC actually falls below ½. This implies that over the very short horizon a falling GDP is a useful early warning indicator for banking crises.

The property price gap is an intermediate case. As seen from Graph 2, it tends to fall ahead of crises and hence the signalling quality decreases and AUROC is insignificant over the short horizon. However, it provides valuable signals in the other years. From year 3 onwards, it is even marginally better than DSR.

4. Conclusions

In this paper, we argue that the statistical procedures used to construct and evaluate EWIs should be aligned with the requirements of policymakers. Among the considerations which seem particularly important are the relative trade-offs between type I and type II errors, as well as the timing and consistency of the EWI signals. Because data which could be used to pin down more narrow ranges for the policymakers' preferences are unavailable, we employ a technique which does not assume particular policy preferences – the ROC curve and its associated AUROC value – for assessing the performance of different EWIs. We also make novel use of these techniques to assess the performance of the EWIs over time.

We find that a new measure of private sector indebtedness, the DSR, as well as the credit-to-GDP gap, significantly dominate the other EWIs. Both of these variables also perform well over time, with the DRS dominating at shorter horizons and the credit-to-GDP gap dominating at longer horizons. The other EWIs have less stable temporal performance and are generally dominated by the DSR and the credit-to-GDP gap.

Summarizing, the ROC curve provides a powerful framework evaluating the predictive abilities EWIs in the absence of detailed knowledge of the policymakers' preferences. This framework can also easily be extended to assess the temporal performance of the indicators. Based on our results, we conclude that such assessments are particularly valuable for detecting various complementarities between different EWIs and, hence, constitute an additional useful tool for policymakers.

Bibliography

….