Reducing Model Risk in Early Warning Systems for Banking Crises in the Euro Area

Virginie Coudert* Julien Idier[ǂ]

Abstract

We assess the performance of early warning systems for detecting banking crises in the euro area and propose a new method to deal with model uncertainty. In a first step, we select a set of macro-financial risk indicators for their signaling ability among a large number of candidates over the period spanning from 1985:Q1 to 2009:Q4. Then, we run all the possible logit models including four of these indicators. We retain two sets of models: a small one only including models with all coefficients significant and with the expected signs, and a large set, obtained by relaxing the selection criteria. In a second step, we calculate the weighted average of the crisis probabilities estimated by the models belonging to the two selected sets. The weight given to each model is proportional to its usefulness at predicting crises either at the panel or the country-level. The simulations performed both over and out of the sample show that aggregating more models yields better results than relying on any single model or only a few of them, as model uncertainty is reduced. Performance is also enhanced by aggregating models' results with country-specific weights relatively to common panel-weightings.

JEL codes E52 G12 C58

Keywords: Macroprudential policy, Banking Crises, Early Warning Indicators.

* Bank of France, Financial Stability Directorate,virginie.coudert@banque-france.fr

ǂ Bank of France, Financial Stability Directorate, julien.idier@banque-france.fr.

1. Introduction

Since the 2008 crisis, economists as well as banking supervisors have given more credibility to the idea of a financial cycle. The mounting phase of the cycle, or "boom" period, is characterized by abundant credit and low risk aversion, which both feed agents' indebtedness and fuels the rise in prices of financial assets as well as those of real estate. Within several years of this regime, the built-up of debt and the bubbles in asset prices pave the way to the next crisis. Once bubbles burst and agents start to deleverage, banks are hit by the fall of asset prices that deteriorate their balance-sheet and the value of collateral for their loans. This sequential pattern has been been documented in the economic literature about the financial cycle (Claessens et al. 2011, Borio, 2012), although it is only since the 2008 crisis that governments have decided to tackle the issue by setting up a macroprudential policy. The aim of this policy is to contain the amplitude of the financial cycle, especially during its boom phase, by imposing more capital requirements on banks when credit growth is gauged excessive at the macroeconomic level.

For this policy to succeed, a major challenge is to be able to assess in real-time at which point of the cycle we stand. To do so, early warning systems (EWS) designed at predicting crises have been revived, particularly within the euro area (Alessi and Detken, 2011, 2014; Shin, 2013; Detken et al., 2014; Ferrari and Pirovano, 2015; Kalatie et al., 2015). First developed to predict the financial crises in the emerging countries (Frankel and Rose, 1996; Kaminsky et al, 1998, Kaminsky, 1999; Burkart and Coudert, 2002; Gourinchas et al. , 2001; Bussiere and Fratzscher, 2006), they have then been applied to large panels of advanced and emerging economies (Demirgüç-Kunt, A. and Detragiache, E. 1998, 2005; Eichengreen and Arteta, 2000; Bordo et al. 2001; Borio and Lowe, 2002). After the 2008 crisis, a number of studies, have also been devoted to assess if EWS could have been able to detect it (Borio and Drehman, 2009; Barrell et al., 2010; Frankel, J. A. and Saravelos, G. 2012, Bussière and Fratzscher, 2006 and 2008, Bussière, 2013).

One of the main difficulties in the setup of EWS comes from financial crises being (hopefully) rare events. This is why it requires considering a panel of countries in order to have a representative sample.. Another difficulty for the EWS used for the conduct of macroprudential policy stems from the long lag that the policy maker has to face before its policy becomes effective. A major challenge is to take the appropriate steps – i.e. raise the countercyclical capital buffers (CCyB) - early enough during the boom period. Being preventive is necessary for at least two reasons: (i) institutionally banks have 12 months to comply with the new level of the CCyB (when increased) and (ii) transmission channels are surrounded by uncertainty but some delays in pass-through are to be expected before capital requirements affect the banks' credit supply. Hence raising CCyB too late, especially just before a crisis is looming, would only worsen the situation for banks by being procyclical. Consequently, an essential prerequisite is to be able to predict crises early

enough during the boom period. That is why we need a forecast horizon of at least 1 to 3 years.

In this paper, we try to tackle these issues when constructing an EWS strategy for detecting the risks of banking crises in the euro area with the objective of using it for the macroprudential policy. We use a panel of 10 euro area countries over the 1985:Q1-2009:Q4 period and proceed in three steps. First, we select the most relevant univariate indicators to predict banking crises within a 1 to 3 years horizon.by adopting a signaling approach similar to Detken et al. (2014) . Second, we proceed to a multivariate analysis. To do this, we run all the possible logit models with four explanatory variables extracted from the previously selected indicators. Third, we select two sets of models on the basis of stringent or relaxed criteria and aggregate them with different weighting schemes reflecting either their performance at the panel or the country level. We then compare the results obtained through the different options both in and out of sample and proceed to robustness checks.

Our contribution to the literature is to propose a method to mitigate model uncertainty by aggregating a large number of models, once a pre-selection of relevant models has been carried out. The originality of the method is to make the set of models as well as their weightings in the aggregation vary over time. This allows us to address the problem of models instability and capture the evolving risk factors. The results outperform those obtained by any single model. By adopting different weights in aggregating the models, we are also able to derive country-specific early-warning systems, even if the logit models are estimated on a panel of countries.

The rest of the paper is organized as follows. Section 2 describes the data, sample and criteria to assess performances; it also provides a set of univariate indicators with predicting properties over the sample. Section 3 presents the multivariate econometric approach, based on aggregating sets of logit models with different weightings. Section 4 evaluates the results in-sample and out-of-sample of the different EWS strategies. Section 5 proceeds to robustness tests. Section 6 concludes.

2. Data, sample, and criteria to assess performances

To build an early warning system (EWS), we need two sets of data: the dates of the crisis episodes and a number of economic variables that possibly release signals by evolving specifically during the pre-crisis periods. The forecast horizon gives us the span of the pre-crisis period. These features apply both to the univariate and the econometric methods.

2.1 Crises, horizon of prediction and pre-crisis periods

The sample is made of quarterly data from 1985Q1 to 2009Q4 for ten countries (Austria, Belgium, Finland, France, Germany, Ireland, Italy, Netherlands, Portugal, Spain). By limiting the panel to the euro area members on a relatively short period, we expect that the sample

is made of economies with similar functioning. We then work on a balanced panel of N countries and T periods, N=10; T=96. To identify crisis periods, we follow a historical approach that considers lists of crises validated by their use in the economic literature. Most of these lists rely on expert surveys (for example see Laeven et Valencia, 2008, 2012). Here, we use the list updated by Babecky et al. (2012a)[1] that is extended up to the 2008 crisis (Table 1). The crisis dates for each country are identified by the same country's central bank. In particular, we note that all euro area countries have experienced a crisis in 2008, except Italy.

Table 1 : Periods of crisis in the 10 euro area countries, 1985-2009

Country	Crisis periods	Country	Crisis periods
Austria	2008Q1-2008Q4	Ireland	1985Q1
Belgium	2008Q1-2008Q4		2007Q1-2010Q4
Finland	1991Q1-1995Q4	Italy	1990Q1-1995Q4
France	1994Q1-1995Q4	Netherlands	2008Q1-2008Q4
	2008Q1-2009Q4	Portugal	2008Q1-2008Q4
Germany	2008Q1-2008Q4	Spain	2008Q1-2008Q4

Note : The dates of crises are those retained by Babecky et al. (2012a) for banking crises.

The dates of crises for each country are associated with their characteristic function C_{nt} equal to 1 if there is a crisis in country n at time t and 0 otherwise.

$$\begin{cases} C_{n,t} = 1 \text{ if there is a crisis in country } n \text{ at time } t \\ C_{n,t} = 0 \text{ otherwise} \end{cases} \quad (1)$$

However, we also need a pre-crisis variable as our aim is to identify variables that behave differently during pre-crises periods, not during crises. The horizon of prediction is set from 12 to 5 quarters, as adopted in Detken et al. (2014) or ESRB (2014). We are interested in characterizing the pre-crisis periods within this horizon $h \in H$, where H = [5, 12] is the set of quarters going from 12 to 5 quarters before the crisis. This rather long delay is justified by the delays needed for implementing macroprudential policies. Moreover, we account for the fact that periods just before crises and in their immediate aftermath can pollute the estimations. To avoid this, we remove them from the sample, marking them as missing values (NA).

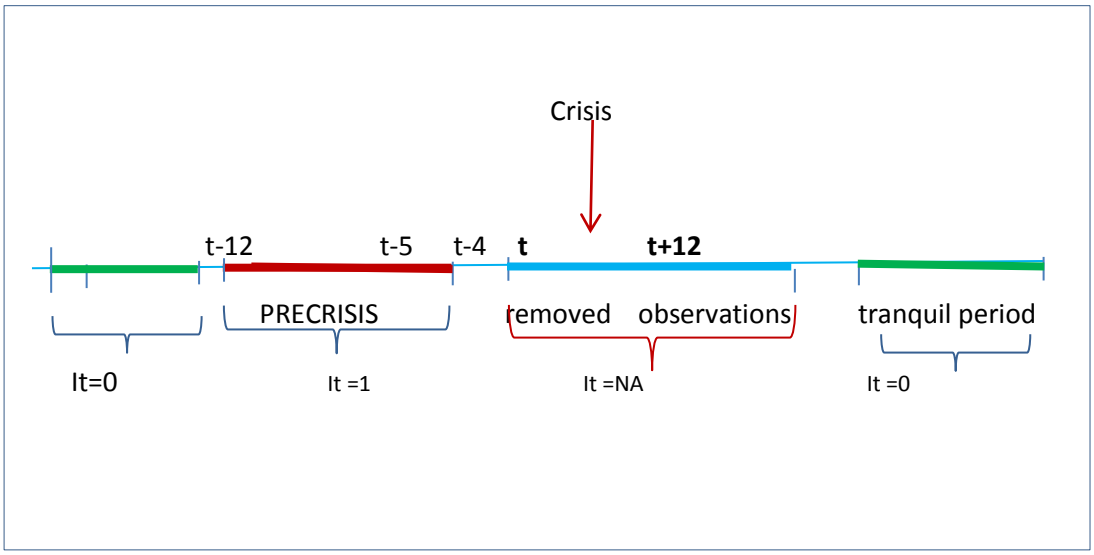
More precisely, we define the pre-crisis indicator $I_{n,t}$ of as:

[1] An improvement and updating of this database until 2015 is still in progress at the euro area level within Eurosystem working groups.

$$\begin{cases} I_{n,t} = 1, \quad \text{if } \exists\, h \in H=[5,12] \text{ such that } C_{n,t+h} = 1 \\ I_{n,t} = NA, \quad \text{if } \exists\, h \in [-12, ..,4] \text{ such that } C_{n,t+h} = 1 \\ I_{n,t} = 0, \quad \text{otherwise} \end{cases} \tag{2}$$

The pre-crisis period indicator $I_{n,t}$ equals 1 when a crisis occurs in country n within the H horizon; it is set to missing values (NA) around the crises (from 4 quarters ahead to 12 quarters after), and set to 0 in all the other periods, that are referred to as "tranquil periods". The construction of the precrisis indicator is illustrated on Figure 1.

Figure 1: Construction of the pre-crisis variable It



Although the data initially covers the 1985:Q1-2009:Q4 period, as all sample countries (but one) went through a crisis in 2008:Q1, the pre-crisis indicator has missing values from 2007:Q1 on. Indeed, as previously stated, we suppress observations up to one year ahead of a crisis, as well as the three subsequent years. This entails removing the years from 2007:Q1 to 2011:Q4 at least, and leaves us with a sample ending in 2006:Q4. Despite ending in 2006:Q4, the sample does take into account the 2008 episode, and we expect that the values of the variables observed during pre-crisis periods, 2005 or 2006, are able to detect the 2008 crisis. This strategy is in line with Detken et al. (2014).

2.2 Macrofinancial indicators and direction of risks

We consider a large set of economic variables, defined on the same sample as potential candidates for early warning indicators. This set accounts for the main risks on macroeconomics, credit, interest rates, real estate and financial markets. The choice is restricted to data available for all 10 considered countries over the whole time sample. All the series and their transformations are presented in Table A1 in the Appendix.

A preliminary step in the signaling process is to specify the direction of the risk. As we search for possible "booms" matching the pre-crisis periods, for a majority of our indicators, the risk increases with the high values of our series. Indeed, excessive values in credit ratios, asset or property prices favor the building-up of imbalances in the economy and are able to bring about financial bubbles that may unwind in future crises. Hence, the risk is on the right-tail of the distribution for all these series, the signal being emitted by the variable crossing its threshold upward. The only exceptions in our variables are the interest rates and the spreads, whose risk is the other way round. Indeed, low interest rates are more likely to be seen in "boom" periods as they enhance credit, deficits and fuel the rise in asset or house prices. Hence, the direction of risk associated with interest rate is on the left-tail of their distribution. To sum up, we are looking for upper thresholds for all variables but the interest rate related ones.

2.3 Assessing the relevance of an indicator by the signal method

The signal approach has long been used for forecasting currency and balance of payments crises (Kaminsky et al., 1998, Kaminsky, 1999) as well as for banking crises (Demirgüç-Kunt, and Detragiache, 1998, 2005) and financial crises (Christensen and Li, 2013). It amounts to counting the number of crises that burst once a given variable hits a critical threshold appropriately chosen. The signal method is key to EWS as it makes it possible to convert continuous variables into binary ones, called "signals" supposed to alert to crises. The method consists in finding the variables and their thresholds so that the thresholds are more frequently hit during the pre-crisis periods than during tranquil ones. We rely on this method for selecting our univariate indicators as well as our econometric models.

To assess the relevance of indicator *Z* and its threshold, the sample is decomposed in four categories of observations : (A) a signal is emitted and a crisis bursts at the H horizon, the crisis is well predicted; (B) a signal is emitted and no crisis occurs within H horizon, it is a false alarm (Type II error); (C) no signal is emitted and a crisis bursts within the H horizon, it is a missed crisis (Type I error); (D) no signal is emitted and no crisis occurs at the H horizon, the tranquil period is well predicted. The number of observations in each category is counted and denoted respectively A, B, C, D as in Table 2. Once the number of observations in each category is counted, one can easily calculate the performance ratios (last range of Table 2). The number of observations A, B, C and D can be calculated for the panel of all countries taken together; another possibility discussed in Section 2.7 is to calculate these numbers for each country.

For each value of θ, the performance of indicator Z can then be assessed by ratios such as the percentage of missed crises T1 (θ,Z) (type I errors), of false alarms T2(θ,Z) (type II errors). The noise to signal ratio T2(θ,Z)/(1-T1(θ,Z)) is also used to assess the global performance. The conditional probability of crisis if a signal is emitted is also useful to compare with the a priori probability of crisis (last column of Table 3). Indeed, we expect that a signal emitted by a relevant indicator at an appropriate threshold will increase the probability of crisis above the probability obtained without any information.

Strictly speaking, the denominations provided in Table 2 and both paragraphs above are not really accurate when the horizon of prediction H spans over more than one period, although they are generally chosen for their appealing simplicity. More specifically, the number A does not exactly refer to the "well predicted crises" but to the "well identified pre-crisis periods" meaning the observations both in pre-crisis periods (I_{nt} =1) and with Z_{nt} >θ. Consequently, the sum of (A+B) is not equal to the number of crises, but to 8 times it. Similarly the "missed crises" are the "missed pre-crisis periods" and the false alarms are the "tranquil periods wrongly identified as pre-crisis periods". However, for the sake of brevity and simplicity, we will continue to refer to terms such as "well-predicted crises" instead of "well-identified pre-crisis periods" in the following sections.

Table 2: Decomposition of the observations in the sample according to variable Z and threshold θ, performance ratios and probabilities of crises

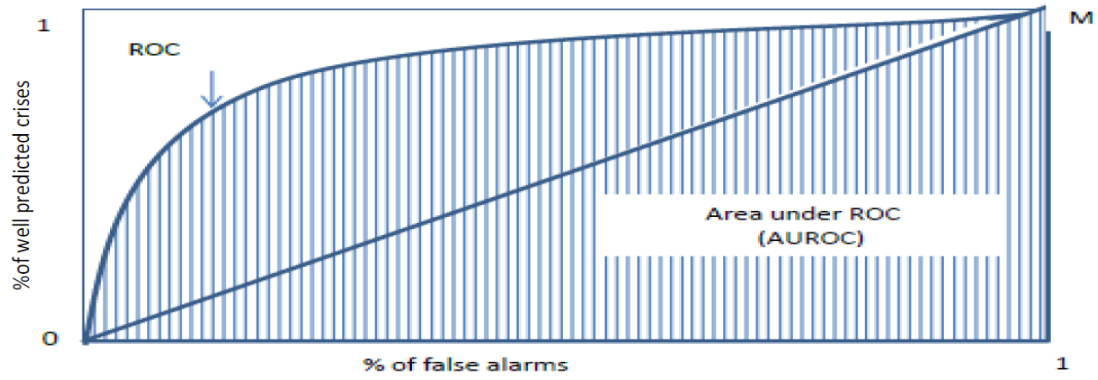
	$\exists h \in H$: a crisis occurs in *t+h*, in country *n*	$\forall h \in H$, no crisis occurs in *t+h*, in country *n*	Probabilities of crises
	Pre-cris indicator Int = 1	Pre-cris indicator Int =0	
Signal emitted $Z_{nt} \geq \theta$	"Crises well predicted" Nb =A	Error of Type 2 "False alarm" Nb=B	Probability of crisis if signal emitted A/(A+B)
No signal $Z_{nt} < \theta$	Error of Type 1 "Missed Crises" Nb=C	"Tranquil period well predicted." Nb= D	
Performance ratios	Proportion of "missed crises" T1(θ,Z) = C/(C+A)	Proportion of "false alarms" T2(θ,Z) = B/(B+D)	A priori probability of crisis (A+C)/NT

The threshold should be set by assessing the cost linked to the two types of errors. The trade-off is between (i) missing too many crises (T1) or (ii) wrongly predicting crises that do not exist (false alarms or T2). The lower the threshold, the more frequent the signal. Hence, by setting the threshold sufficiently low, one can easily predict the whole set of crises, but this may generate numerous false alarms. Inversely the higher the threshold, the less signals the indicator emits, at the risk of missing more crises.

When progressively lowering the threshold on the entire range of variation of Z, from its minimum to its maximum value, we can increase continuously the number of emitted signals. The percentage of well predicted crises then goes from 0% (with 0% of false alarms) to 100% (yielding also 100% of false alarms as all values emit a signal). This trade-off is represented on Figure 1, by shifting from point O, where no signal is emitted, to M, where a signal is emitted at each period. According to the threshold retained, the same indicator provides the whole range of results. The receiver operating curve (ROC) linking O to M represents the relevance of the indicator (Figure 1). As a relevant indicator should detect a

high percentage of crises with few false alarms, it should display a ROC well above the bisector.

Figure 2 : Trade-off between two types of errors



Note : Point O: the threshold is set at maximum of the indicator, no signal is emitted (0% of predicted crises, 0% of false alarms); Point M: the threshold is set at minimum of the indicator, the signal is emitted at each period.

The relevance of the indicator can then be measured by the area under the ROC, ie the AUROC, which is shown on the hatched area in Figure 1. By construction, the AUROC is always between 0 and 1, and would be equal to 0.5 for a random signal. Therefore, to be relevant, an indicator must have an AUROC greater than 0.5, otherwise, it gives no information. The advantage of the AUROC criterion is to be independent of a particular threshold. Consequently, we will use this criterion when selecting our indicators, as we will first eliminate all the variables with an AUROC smaller than 0.5.

2.4 Policy maker's preferences and determination of threshold

Although useful for preselecting indicators among a great number of potential ones, the AUROC criteria is not sufficient because it does not provide any particular threshold. Nevertheless, the thresholds are key to the EWS approach, as without them, one cannot say if signals have been emitted or not. This is why we need another approach to select the thresholds. One standard way is to minimize the policy maker's loss when making errors in predicting the crises.

To select the thresholds we minimize the policy maker's loss when making errors in predicting the crises. The policy maker's loss function L is defined as the weighted average of the two types of errors generated by the signal given by Z crossing a given threshold. The weighting parameter μ varying between 0 and 1 indicates the policy maker's preferences for avoiding type I errors compared to those of type 2.

$$L(\mu, \theta, Z) = \mu \mathrm{T1}(\theta, Z) + (1 - \mu)\mathrm{T2}(\theta, Z) \quad (4)$$

where *T*1(θ,Z) denotes the percentage of missed crises *T*2(θ,Z), the percentage of false alarms obtained for a given θ threshold, μ is a parameter that indicate the preference to avoid type 1 errors. The higher μ, the more costly it is to miss predicting a crisis. In this section, we set the μ parameter arbitrarily at 0.5. It will be allowed to vary in Section 5 to test for results sensitivity.

Once the μ parameter is fixed, the optimal threshold $\bar{\theta}$ can be easily determined by iteration through the minimization of the loss function.

$$\bar{\theta}(\mu, Z) = \operatorname{argmin}_{\theta} L(\mu, \theta, Z) \quad (5)$$

Once the threshold is optimized, we can determine the loss borne by the policy maker when using a given indicator Z associated with its critical threshold:

$$L(\mu, Z) = \quad L(\mu, \bar{\theta}, Z) \quad (6)$$

If a signal is emitted every time, the loss function will be equal to (1-μ); if no signal at all is released, the loss function will be equal to μ. Hence the policy maker has the possibility of lowering its loss to Min[μ, (1-μ)] independently of the information contained in any variable Z. Then, the "usefulness" $u(\mu, Z)$ of variable Z can be measured by the reduction in the loss function obtained by considering the signal emitted by Z instead of getting Min[μ, (1-μ)] with no information.

$$u(\mu, Z) = Min[\,\mu, (1-\mu)] - L(\mu, Z) \quad (7)$$

2.5 Evaluation at a country-level using panel-country data

When considering a panel of countries for selecting indicators, we are left with the choice of which type of information will seem relevant to the national policy maker. She may optimize the prediction by considering the value of the loss function obtained over the whole panel of countries (like in Equation 4), or over her own country only.

In this latter case, the loss function, denoted $L_n(\mu, \theta, Z)$, will be country-specific, depending on the two types of errors $\mathrm{Ti}(n, \theta, Z)$, i=1,2, obtained by the indicator Z for country *n* at θ threshold

$$L_n(\mu, \theta, Z) = \mu \mathrm{T1}(n, \theta, Z) \quad + (1-\mu)\mathrm{T2}(n, \theta, Z) \quad (8)$$

Where T1(n, θ, Z) (T2(n, θ, Z)) is the percentage of missed crises (false alarms) for country *n* by using the θ threshold for the Z variable. Hence the differences in the country-specific loss functions stem from the various relevance of indicator Z across countries (at the same θ threshold), not from different preferences of the policy makers, as μ is assumed to be the same across countries.

If crises were not rare events, one could optimize the θ threshold for Z variable by only using the country-specific observations. In practice, the low number of crisis events makes it impossible to derive a robust threshold from optimization of the loss function in a single country. That is why the optimal threshold has to be common to all countries.

Hence we consider that the optimized country's loss function is obtained with the $\bar{\theta}$ threshold previously optimized at the panel-level.

$$L_n(\mu, Z) = \quad L_n(\mu, \bar{\theta}, Z) \tag{9}$$

The usefulness of an indicator at the country-level is then deduced by:

$$u_n(\mu, Z) = Min[\,\mu, (1-\mu)] - L_n(\mu, Z) \tag{10}$$

2.6 Selection of indicators

We now select the univariate indicators among the initial set of 67 variables (Table A1 in the Appendix). We proceed in two steps. In the first step, we eliminate all variables whose performance in terms AUROC is smaller than 0.50 over the sample of 10 countries. This leaves us with 44 indicators that perform better than a random draw. For those indicators, we compute the critical threshold θ that minimizes the policy makers' loss function over the panel of 10 countries with μ=0.5.

We then require that the indicator performs well at a country-level. In the present case, as the EWS will be used for French macroprudential policy, we consider their performances at the French level. Hence, in the second step, we retain only the indicators with a positive usefulness for France. 32 indicators fulfill this criterion ; we rank them according to the usefulness of their signal in Table A2 in the Appendix. The 3-year change in monetary aggregate M3 and the total credit to GDP gap to its long term trend are the two best performing indicators. More generally, the results show that credit and money variables rank among the best indicators. Interest rates and real estate (prices and loans) also have a prominent place in the list.

3) The econometric approach: averaging logit models

The univariate approach developed above leads to identify several relevant indicators. We now combine them through logit models.

3.1 Logit models, benchmark models and the "Basel gap"

In the logit estimation, the left-hand side (LHS) variable is the same pre-crisis indicator variable $I_{n,t}$ as defined by Equation (2), with the same horizon of prediction. We also keep excluding the observations in the immediate neighborhood of crises. This strategy matches the one described in the ESRB Occasional paper on the operationalization of the CCB (Detken et al., 2014).

The basic logit equation to estimate is the following:

$$I_{n,t} = F\left[\alpha + \sum_{k=1}^{K_0} \beta_k X_{k,n,t-1} + \varepsilon_{n,t}\right] \tag{11}$$

where F is a logistic function, $\mathrm{F(Z)} = \frac{\beta_k e^Z}{1+e^Z}$, and K_0 is the number of variables to be included in the regression, α and β_k, parameters to estimate. The one–quarter lags on the explanatory variables $X_{k,n,t-1}$ do not reflect the horizon of forecast, since this is taken into account by the leads in the dependent variable (5 to 12 quarters ahead of the crises); they only account for the delay in the avaibility of data for the policy maker.

As the logistic function is monotonously increasing, and ranging between 0 and 1, it matches a cumulated distribution function. Hence the fitted value of the logit estimation can be interpreted as the estimated conditional probability of crises.

$$\hat{p}_{n,t} = \mathrm{Prob}\left[I_{n,t} = 1|\{X_k\}\right] = F\left[\hat{\alpha} + \sum_{k=1}^{K_0} \hat{\beta}_k X_{k,n,t-1}\right] \quad (12)$$

This probability of crises can be dealt with through the signaling approach just like a univariate indicator. We hence compute the policy maker's loss function and the critical threshold probability θ and we can also assess the performance of the model by calculating its usefulness.

The logit regression is run on panel data without any country effects. Indeed it is not possible to include fixed effects as some countries in the sample have experienced no crises during the period under review, hence their null dependent variable would be correlated with the fixed effect.

The key issue here is to select the relevant indicators X_k to include in the model among numerous potential variables. Putting all the potential variables in the regression at the same time would lead to multi-colinearity and biased results. Putting only several variables would be arbitrarily in the absence of a clear criterion. Given the high model uncertainty, it is reasonable to run a whole set of models before either picking the best ones or averaging results across a set of models, which is the strategy that we choose here.

Detken et al. (2014) among others have used the methodology described above to estimate a number of logit models as Equation (11) over a balanced sample of European Union (EU) countries. Their conclusions show that the best performing model over this pooled-sample includes 4 right hand-side (RHS) variables: the total credit to GDP gap, the debt service ratio, the equity prices (as a year-on-year change), the house price to income ratio.

We start by estimating a similar model. This model, that we will refer to as the benchmark model or Model 1 in the following, includes four explanatory variables: (i) the bank credit-to-GDP gap, equal to the difference between the bank credit to GDP ratio and its long term trend (ii) the residential property price-to-income ratio (annual change), (iii) the three-year real equity price growth and (iv) the debt service ratio defined as in Drehmann and Juselius (2012).

As a matter of fact, the credit gap is a key variable to consider when assessing the financial cycle (BCBS, 2010b; Drehmann et Juselius, 2014; Dembiermont et al., 2013; Drehmann and Tsatsanoris, 2014). As this variable stands out as the most reliable one in a number of

studies, it has been recommended by the BCBS in order to evaluate the appropriate level of the countercyclical buffers and hence dubbed the « standardized Basel gap ». However, we prefer the bank credit gap to the total credit gap, as it seems more related to the macroprudential instruments, the CCyB, acting on banks; however, the total credit to GDP gap is also considered among the other variables. We include the bank credit to GDP gap as an explanatory variable in all the logit models as a benchmark indicator.

3.2 Selecting the sets of models to aggregate

If we consider all indicators as possible RHS variables in Equation (11), we have a set of logit models $m \in \Omega = (1,\ldots, M)$. Each model m is defined by the set K_m of its RHS variables $\{X_{k,}\} \in K_m$, taken among the K possible candidates. The equation for model m is denoted by :

$$I_{n,t} = F[\alpha_m + \sum_{k \in K_m} \beta_{m,k} X_{k,n,t-1} + \varepsilon_{n,t}] \quad (13)$$

with $0 < |K_m| \leq K$.

As this strategy may lead to a large number of models to consider, we limit their number in two ways. First, we restrict the set of the possible RHS variables to the univariate indicators selected previously to which we add two other variables: the first one is the equity price 3-year growth because it is also significant in the benchmark model ; the second one is the annual real GDP growth, to be sure not to miss a macroeconomic signal. We also remove from the set of RHS variables all those with a trend, as the presence of a trend makes it more and more likely that a given threshold is crossed as the time goes on. This leads us to drop all simple ratios, like credit over GDP, and keep only their transformation, as growth rate or gap against trend. We then have a set of 29 possible RHS variables.

Second, among this set of 29 indicators, we only consider the combinations of 4 variables: the first one invariably being the bank credit gap and the three others being picked out of the 28 remaining indicators. This specification with 4 RHS variables is in line with the benchmark model. This setup implies estimating 3276 logit models. In order to get reasonable results and avoid any misspecification issues, we only retain models fulfilling a number of criteria. More specifically, we select the two following sets of models.

The first set Ω_1 is restricted to models meeting stringent criteria: (i) each of the four estimated coefficients has to be significant at the 95% level; (ii) each of them has also to match the expected sign regarding the risk the indicator it is supposed to gauge, for example, positive, for debt ratios, negative for interest rates (as discussed in Section 2). We systematically include the benchmark model in this set even if its coefficients are not significant.[2] Applying such stringent criteria drastically reduces the set of available models from 3276 to 6. The variables entering into these six models and the estimated coefficients are provided in Table 3 .

[2] However, if no model at all meet the stringent conditions, we will consider that the set Ω_1 is empty (See Section 4.2 below)

Table 3. The 6 models in the set Ω_1 selected with the stringent criteria, in-sample estimations

	Model 1	Model 2	Model 3	Model 4	Model 5	el 6Mod	Rate of appearance in selected models
GAP400_CB2GDP	0.0015 *0.013*	0.035 *0.014*	0.031 *0.014*	0.036 *0.015*	0.034 *0.016*	0.044 *0.013*	100%
D12_EQPR	0.007 *0.0016*	0.0044 *0.0015*	0.005 *0.0015*	0.0044 *0.0014*	0.0038 *0.0015*	0.004 *0.0014*	100%
SLOPE		0.264 *0.07*	0.174 *0.060*	0.226 *0.069*	0.178 *0.066*	0.149 *0.069*	83%
D4_RREP2INC	0.081 *0.019*	0.068 *0.023*					33%
DSR	0.191 *0.044*						17%
GAP400_RREPR			0.032 *0.016*				17%
D4_RREP2RENT				0.029 *0.012*			17%
GOLDEN1					0.091 *0.041*		17%
D4_M3R						0.041 *0.018*	17%

Note: Model 1 is the benchmark model in line with Detken et al. (2014), i.e. a model automatically selected in the selection process, even with non significant coefficients. The figures below the coefficients are the standard errors. GAP400_CB2GDP = Bank credit gap to GDP against the trend obtained with a hp filter 400 000; this variable enters in all models by construction; D12_EQPR is 3-year growth of equity prices; slope = yield curve slope 3M 10Y multiplied by (-1); D4_RREP2INC = yoy residential real estate price to disposable income; DSR = debt service ratio à la Drehmann et Juselius (2012); GAP400_RREPR = residential real estate prices gap against the trend obtained with hp filter 400 000; D4_RREP2rent = yoy residential real estate price to rent; GOLDEN1 = golden rule as real yoy GDP vs real 10-year yield; D4_M3R is yoy growth of M3.

The second set of models Ω_2 is larger, as it is selected through more relaxed criteria. It is made of all possible models with three in the four estimated coefficients significant at the 95% level and the expected sign. Consequently, one of the four variables has no constraint on its coefficient. As the criteria are more relaxed, the number of models is larger, amounting to 611 in the sample. Hence, the composite crisis probability (obtained from the aggregation of models that is explained in the next section) takes into account more heterogeneous risk indicators. By construction, the set Ω_1 is a subset of Ω_2.

3.3 The risk factors involved

One key question is to acknowledge which risk factors these selected models account for. To answer this question, we report the frequencies of occurrence of each variable among the two sets of models in Table 2. The RHS variables that appear in the selected logit

models can be considered as the most significant risk factors over the pooled-sample. Outside of the bank credit gap to GDP ratio that is included in all models by construction, one variable stands out by appearing in all the retained models: it is the 3-year change in equity price. By measuring variations over a 3-year period, this variable is able to capture the building-up of imbalances on the stock market. The slope of the yield curve is also a key variable as it enters in 83% of models; as the risk measured in this variable is left-tailed, the more risky situations are found with very low long term rates relatively to short ones. Then, a few other variables are retained in 17% of models to measure real estate risk, interest rates and growth of money aggregates. As expected, we retrieve the four variables highlighted in the benchmark model.

The performances of the models are rather satisfying over the pooled- sample. Their AUROCs range from 0.66 to 0.71 with a median of 0.68 when 6 models are retained with stringent criteria; between 0.59 and 0.83 with a median of 0.68 for the set Ω_2 of 611 models selected with the relaxed criteria.

Table 2: Statistical appearance of risk factors in the two sets of selected models Ω_1 and Ω_2 (*)

Name	Unit	Set Ω_1 restrictive criteria 6 models	Set Ω_2 relaxed criteria 611 models
Bank credit to non financial private sector	Real – % GDP – gap to long-term trend	100%	100%
Equity price index	Real, 3-y change - %	100%	30%
Slope of the yield curve	%	83.33%	22%
House price-to-income ratio	Y-o-y change	33.3%	15%
Debt service ratio, non-financial sector	%	16,67%	11%
Monetary aggregate M3	Real, y-o-y change - %	16.67%	19%
Residential property price	Real, gap to long-term trend	16.67%	8%
Interest rate gap to GDP (Golden rule)	%, real bond yield minus 1-year real GDP growth	16.67%	10%
Ratio of house price to rent price	y-o-y difference	16.67%	12%
Loans for house purchase	Real, 3-y change - %	0%	17%
Debt service to income ratio, non financial corporations	%	0%	16%
Monetary aggregate M3	Real, 2-y change - %	0%	15%
Monetary aggregate M3	Real, 3-y change - %	0%	12%
Interest rate gap to GDP (Golden rule)	%, real bond yield minus 3-year real GDP growth	0%	10%
Total Credit to Households	Real, 1-y change - %	0%	8%
Total Credit to non-financial Corporations	Real, 1-y change - %	0%	8%
Residential property price	Real, y-o-y change - %	0%	8%
Loans to for house purchase	Real, 1-y change - %	0%	8%
Residential property price	Real, 2-y change - %	0%	8%
Total Credit to Households	Real, 2-y change - %	0%	7%
Interest rate gap to GDP (Golden rule)	%, real bond yield minus 2-year real GDP growth	0%	7%
Total Credit to non-financial Corporations	Real, gap to long-term trend	0%	7%
3-month interest rate	%	0%	6%

Calculations: Banque de France. Note: (*) Set Ω_1, restrictive criteria : all four variables are significant at 95% with the expected sign; Set Ω_2, relaxed criteria : three in four variables are significant and with the expected sign. Unmentioned indicators did not appear in the selected models. In grey, the variables common with the benchmark ESRB(2014) model.

3.4 Two options for aggregating the models: usefulness at panel-level or country-level

There are several ways to proceed to this aggregation, as described in Holopainen and Sarlin (2015). For example, a strategy followed by Babecky et al (2012a) is to select the variables that are the most significant in the largest number of models (considering their Student statistics). To do that, they construct a "posterior inclusion probability" (PIP) for

each variable that is equal to the probability that the coefficient β_{mk} is significantly different from 0 in all models. Here our strategy relies on averaging the models results by giving more weight to the most performing ones, the performance being measured by the usefulness as detailed below.

Once the set of models $\overline{\Omega}$ has been selected (Ω_1 or Ω_2), we calculate the probability of crises of each model $m \in \overline{\Omega}$, denoted $\hat{p}_m$, as the fitted value of Equation (13) for country *n* at time *t* :

$$\hat{p}_{m,n,t} = F\left[\widehat{\alpha_m} + \sum_{k \in K_m} \hat{\beta}_{m,k} X_{k,n,t-1}\right] \quad (14)$$

We then are able to calculate the policy maker's loss function $L(\mu, \theta, \hat{p}_m)$ at the panel-level given the μ parameter and for any θ threshold:

$$L(\mu, \theta, \hat{p}_m) = \mu \mathrm{T1}(\theta, \hat{p}_m) + (1-\mu)\mathrm{T2}(\theta, \hat{p}_m) \quad (15)$$

where $\mathrm{Ti}(\theta, \hat{p}_m)$ is the ratio of type i (i=1,2) errors when $\hat{p}_m$ crosses the θ threshold.

By optimizing this loss function at the panel-level, we find the critical threshold $\bar{\theta}$, which is the cut-off probability to release a crisis signal.

$$\overline{\theta(\mu, \hat{p}_m)} = \operatorname{argmin}_\theta L\left(\mu, \theta, \hat{p}_m\right) \quad (16)$$

This allows us to calculate the usefulness of each model at the panel-level.

$$u(\mu, \hat{p}_m) = Min[\,\mu, (1-\mu)] - L(\mu, \hat{p}_m) \quad (17)$$

The usefulness can also be assessed at the country-level as indicated in Section 2.5. To do this, we calculate the country's loss functions $L_n(\mu, \overline{\theta}, \hat{p}_m)$ by applying the same critical threshold $\bar{\theta}$ as calculated at the panel-level in Equation (17).

We denote the usefulness of model *m* at the country level with a *n* subscript: $u_n(\mu, \hat{p}_m)$.

$$u_n(\mu, \hat{p}_m) = \mathrm{Min}[\,\mu, (1-\mu)] - L_n(\mu, \overline{\theta}, \hat{p}_m) \quad (18)$$

The method consists in averaging all the probabilities of crisis obtained from the selected models $m \in \overline{\Omega}$ by giving more weight to the most useful models. Therefore the weight of each model is proportional to its usefulness. As the usefulness of models can be assessed both at the panel-level and at the country-level, we use two alternative weighting schemes and therefore obtain two composite probabilities of crises. The first one $\hat{P}^P$ gives more weights to the best performing models at the pooled-level and the other one, $\hat{P}^C$, has its weights tailored at the country-level performance.

$$\hat{P}^J_{n,t} = \sum_{m \in \overline{\Omega}} w^J_{m,n} \hat{p}_{m,n,t} \quad \textit{for J=P,C.} \quad (19)$$

Where $w^J_{m,n}$is the weight given to model *m* for aggregating country *n's* estimated probabilities in option J, J=P or C; the index P refers to the pooled- level and C to the country -level.

The pooled weights $w_{m,n}^{P}$ are the same for all countries and depend on the usefulness of the model *m* over the pooled sample. [3]

$$w_{m,n}^{P} = w_{m}^{P} = \frac{u(\mu,\hat{p}_m)}{\sum_{m\in\bar{\Omega}} u(\mu,\hat{p}_m)} \tag{20}$$

The country-specific weights $w_{m,n}^{C}$ vary across countries and depend on the usefulness of the models $u_n(\mu, m)$ assessed separately over each country.

$$w_{m,n}^{C} = \frac{u_n(\mu,\hat{p}_m)}{\sum_{m\in\bar{\Omega}} u_n(\mu,\hat{p}_m)} \tag{21}$$

From the previous step, we get two aggregated series of crises probabilities: $\hat{P}^{P}$ and $\hat{P}^{C}$ obtained by averaging the selected models with their usefulness either at the pooled or the country-level. We then calculate the two thresholds to be applied to these probabilities by optimizing the policy makers' loss function at the panel-level in both cases.

$$\overline{\theta}\ (\mu, \hat{P}^{J}) = \ \mathrm{argmin}_{\theta}\, L\left(\mu, \theta, \hat{P}^{J},\right), J=P, C \tag{22}$$

The aggregation strategy presented above presents three main advantages. First, and most importantly, it mitigates model uncertainty by taking into account a number of different models. Second, it also makes it possible for countries to differ in terms of risk factors sensitivity, while mixing pooled and country-level information. Indeed the country-specific probability $\hat{P}_{n,t}^{C}$ also draws its legitimacy from the fact that all the models considered in the aggregation answer to criteria on a pooled-information basis (significance and sign of their coefficients) which ensures their validity over the whole panel. Third, the weight given to each model changes over time according to its usefulness, hence the weighting scheme can be updated continuously according to the time-varying performances of the selected models (if the exercise is done in real-time). This is a valuable property as risk factors are likely to vary over time. Of course, any models can be re-estimated on a regular basis, but the strategy presented here is more flexible: the coefficients of each model are not only re-estimated at each period; the set of selected models itself changes over time.

4 Assessing the performance of the econometric approach

We now check whether the aggregated probabilities of crises provided by the models are able to emit relevant signals of crises. To do so, we compare the signals obtained with the two aggregation strategies (pooled or country-level) and the two sets of models (Ω_1 or Ω_2). We begin by a standard over-the-sample evaluation then go on with out-of-the sample or "real-time" simulations.

[3] We restrict model selection to models with positive usefulness, since usefulness can be negative if one logit performs worse than a pure random model.

4.1. In-sample evaluation

Table 3 displays the results obtained in sample by aggregating the models over the two sets Ω1 or Ω2. Two major findings stand out from these results. First, performance is greatly improved by aggregating more different models. This is shown by the much better results obtained by averaging the models over the larger set Ω_2 when comparing the loss functions. Adopting a larger set of models, Ω_2 decreases by around 25% on average the value of the loss function for both options relatively to the small set Ω_1. Hence, it seems rationale to relax model selection criteria in order to bring about better results. Second, tailoring the models' weight on country-specific usefulness improves model performances when using a large set of models, while it yields about the same results with the small set of models. The advantage of these country-specific signals is that they also account for the pooled information, as the set of models involved has been selected on the basis of the significance and sign of their coefficients over the whole panel.

Table 3. In-sample results for the aggregated models, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme and the set of models, μ=0.5

Options for the weightings scheme: models' usefulness calculated at	*Small set of models Ω_1 (*)*			*Large set of models Ω_2 (**)*		
	T1	T2	*L*	T1	T2	*L*
Panel -level	0.4	0.200	*0,300*	0.28	0.21	*0,251*
Country-level	0.45	0.174	*0,312*	0.208	0.195	*0,202*

Note. (*) selected through stringent selection criteria (6 models); (**) selected through relaxed selection criteria (611models)

The better performance achieved by aggregating more models needs to be investigated further. Is it a random result or can it be checked and justified? To address this issue, we compare the former results with the performances achieved by each single logit model in Ω_1. Each of these models fulfills the condition of four indicators significant with the expected sign. Results show that single models have less good performances than the aggregated ones (Table 4). Only the benchmark model, model 1, slightly outperforms a combination of a small set of models (Ω_1) when the aggregation is made at the country-level. However, the value of the loss function obtained from each model is much higher than when a large set of models Ω_2 is averaged. In other words, no single model is able to do better than a large combination of models.

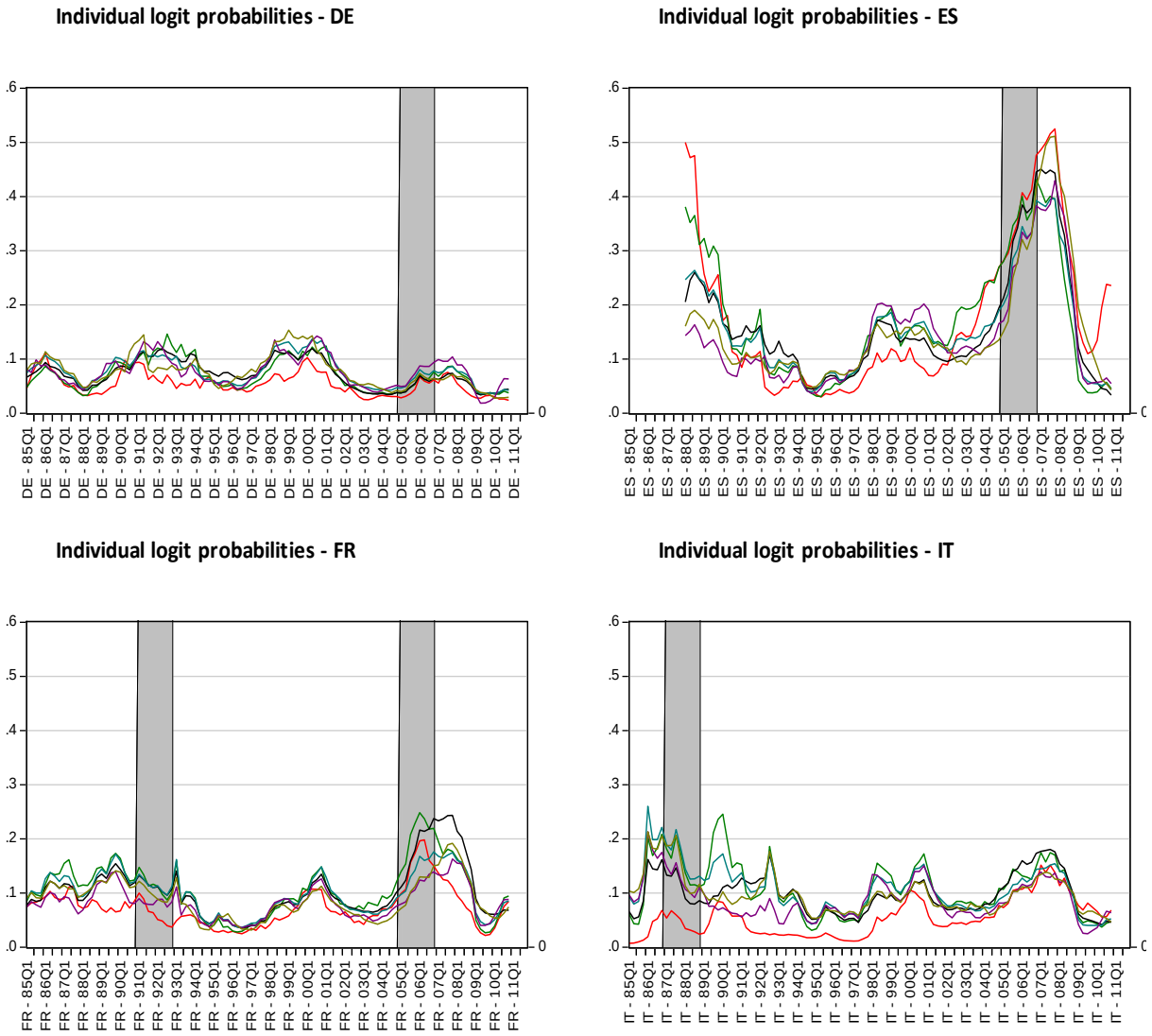
Table 4. Value of the loss function obtained from each individual logit model in the set Ω_1 (*), in sample

	Model 1(**)	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
Loss function	0.301	0.337	0.370	0.359	0.340	0.362

Note : (*) Loss function obtained from each individual logit model satisfying the stringent selection criteria (4 significant and expected sign coefficients). (**) Model 1 is the only one for which the Bank Credit-to-GDP gap is not significant, however not excluded given its benchmark status.

One way to explain the weaker performances obtained by single models compared to averaging results of models is to admit that increasing the number of models reduces model uncertainty. Figure 3 depicts the respective crisis probabilities estimated by the 6 models for France, Germany, Italy and Spain. Even if the 6 probabilities exhibit strong co-movements, there are notable differences across models that may lead to different assessments regarding the threat of a banking crisis. Indeed, the different combinations of factors point to different risks that could ultimately lead to a banking crisis. If we define uncertainty by the width of the range of probabilities given by the 6 models for a given date and a given country, one salient feature is that uncertainty is especially high when the probability of crisis increases. This peak in uncertainty when crises are about to burst clearly calls for a multiplicity of models to better monitor the risks of financial crises.

Figure 3: Crisis probabilities estimated with the 6 logit models in the set Ω_1 (in sample), for Germany, Spain, France and Italy.



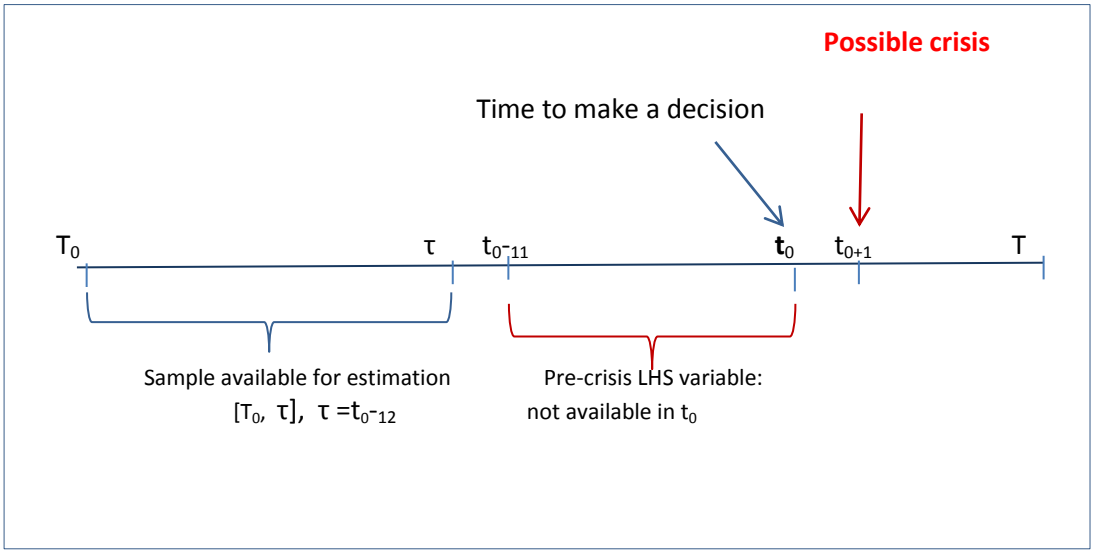
Note: crisis probabilities obtained with the 6 models in the set Ω_1 selected for their 4 significant and expected signed coefficients

4.2 Real-time evaluation of the monitoring strategy

4.2.1. Principles for the real-time simulations

To understand the lags a policy maker has to cope with when predicting a crisis, we have to remember that the LHS variable, being a pre-crisis indicator, is available only with a 12-quarter delay. Let us suppose that in time t_0, we are just a quarter ahead of a possible crisis; as we do not know it, the pre-crisis variable cannot be defined from t_{0-11} to t_0. This situation is depicted on Figure 4. Then the largest period for estimation spans from T_0, the beginning of the sample (1985Q1), up to t_{0-12}.

Figure 4: Available information in real-time at time t



To leave enough observations for the estimation, we start the out-of–sample exercise in 2003Q1 until 2009Q4. Let us describe thoroughly the different steps to estimate the first simulation as if it took place in t_0=2003q1. For the reasons indicated above, we have to end the first model estimation at date t_{0-12} =2000q1.

Let us call $\hat{p}^{\tau}_{m,n,t}$ the probability of pre-crisis obtained for time *t* with the model *m* estimated until time τ= t_{0-12}. It is expressed as:

$$\hat{p}^{\tau}_{m,n,t} = F\left[\widehat{\alpha^{\tau}}_m + \sum_{k\in K_m} \widehat{\beta^{\tau}}_{m,k} X_{k,n,t-1}\right] \qquad (23)$$

Where $\widehat{\alpha^{\tau}}_m$ and $\widehat{\beta^{\tau}}_{m,k}$ are the parameters obtained by estimating model *m* from T0=1985Q1 to τ. We thus get the predicted probabilities $\hat{p}^{\tau}_{m,n,\tau+h}$, h= 1 to 12. The last one τ+12 provides us with the needed prediction for t_0=2003q1, but we also look at the predictions for the shorter horizons. We then aggregate the probabilities obtained from the different models taking into account the usefulness of the models computed over the

sample [T0, τ] successively at the panel and the country-levels. Similarly, we estimate the thresholds over the same sample [T0, τ].

Once the first simulation is made for 2003q1, we proceed in exactly the same way for 2003q2 by adding one quarter to the estimation sample. We end the process in 2009q4. This provides us with 28 forecasts for 10 countries for each horizon (h=1 to 12 quarters ahead); in fact, the number of forecasts is smaller, as we have removed observations surrounding crises (since the dummy is set to NA in those periods as explained in Section 2). Taking into account crisis dates, we end up with 164 forecasts in total for the 10 countries among which we have 64 pre-crisis quarters. As the sample dates indicate, the out of sample evaluation is merely a test of the signaling properties of the models for the 2008 crisis.

4.2.2 Illustration for France and Germany

To illustrate the real time evaluation, we start by putting ourselves in the shoes of a policy-maker before the 2008 crisis, for example in 2005Q1 (then, 2006Q1, 2007Q1). She has to decide whether to implement or not macro-prudential tools in her country. Being in 2005Q1 means that the policy maker can only estimate the logit models up to 2003Q1 since the "pre-crisis" dummy used in these models is not defined after this date. To assess the model results under this real-time constraint, we successively proceed to the aggregation of the two sets models (Ω_1 or Ω_2) with the two options in 2005Q1, 2006Q1, 2007Q1.

First, a surprising result is that the set of models Ω_1 is an empty one, and therefore not usable. Indeed, up to 2008Q1, no model satisfies the stringent selection criterion (all 4 variables significant with the expected sign). Therefore, it is unavoidable for the policy maker to relax the model selection criteria as we have done in the previous in sample analysis and use the set Ω_2 (that includes all models with 3 in 4 significant variables with the expected sign).

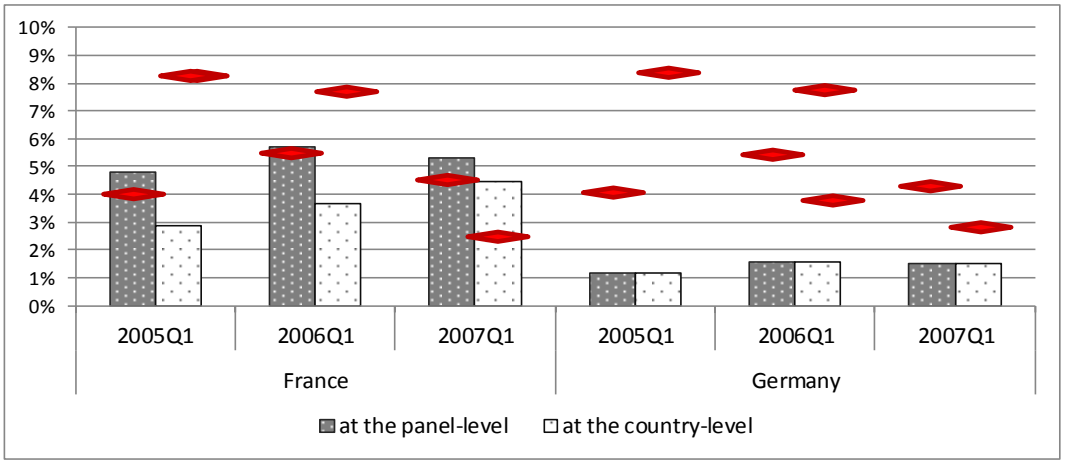
Second, on the contrary, the larger set Ω_2 is well furnished with models, as it includes 166 of them in 2005Q1, 386 in 2006Q1 and 632 in 2007Q1.

Figure 5 presents the corresponding aggregated probabilities of crisis given by these models for France and Germany when aggregated according to the models' usefulness at the panel and country-levels. As regards France, the panel-weighted probabilities give very satisfying results as the signal is released as early as 2005Q1; on the contrary, the signal is postponed until 2007Q1 if using the country-weighted aggregation.

In the case of Germany, two features stand out. First, both methods give the same probabilities of crises. This is due to Germany not having experienced any crisis prior to 2008; hence it is not possible to calculate the usefulness of an indicator over the German sample in this real-time estimation carried out to predict the 2008 crisis. Second, the method fails to deliver any clue of the coming crisis. This may come from the fact that macro financial indicators were not showing so large imbalances in this country prior to

2008, which is also in line with the 2008 crisis being much less severe in Germany than in some other countries of the sample.

Figure 5. Crisis probabilities and thresholds estimated in real-time for the two aggregation options on the set of models Ω_2 (μ =.5)



4.2.3 Overall results for out-of-sample evaluation

Applying the model selection criteria in real-time leads to the rise and death of models. This is a particular strong feature, showing that model uncertainty could impair the robustness of an early warning system over time. The number of selected models grows up from about 200 in the early 2000s to around 600 just before the 2008 crisis (Figure A1 in the Appendix). Once the imbalances leading to a severe crisis start to build-up, they affect a large set of risks, more indicators are turning red and multivariate signals become stronger.

Table 5 presents the out of sample results for the two aggregation options for the panel of countries between 2003Q1 and 2009Q4. Here, the panel-level weighting scheme seems to outperform the country-level aggregation, if we consider the loss function. This matches our previous finding for France, showing that the panel-weighted models would have been better to predict the 2008 crisis but contradicts the in-sample results that were better with country-level weightings. Consequently, the in-sample and out-sample results leave us with mixed evidence concerning the option to follow. As there are no clear-cut conclusions regarding the best aggregating strategy, we consider it useful to systematically run the two aggregating options to signal possible crises.

Table 5 Out-of-sample results for the aggregated models in the set Ω_2, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme, (μ=0.5), real-time simulations

Options for the weightings scheme: models' usefulness calculated at	*Aggregation on the large set of models Ω_2 (*)*		
	T1	T2	*L*
Panel -level	0.44	0.39	***0,42***
Country-level	0.64	0.28	*0,46*

Note. (*) selected through relaxed selection criteria.

5 Robustness checks

In this section, we provide complementary results regarding the value of the µ parameter and the way to calculate thresholds as well as robustness checks. First, we assess the results with different values of the policy maker's µ parameter, reflecting her level of risk aversion towards missing crises. This step is necessary because the µ parameter is very difficult to calibrate and may also vary over time. Second, we propose an alternative method for calculating the thresholds to apply on aggregated results: instead of optimizing the cut-off levels of the aggregated probabilities (as we have done in the previous sections), we now compute the weighted average of single models' optimal thresholds. Third, we check for the impact of the dummy crisis variable on the results: we thus estimate all the models again as well as the ensuing weightings with an alternative crisis dummy.

5.1 Alternative values for the µ parameter

For assessing the results with alternative values of µ, we rely on the same models' simulations, in and out-of-the sample, as previously. We therefore start from the same sets of models Ω_1 and Ω_2 for the in-sample results and Ω_2 for the real-time simulations. The only differences stem from (i) the way the models are aggregated because the usefulness of models changes according to the µ parameter; (ii) the optimal threshold that is lowered as the µ aversion to miss crises increases. This latter difference makes higher values of µ release more true signals at the cost of more false alarms.

We display the results of the in-sample simulations for the two alternative values of µ =0.6 and 0.7 while reminding the previous ones obtained with µ = 0.5 (Table 6). Both our main previous findings are comforted by these results. First, averaging the models' probabilities over a larger set of models provides much better performance regardless of the value of µ and the aggregation method. This is shown by the lower values of the loss functions obtained by aggregating the large set of models Ω_2 in the two last rows of Table 6. Second, the usefulness of models at the country-level provides a better weighting method in the large set of models, as it reduces the value of the loss function relatively to a panel-

weighing, whatever the value of μ. As regards the out-of-sample simulations, they provide the same kind of results as before: the panel-weighting method performs better for μ = 0.6, as for μ =0.5 (Table A4 in the Appendix). Nevertheless, the country-specific weightings provide better results for μ =0.7, which enhances the interest of this aggregation method.

Table 6. In-sample results for the aggregated models, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme, the set of models, and the *μ* parameter.

Weightings schemes : models' usefulness calculated at	*μ= 0.5*			*μ*= 0.6			*μ= 0.7*		
	T1	T2	*L*	T1	T2	*L*	T1	T2	*L*
Small set of models Ω_1 *(*)*									
Panel -level	0.40	0.200	*0,300*	0.325	0.275	*0,305*	0.075	0.66	*0,251*
Country-level	0.45	0.174	*0,312*	0.25	0.41	*0,314*	0.0	0.926	*0,278*
Large set of models Ω_2 *(**)*									
Panel -level	0.28	0.22	*0,251*	0.283	0.221	*0.258*	0.044	0.717	*0,247*
Country-level	0.208	0.196	0,202	0.059	0.351	0,176	0.0	0.449	0,135

(*) selected through stringent selection criteria (6 models); (**) selected through relaxed selection criteria (611models)

Table 7. Value of the loss function obtained from each individual logit model in the set Ω_1, in sample, with different values of μ.

	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
μ=0.5	0.301	0.337	0.370	0.359	0.340	0.362
μ=0.6	0.310	0.315	0.338	0.324	0.284	0.347
μ=0.7	0.293	0.270	0.282	0.269	0.225	0.274

We now consider the values of the loss function obtained by the stringently selected individual models Ω_1 (Table 7). This allows us to confirm the conclusion drawn above. All single models are outperformed by their aggregation on a large set, irrespective of the value of μ. The most disturbing point about these single models' results is that the best performing one changes according to the aversion μ of the policy maker to miss a crisis. This is particularly upsetting as the μ parameter is quite impossible to estimate and set at the discretion of the econometrician. The benchmark model, Model 1, that stands out as the most performing one for μ =0.5, is outperformed by Model 5, as soon as μ =0.6. More worryingly, it is the worst of the six when μ is set to 0.7, ie when the policy maker is keener to avoid crises. In these conditions, on the top of knowing that most single models are not stable through time, our doubts over the true value of the μ parameter makes it very problematic to rely on a single model to predict crises. This clearly highlights the great uncertainty surrounding the appropriate model to retain, and comforts us in our approach to aggregate a large set of models.

5.2 An alternative method for setting critical thresholds

Up to now, we have set the two thresholds for the aggregated probabilities P^P and P^C by optimizing the loss function at the panel-level. In this section, we adopt another method for setting the thresholds which mirrors the way we have constructed the aggregated probabilities.

More specifically, we rely on the same average probabilities P^P and P^C but only modifies the cut-off levels that release signals. To do so, we proceed in two steps. First we calculate all the critical thresholds $\bar{\theta}(\mu, \hat{p}_m)$ for the probabilities $\hat{p}_m$ obtained for all models m by optimizing the policy makers' loss function at the panel level. Second, we aggregate all the models' thresholds using either the weighting scheme resulting from the panel or country-level models' utility.

To derive the new "panel weighted threshold", $\tilde{\theta}(\mu, P^P)$ applied to the panel-level probability we hence calculate the average of the models' thresholds by weighting them with the panel-level weights w_m^P defined in Equation (21):

$$\tilde{\theta}(\mu, P^P) = \sum_{m \in \bar{\Omega}} w_m^P \bar{\theta}(\mu, \hat{p}_m) \quad (24)$$

Turning to the country-weighted threshold, $\tilde{\theta}_n(\mu, P^C)$ applied to P^C, we define it as the

average of models' thresholds weighted by the models' country-specific weights $w_{m,n}^C$ defined in Equation (22)

$$\widetilde{\theta}_{\mathrm{n}}(\mu, \mathrm{P^C}) = \sum_{\mathrm{m} \in \bar{\Omega}} \mathrm{w_{m,n}^C}\, \overline{\theta}(\mu, \hat{\mathrm{p}}_{\mathrm{m}}) \quad (25)$$

In the former setting, the two cut-off probabilities were common to all countries. In this new framework, the country–weighted threshold is allowed to vary across countries, in order to better reflect the relevance of the different models for each country.

The results obtained for these new thresholds in the sample are displayed on Table 8. The two main outcomes found previously are comforted by this exercise. First, aggregating models on a larger set of models gives better risk predictions. Second, the country-weighted aggregation improves upon the results relative to the panel-weighted one.

Another issue is to gauge these results relatively to those obtained previously through optimizing the thresholds. To do this, we compare the loss functions found on Table 8 with those reported on Table 3. At the panel-level, this new method of setting thresholds definitely underperforms the former one. This is not surprising, since the former method relied on an optimized threshold, so no other threshold is able to give better results on the loss function at least in the sample. However, at the country-level, the country-specific cut-offs outperform the optimized one, because the optimization was made under the constraint of a single level for all countries. Consequently, tailoring both the crisis probability and its cut-off at the country-level appears to be a valuable approach to account for heterogeneity. It is therefore worthwhile to implement this alternative method.

Table 8. In-sample results for the aggregated models, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme and the set of models, μ=0.5, for alternative aggregated thresholds (1)

Options for the weightings scheme: models' usefulness calculated at	*Small set of models* Ω_1 *(*)*			*Large set of models* Ω_2 *(**)*		
	T1	T2	*L*	T1	T2	*L*
Panel -level	0.275	0.385	*0,33*	0.28	0.24	*0,26*
Country-level	0.187	0.426	*0,302*	0.134	0.256	*0,195*

Notes. (1) The alternative thresholds are calculated by averaging the models' critical thresholds with either panel-weighted or country-weighted utilities; (*) selected through stringent selection criteria (6 models); (**) selected through relaxed selection criteria (611models)

Turning to real-time simulations to predict the 2008 crisis, we face the same obstacles as previously described and results are still blurred (Table A.5 in the Appendix). First, the country-level aggregation offer equivalent risk prediction for high values of μ, but not for μ=0.5. Second, contrary to the in-sample simulations, the alternative thresholds do not enhance the performances, neither at a panel nor at a country-level.

5.3 Robustness checks over the crisis dummy variable

As the results are contingent on the crisis episodes recorded in the dummy variable, we proceed to a robustness check by employing an alternative dummy variable. We now use the ESRB crisis dummy as in ESRB (2014) and run the in and out of sample estimations again with this new dependent variable. One key difference with the former crisis dummy is that neither Austria, Belgium nor Germany is supposed to have experienced a crisis in 2008 in this new setting. Otherwise, datation is similar, there were crises in 1991Q1-1992Q2 in Finland; 1993Q3- 1995Q3 and 2008Q3- 2010Q4 in France; 2000Q1-2003Q4 in Germany; 2008Q3- 2010Q4 in Ireland; 1994Q1- 1995Q4 in Italy but not in 2008 as previously; 2002Q2-2003Q4 and 2008Q3- 2010Q4 in the Netherlands; 1999Q1- 200Q1 and 2008Q4- 2010Q4 in Portugal; 1978Q1-1982Q3 and 2009Q2- 2010Q4 in Spain.

The in-sample results reinforce those previously found (see Table A6 in the Appendix). First, aggregating a large set of models obtained with the relaxed criteria yields much better results than restricting the set of models to stringent criteria. Second, using a country-weighting scheme to aggregate the models also improves the predicting performance for both sets of models, although it was true only for the large set with the former dummy.

Interestingly, the real-time simulations yield much better predictions for the 2008 crisis than those performed previously with the former dummy variable (Table A7 in the Appendix). Besides the fact that the set of models selected with stringent conditions is no longer empty, the loss function is lower for all values of μ. This can be seen when comparing the results with the former ones reported on Table A4. In particular, the fact that we were not able to forecast a crisis in 2008 in Germany with the previous dummy now

turns to be a good thing, for the 2008 observations that were tagged as crises with the previous dummy for this country are classified as tranquil periods with the new one. As a matter of fact, it is beyond the scope of this paper to decide which crisis dummy is the more appropriate, for this depends on the severity of crises assessed at the country-level.

In addition, the real-time simulations obtained with this alternative crisis dummy actually comfort our previous conclusions found with the in-sample results, which contrasts with the blurred outcomes drawn from the former real-time simulations. First, aggregating probabilities over a greater number of models heightens the signaling power (except for μ=0.5 at the country-level). Second, country-weighted aggregation outperforms the panel-level weighting. Third, the number of models involved in the aggregation tends to rise in 2006 and 2007 when approaching financial turmoil.

6. Conclusion

In this paper, we present a monitoring strategy for bank crises, based on early warning properties of indicators. This strategy takes into account numerous risk factors. One main difference with the related literature is that we rely on a large number of models, instead of a single one.

After selecting a set of risk indicators on the basis of their abilities to predict the banking crises in 10 euro area countries, we run all possible logit models combining four of these factors. Once the models have been estimated over the panel of countries, we select two sets of them: a small one following a stringent criterion, restricted to those with all variables significant and with the expected sign, as well as a larger set obtained through relaxed criteria, requiring only three variables in four being significant and with the expected sign. We then proceed with a weighted average of all the probabilities estimated by the different models across the two sets. To do so, we set the models' weights as proportional to their usefulness, which is a measure of their performance at predicting crises. The more useful is a model, the heavier its weight in the aggregated result. As the performance of models can be assessed either at the panel-level or at the country-level, we propose two options for the weighting scheme: one common to all countries, based on the usefulness of the models to predict crises on the whole panel; the other one, country-specific, resulting from the usefulness at the country-level.

Four main features stand out from the paper. First, aggregating a large number of models greatly improves the signaling performance over the sample – the loss function is reduced by 25% on average compared to the best performing model. In addition, averaging models allows us to avoid the unpleasant consequences of models' instability through time. Indeed, our real-time simulations show that the best performing model not only varies over time, it also depends on the policy maker's aversion to miss predicting a crisis, which is an unobserved parameter. On the whole, averaging models enables us to mitigate the uncertainty surrounding any single model.

Second, aggregating numerous models also appears the best strategy for the real-time simulations. Indeed, when we have estimated the models to replicate the policy maker's conditions before the 2008 crisis, we found that no model at all had its four variables significant with the expected signs at that time. Hence, retaining models on the basis of stringent criteria would not have been possible in real-time. Therefore, the only way is to take into account a large number of models selected with relaxed criteria. As a matter of fact, the results obtained using a large set of models are quite satisfying to predict the 2008 crisis at a reasonable horizon in most countries. Accounting for all possible risk factors hence appears as a good strategy in troubled times, when the sources of risk are evolving.

Third, we account for different risk factors across countries by tailoring country-specific weightings when aggregating the models, while we still use all the information at the panel-level to estimate the models. This strategy, mixing pooled and country level, is consistent with both the fact that countries differ in terms of risk factors sensitivity, and that estimation is improved by considering a panel of countries.

Fourth, the approach also enables us to address the issue of risk factors changing over time by allowing for flexible weighing schemes and changing sets of models. Indeed, in the real-time simulations, we continuously update the weightings and the sets of models according to their time-varying performances. This is a valuable property as risk factors are known to vary over time. Overall, once model uncertainty is acknowledged, we rely on a strategy involving the most possible risk factors at each time, while accounting for changes in these risk factors and their weightings over time.

References

Alessi, L. and Detken, C. 2011. Quasi Real Time Early Warning Indicators for Costly Asset Price Boom/bust Cycles: A Role for Global Liquidity. European Journal of Political Economy, 27(3), 520–533.

Alessi, L. and Detken, C. 2014. Identifying excessive credit growth and leverage.ECB WP 1723.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídová, K. and Vašíček, B. 2012a. Banking, Debt and Currency Crises : Early Warning Indicators for Developed Countries. ECB WP 1485.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídová, K. and Vašíček, B. 2012b. Leading Indicators of Crisis Incidence: Evidence from Developed Countries. ECB WP 1486.

Barrell, R., Davis, E. P., Karim, D. and Liadze, I. 2010. Bank Regulation, Property Prices and Early Warning Systems for Banking Crises in OECD Countries. Journal of Banking & Finance, 34(9), 2255–2264.

Basel Committee on Banking Supervision, BCBS 2010a. Basel III: A global regulatory framework for more resilient banks and banking systems. http://www.bis.org/publ/bcbs189.pdf

Basel Committee on Banking Supervision, BCBS 2010b. Guidance for national authorities operating the countercyclical capital buffer. BCBS Paper No. 187.

Bordo, B Eichengreen, D Klingebiel and M S Martinez-Peria (2001):"Financial crises: lessons from the last 120 years", Economic Policy, April.

Borio, C. 2012. The financial cycle and macroeconomics: what have we learnt? BIS Working Papers No 395

Borio, C and M Drehmann. 2009. Assessing the risk of banking crises – revisited, BIS Quarterly Review, March, pp 29–46.

Borio, C. and Lowe, P. 2002. Assessing the Risk of Banking Crisis. BIS Quarterly Review,December, 43–54.

Burkart, O. and Coudert, V. 2002. Leading indicators of currency crises for emerging countries. Emerging Markets Review, Elsevier, vol. 3(2), pages 107-133, June.

Bussière, M., 2013. "In Defense of Early Warning Signals," Working papers 420, Banque de France.

Bussiere, M. and Fratzscher, M. 2006. Towards a New Early Warning System of Financial Crises. Journal of International Money and Finance, 25(6), 953–973.

Bussière, M. & Fratzscher, M. 2008. "Low probability, high impact: Policy making and extreme events," Journal of Policy Modeling Volume 1, January–February 2008, Pages 111–121.

Christensen, I. and Li, F. 2013. "A Semiparametric Early Warning Model of Financial Stress Events" Bank of Canada Working Paper n°2013-13.

Claessens, S, M A Kose and M E Terrones. 2011. Financial cycles: What? How? When? IMF Working Paper WP/11/76.

Dembiermont, C, M Drehmann, and Muksakunratana, S. 2013 : "How much does the private sector really borrow - a new database for total credit to the private nonfinancial sector", BIS Quarterly Review, March.

Demirgüç-Kunt, A. and Detragiache, E. 1998. The Determinants of Banking Crises in Developing and Developed Countries. IMF Staff Papers, 45(1), 81–109.

Demirgüç-Kunt, A. and Detragiache, E. 2005. Cross-Country Empirical Studies of Systemic Bank Distress: A Survey. IMF Working Paper No. 05/96.

Detken, C. et alii, 2014. Operationalising the countercyclical capital buffer: indicator selection, threshold identification and calibration options. ESRB Occasional Paper No. 5.

Drehmann, M. et Juselius, M., 2012. Do debt service costs affect macroeconomic and financial stability?, BIS Quarterly Review, September.

Drehmann, M. and Juselius, M. . 2014. Evaluating early warning indicators of banking crises: Satisfying policy requirements. International Journal of Forecasting, vol. 30(3), pages 759-780.

Drehmann, M. and Tsatsaronris, K, 2014. The credit-to-GDP gap and countercyclical capital buffers: questions and answers, BIS Quarterly Review, March 2014

Eichengreen, B., Arteta, C., 2000. Banking crises in emerging markets: presumptions and evidence. Centre for International Development and Economics Research Working Paper, C00-115, August.

Ferrari, S. & Pirovano, M, 2015. Early warning indicators for banking crises: a conditional moments approach. MPRA Paper 62406, University Library of Munich, Germany

Frankel, J. A. and Rose, A. K. 1996. Currency Crashes in Emerging Markets: An Empirical Treatment. Journal of International Economics, 41(3–4), 351–366.

Frankel, J. A. and Saravelos, G. 2012. Can Leading Indicators Assess Country Vulnerability? Evidence from the 2008–09 Global Financial Crisis. Journal of International Economics, 87(2), 216–231.

Gourinchas, Pierre-Olivier, Valdes, Rodrigo, Landerretche, Oscar, 2001. Lending booms: Latin America and the world. NBER Working Papers 8249.

Haut Conseil de Stabilité Financière (HCSF). 2015, Le coussin de fonds propres contra-cyclique: procédure de mise en œuvre. September.

Holopainen, M and P. Sarlin 2015. Toward robust early-warning models:A horse race, ensembles and model uncertainty, Bank of Finland Working Paper n°6.

Kalatie, S. Laakkonen, H. and Tölö, E. 2015, Indicators used in setting the countercyclical capital buffer. Discussion Papers 8/2015. Bank of Finland.

Kaminsky, Graciela, 1999. Currency and banking crises: the early warnings of distress. IMF Working Paper No. 99/178.

Kaminsky, G. L., Lizondo, S. and Reinhart, C. M. 1998. The Leading Indicators of Currency Crises. IMF Staff Papers, 45(1), 1–48.

Laeven, L. and Valencia, F. 2008. Systemic Banking Crises: A New Database. IMF Working Paper No. 08/224.

Laeven, L. and Valencia, F. 2012. Systemic Banking Crises Database: An Update. IMF Working Paper No. 12/163.

Schüler,Y., Hiebert, P. and Peltonen, T. 2015 Characterising the financial cycle: a multivariate and time-varying approach, ECB WPS 1846. No 1846 / September 2015 No 1846 / September 201

Shin, H. 2013. Procyclicality and the Search for Early Warning Indicators. IMF Working Paper No. 13/258.

Appendix. Table A1. List of indicators tested

Indicators	Transformation	Source
Total credit to non financial private sector	Real – % GDP Real – % GDP – gap to long-term trend y-o-y change, 2-y change, 3-y change - %	BIS
Total credit to non financial firms	Real – % GDP Real – % GDP - Gap to long-term trend	BIS
Total credit to households	Real – % GDP Real – % GDP - Gap to long-term trend y-o-y change, 2-y change, 3- change - %	BIS
Banking credit to the private sector	Real – % GDP Real – % GDP - Gap to long-term trend y-o-y change , 2-y change, 3-y change - %	ECB
Loans for house purchases	Real – % GDP Real – % GDP - Gap to long-term trend y-o-y change, 2-y change, 3-y change - %	ECB
Debt service [2] to income ratio, households and non financial firms	% income	ECB
Debt service to income ratio, non financial firms	% disposable income	ECB
Debt service to income ratio, households	% disposable income	ECB
Households' debt	% gross disposable income	ECB
GDP	Real, y-o-y change - % 2-y change, 3-y change - %	ECB
Consumer price index	Y-o-y change, 2-y change, 3-y change - %	ECB
Monetary aggregate M3	Real, y-o-y change - % 2-y change, 3-y change - %	ECB
Current account	% GDP	ECB
Public Debt	% GDP	ECB
Unemployment ratio	%	ECB
Long-term government bond yield (*)	in % , nominal and real	Bloomberg
3-month money market interest rate (*)	In %, nominal % and real	ECB
Slope of the yield curve (10 Y – 3 M) (*)	b.p.	ECB
Real effective exchange rate	Index, y-o-y change - % 2-y change, 3-y change - %	ECB
Residential property index	Real, y-o-y change - % 2-y change, 3-y change - % Gap to long-term trend	OECD
Ratio of real estate price to disposable income per head	Index based 100 in 2010 Index based 100 at the mean of each country Y-o-y change	OECD
Ratio of house price to rents	Y-o-y change	OECD
Rent index	Real, y-o-y change, - %	OECD
Stock price index	real, y-o-y change, 2-y, 3-y change - %	OECD
Golden rule (gap of real long term interest rate to real GDP)	b.p. over 1 year, 2 y, 3 y	ECB

Total credit to non financial sector includes all debts of the private non-financial sectors (households and firms) whatever (i) the instrument, loan, bond, securitization. (ii) the type of lender : banks, households, firms (iii) the geographical area : external and domestic debt. The gap of this series in % of GDP to its long-term trend is the "Basel ratio»

(*) indicates series with a left-hand side risk.

Table A 2 : Univariate indicators selected, ranked by usefulness

Indicator	Treshold	Auroc	T1	T2	RUS
Monetary aggregate M3 - Real, 3-y change - %	13,37	0,66	0,13	0,35	0,53
Total credit to the private non financial sector - Real – % GDP - Gap to long-term trend	6,03	0,62	0,25	0,22	0,53
Total credit to households - Real – % GDP	40,64	0,7	0,5	0	0,5
Slope of yield curve (10Y-3M) b.p. (*)	1,28	0,58	0,06	0,45	0,49
Total credit to the private non-financial sector - Real – % GDP	126,9	0,73	0,5	0,02	0,48
Debt service, non-financial firms %	28,31	0,63	0	0,52	0,48
Residential property price index - Gap to long-term trend	17,21	0,6	0,5	0,02	0,48
Debt service, households and non-financial firms %	16,12	0,65	0,5	0,03	0,47
Loans to for house purchase - Real, 1-y change - %	9,84	0,56	0,5	0,05	0,45
Long-term government bond yield - Nominal % (*)	4,07	0,82	0,5	0,05	0,45
Loans to for house purchase - Real, 2-y change - %	21,53	0,56	0,56	0	0,44
Total credit to households - Real, 3-y change - %	16,89	0,6	0,19	0,4	0,41
Total credit to non-financial firms - Real – % GDP - Gap to long-term trend	1,47	0,51	0,06	0,53	0,4
Ratio of house prices to rent prices – nominal- 1-y change - %	8,34	0,59	0,5	0,12	0,38
Loans to for house purchase - Real, 3-y change - %	31,68	0,54	0,63	0	0,38
Residential property price index - 2-y change - %	15,61	0,58	0,5	0,12	0,38
Golden rule – 1-y	-0,25	0,68	0,5	0,12	0,38
Total credit to non financial firms - Real – % GDP	87,67	0,65	0,56	0,07	0,37
Ratio of house prices to disposable income per head – nominal- 1-y change	6,5	0,6	0,5	0,13	0,37
Residential property price index - 3-y change - %	21,85	0,56	0,5	0,13	0,37
Residential property price index - 1-y change - %	8,39	0,57	0,5	0,13	0,37
Long-term government bond yield - Real - % (*)	2,5	0,75	0,5	0,15	0,35
Total credit to non -financial firms - Real, variation 1 an - %	2,37	0,53	0	0,67	0,33
Monetary aggregate M3 - Real, 2-y change - %	11,35	0,67	0,44	0,23	0,33
Total credit to households - Real, 2-y change - %	11,84	0,63	0,31	0,37	0,32
Monetary aggregate M3 - Real, 1-y change- %	7,42	0,66	0,69	0	0,31
Total credit to households - Real, 1-y change - %	6,98	0,62	0,5	0,25	0,25
Banking credit to the private non financial sector - Real – % GDP - Gap to long-term trend	5,02	0,62	0,69	0,07	0,25
3-month money market interest rate - Nominal - % (*)	3,23	0,68	0,56	0,2	0,24
3-month money market interest rate - Real - % (*)	1,06	0,66	0,63	0,15	0,22
Golden rule - 3-y	2,85	0,53	0,38	0,47	0,16
Golden rule -2-y	1,15	0,6	0,5	0,38	0,12

Note. AUROC is calculated over the panel. RUS is the relative usefulness = Usefulness/ Min (mu,(1-mu)), here mu = 0.5.

Table A4. Out-of-sample results for the aggregated models in the set Ω_2, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme and the μ =parameter, real-time simulations

Weightings schemes : mod usefulness calculated at	*μ= 0.5*			μ= 0.6			*μ= 0.7*		
	T1	T2	*L*	T1	T2	*L*	T1	T2	*L*
Panel -level	0.44	0.39	0,42	0.20	0.42	0.290	0.06	0.46	0,186
Country-level	0.64	0.28	0,46	0.35	0.33	0,34	0.03	0.46	0,162

Table A5. Out-of-sample results for the aggregated models, percentage of missed crises (T1), false alarms (T2) and loss function (L) depending on the weighting scheme with the set of models Ω_2 , for alternative aggregated thresholds (1), real-time simulations

Options for the weight scheme: models' usefu calculated at	*μ= 0.5*			μ= 0.6			*μ= 0.7*		
	T1	T2	*L*	T1	T2	*L*	T1	T2	L
Panel -level	0.47	0.49	*0,48*	0.37	0.60	*0,46*	0.33	0.66	*0,43*
Country-level	0.57	0.51	*0,54*	0.40	0.55	0.46	0,35	0.63	*0,43*

Notes. See Table 10.

Table A6. In-sample results, percentage of missed crises (T1), false alarms (T2) and loss function (L), depending on preference parameter μ, in sample with an alternative dummy crisis

Weightings schemes : m usefulness calculate	*μ= 0.5*			μ= 0.6			*μ= 0.7*		
	T	T	*L*	T			T	T	*L*
Aggregation over a small set of models Ω1 (*)									
Panel -level	0.47	0.06	0,272	0.17	0.43	0.275	0.00	0.73	0,220
Country-level	0.03	0.46	0,247	0.01	0.57	0,236	0.00	0.60	0,181
Aggregation over a large set of models Ω2 (**)									
Panel -level	0.14	0.36	0,249	0.14	0.35	0.225	0.00	0.58	0,176
Country-level	0.11	0.29	0,202	0.04	0.38	0,178	0.00	0.49	0,148

(*) selected through stringent criteria (25 models); (**) selected through relaxed selection criteria (524 models).

Table A7. Out-of sample results, percentage of missed crises (T1), false alarms (T2) and loss function (L), depending on preference parameter μ, with an alternative dummy crisis

Weightings schemes : m usefulness calculate	*μ= 0.5*			μ= 0.6			*μ= 0.7*		
	T1	T2	*L*	T1	T2	*L*	T1	T2	*L*
Aggregation over a small set of models Ω1 (*)									
Panel -level	0.05	0.58	0.318	0	0.68	0.270	0	0.77	0.232
Country-level	0.00	0.44	0.220	0	0.62	0.247	0	0.69	0.209
Aggregation over a large set of models Ω2 (**)									
Panel -level	0	0.60	0.303	0	0.64	0.254	0	0.67	0.202
Country-level	0.10	0.47	0.285	0	0.57	0.227	0	0.64	0.194

Notes. (*) selected through stringent criteria; (**) selected through relaxed selection criteria.