Working Paper Series | 40 | 2019

Learning from trees: A mixed approach to building early warning systems for systemic banking crises

This paper develops an early warning model of systemic banking crises that combines regression tree technology with a statistical algorithm to improve its accuracy and overcome some drawbacks of more standard models.

Carmine Gabriele
European Stability Mechanism

European Stability Mechanism

Disclaimer
This Working Paper should not be reported as representing the views of the ESM. The views expressed in this Working Paper are those of the author(s) and do not necessarily represent those of the ESM or ESM policy.

Learning from trees: A mixed approach to building early warning systems for systemic banking crises measures

Carmine Gabriele[1] European Stability Mechanism

Abstract

Banking crises can be extremely costly. The early detection of vulnerabilities can help prevent or mitigate those costs. We develop an early warning model of systemic banking crises that combines regression tree technology with a statistical algorithm (CRAGGING) to improve its accuracy and overcome the drawbacks of previously used models. Our model has a large set of desirable features. It provides endogenously-determined critical thresholds for a set of useful indicators, presented in the intuitive form of a decision tree structure. Our framework takes into account the conditional relations between various indicators when setting early warning thresholds. This facilitates the production of accurate early warning signals as compared to the signals from a logit model and from a standard regression tree. Our model also suggests that high credit aggregates, both in terms of volume and as compared to a long-term trend, as well as low market risk perception, are amongst the most important indicators for predicting the build-up of vulnerabilities in the banking sector.

Keywords: Early warning system, banking crises, regression tree, ensemble methods.

JEL codes: C40, G01, G21, E44, F37

[1] Email: c.gabriele@esm.europa.eu

Disclaimer

This Working Paper should not be reported as representing the views of the ESM. The views expressed in this Working Paper are those of the author(s) and do not necessarily represent those of the ESM or ESM policy. No responsibility or liability is accepted by the ESM in relation to the accuracy or completeness of the information, including any data sets, presented in this Working Paper.

ISSN 2443-5503
ISBN 978-92-95085-74-9

doi:10.2852/673365
EU catalog number DW-AB-19-006-EN-N

Learning from trees: A mixed approach to building early warning systems for systemic banking crises [♣]

Carmine Gabriele[*,♦]

Banking crises can be extremely costly. The early detection of vulnerabilities can help prevent or mitigate those costs. We develop an early warning model of systemic banking crises that combines regression tree technology with a statistical algorithm (CRAGGING) to improve its accuracy and overcome the drawbacks of previously used models. Our model has a large set of desirable features. It provides endogenously-determined critical thresholds for a set of useful indicators, presented in the intuitive form of a decision tree structure. Our framework takes into account the conditional relations between various indicators when setting early warning thresholds. This facilitates the production of accurate early warning signals as compared to the signals from a logit model and from a standard regression tree. Our model also suggests that high credit aggregates, both in terms of volume and as compared to a long-term trend, as well as low market risk perception, are amongst the most important indicators for predicting the build-up of vulnerabilities in the banking sector.

Keywords: Early warning system, banking crises, regression tree, ensemble methods

JEL Codes: C40, G01, G21, E44, F37,

[♣] I thank Daragh Clancy, Antonello D’Agostino, Aitor Erce, Ilaria Mameli, Lorenzo Ricci, Rolf Strauch, and seminar participants at the European Stability Mechanism. I also thank for the useful comments the participants of the seminar and poster session where this work was presented. The views herein are the author’s and not those of the European Stability Mechanism.

[*] European Stability Mechanism, Circuit de la Foire Internationale 6a, L-1347, Luxembourg, and University of Luxembourg.

[♦] Corresponding author: C.Gabriele@esm.europa.eu.

1. Introduction

The global financial crisis has led researchers and policymakers around the world to put considerable effort into understanding and preventing systemic banking crises. In doing so, the empirical literature concerned has been focusing on developing early warning systems (EWS) which seek to predict the build-up of dangerous vulnerabilities within banking systems.

The aim of early warning models is not to forecast crisis events. Economic and financial crises are usually triggered by unpredictable shocks. The goal of an EWS is to predict whether a system is vulnerable, to the extent that a sudden adverse shock may lead to a crisis. With regard to banking crises, EWSs turn out to be very useful in giving early signals to the authorities, allowing them to activate pre-emptive actions early enough (e.g., macroprudential measures). The role of policy institutions (national or supranational) is not to predict shocks, but to make sure that when they hit, the economic and financial systems are resilient enough to withstand them.

Typically, early warning models rely on the univariate signalling approach, which allows identifying critical thresholds in a set of indicators. However, this methodology is too simplistic and all the indicators, with their critical thresholds, are unrelated and can deliver counterintuitive signals. Early warning models with a higher degree of complexity rely on the logistic regression technique. However, regression-based models are unable to capture important nonlinearities and complex interactions between macroeconomic and financial variables that may exist in the run-up to crises.[1]

Binary regression trees overcome some of the listed issues as they enable to identify rules-of-thumb with endogenously-determined thresholds in a (ex-post) multivariate framework for a set of interrelated indicators.[2] On the one hand, this means that the ordering of the indicators to look at, and their critical thresholds, are determined by the algorithm and not arbitrarily. On the other hand, every *rule* subsequent to the first one comes from a sub-sample, and this causes the non-global optimality of the final prediction.

The easy interpretability of the results of a binary regression tree, coming from the decision tree structure, is one of the advantages of this methodology, together with other desirable features which we will describe in the next sections. However, decision trees also come with some drawbacks, amongst which the most important is the low out-of-sample accuracy. This is due to the tendency of decision trees to

[1] Potentially, regression models can capture nonlinearities by using interaction terms. However, with a large number of indicators, the number of potential interaction terms becomes very high, unless one does not choose arbitrarily what interaction term to use. In this latter case, there would be a difference between arbitrarily chosen interaction terms for a regression-based model, and data-driven interactions for tree-based models.

[2] By ex-post multivariate framework, we mean that every split of the sample that attempts to separate the events from the non-events, is performed by looking at all the indicators, but eventually using just one of them. We call it ex-post multivariate, because the splitting algorithm is recursive and repeats the same procedure on the sub-sample generated by the previous spits, ending up with a prediction that is based on a series of rules-of-thumb computed in a univariate way, but with each of them which is true conditional to the previous ones.

overfit the data and produce unstable results, as they change when we add new variables or new data points.

An early warning model usually includes a dependent variable, listing the series of events that we are aiming to predict, a set of early warning indicators, chosen by the researcher/analyst according to the literature and expert judgement, and a methodology that uses the indicators to predict the events being studied. Early warning models are not models that predict crises. We use them in order to understand whether imbalances are building up in the economy in such a way that the system becomes more vulnerable and therefore more prone to a crisis. A good early warning model is able to issue accurate warning signals, and, at the same time, it is able to show us where the vulnerabilities are likely to come from in an intuitive way.

Our objective in this paper is to build an early warning model that keeps all these features, so that it achieves the right balance between accuracy and interpretability.

In this paper, we combine some alternative methodologies with the classic techniques available in the literature to build an early warning system for systemic banking crises. We use an ensemble method developed by Savona and Vezzoli (2012) called CRAGGING, which aggregates the results of many "learners" (Regression Trees in this case), and we adapt it to issue early warning signals of systemic banking crises. The paper has the following outline. In section 2, we give an overview of the literature where this paper is placed. In section 3, we describe the data we use, while in section 4, we describe the methodology with a sub-section dedicated to the decision trees and another sub-section dedicated to a methodology that helps to overcome the regression trees drawback, i.e. the ensemble methods. Section 5, outlines the results and some implications, while in section 6, we will describe some robustness exercises we have carried out, before closing the paper in section 7.

2. **Related literature**

Most of the literature on early warning models focuses on two main approaches. The first is the univariate signalling approach (Kaminsky and Reinhart, 1999; Borio and Lowe, 2002; Borio and Drehmann, 2009), which essentially maps the historical time series of a single indicator into past crises and extracts a threshold value for each indicator (independently), above which crises are more likely to happen.

A "second generation" of early warning models estimates the probability of being in a pre-crisis period using a set of several potential early warning indicators jointly. This approach is multivariate and parametric (whereas the signalling approach is non-parametric). A leading example in the literature is given by the logit model (Demirgüç-Kunt and Detragiache, 1998; 2005, Bussiere and

Fratzscher, 2006), but recently more formal procedures such as Bayesian model averaging have also been implemented (Babecký et al., 2012).

In order to overcome some of the drawbacks of the two main techniques used in the literature, some have started using Classification and Regression Tree (CaRT) technology to build early warning models. It is a methodology borrowed from computer sciences and only during the last decade, it started being applied to economic and financial studies. A number of papers use CaRT to explore the triggers of sovereign debt crises (Manasse et al., 2003; Manasse and Roubini, 2009; Savona and Vezzoli, 2012 and 2015), currency or balance of payment crises (Ghosh and Ghosh, 2002; Frankel and Wei, 2004) and banking crises (Dattagupta and Cashin, 2011; Davis and Karim (2008), Davis et al., 2011). Alessi and Detken (2014) apply the random forest (RF), a popular statistical method based on the aggregation of the results of a large amount of single decision trees, to the issue of identifying excessive credit and the associated build-up of systemic risk in the banking system. The RF is an extremely accurate predictor and a solid basis for the selection of the relevant indicators. One problem stemming from the exclusive use of the RF is that it does not provide a tree structure like the simple CaRT does. This means that the results come out of a complicated black box and lack of interpretability, especially for monitoring and policy purposes.

We develop an early warning system for systemic banking crises by combining CaRT technology with the CRAGGING algorithm. This ensemble method helps to overcome the drawbacks of the CaRT method (i.e., lack of robustness and poor out-of-sample performances).

Ensemble methods are statistical tools that allow the combination and the aggregation of results coming from large numbers of "single learners" (the generic name for the basic models aggregated using ensemble methods, single decision trees in our case) and that, therefore, help to overcome the weaknesses of single models by aggregating them. Some examples of ensemble methods are bootstrapping and aggregating (BAGGING), Boosting, RF, and the CRAGGING. Alessi and Detken (2014) apply the RF to select the most important variables in identifying systemic banking crises events and then use only those variables to run a single classification tree. An issue with this methodology is that the ordering coming from the RF does not always match the ordering coming from a single decision tree. For instance, the most important variable is not necessarily going to be at the first node of a decision tree run using the full sample.

Savona and Vezzoli (2012) introduced a methodology that makes the BRT results more robust. They use the V-fold cross validation to build a new dependant variable by averaging the predictions of large a number of BRTs estimated by rotating the sub-samples and all their possible permutations of these latter. With the CRAGGING approach, we employ an ensemble method to build a new dependent variable, which carries information from a large number of trees and which is "trained" to do out-of-sample prediction. Therefore, after building this new dependent variable, using it to run a regression tree should help us improve upon the out-of-sample performance of a standard regression tree.

3. Data

We use a quarterly (unbalanced) panel over the period that goes from 1980 Q1 to 2015 Q4 for a sample of 15 European Union countries.[3] This panel has the positive feature of including a homogeneous set of economies and banking systems.

The dependent variable

We study a binary dependent variable that captures systemic banking crises events. In the early warning models literature, there are mainly two ways of defining the dependant variable. Indeed, it is possible either to use a continuous stress indicator, like the ECB's Composite Indicator of Systemic Stress, or to use a discrete crises database. We use a binary dependent variable that ranges from 1970 Q1 to 2012 Q4[4] that defines two different states: crisis periods and tranquil periods. This dataset was developed by the European Systemic Risk Board in order to better understand how to implement the countercyclical capital buffer. This dataset is an updated and amended version of the banking crises dataset built by Babecky et al. (2012), where the authors identify the quarters in which European Union countries' banking sectors are in a crisis between 1970 Q1 and 2010 Q4. They do it by combining already existing databases with academic studies and in some cases by supplementing these data with the data coming from a comprehensive survey among country experts. According to the ESRB database, systemic banking crises are periods in which:

1) The banking system shows signs of financial distress (non-performing loans above 20% of GDP or bank closures amounting to at least 20% of banking system assets);

2) Public intervention takes place, in response to (or to prevent) losses in the banking system.

The resulting dataset has been amended as follows:

1) Non-systemic crises were excluded;

2) Systemic banking crises that were not associated with a domestic credit/financial cycle were excluded;

3) Periods where domestic developments related to the credit/financial cycle could have caused a systemic banking crisis had it not been for policy action or an external event that dampened the financial cycle – henceforth "near misses" – were added.

Tables 1 and 2 in the annex show the dates and length of the crisis events we use in the analysis and provide some descriptive statistics of these events.

[3] There are crisis episodes earlier than 1980, but many indicators we use do not have such long time series. The countries we include in the sample are Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden and the United Kingdom. Our sample choice was dictated by data availability.

[4] We only start using it after 1980 because of the availability of most of the early warning indicators.

[Tables 1 and 2]

As shown in the tables 1 and 2, more than half of the crisis periods are concentrated in the years from 2008 onwards (i.e. the global financial crisis), while the remaining 46% happen between 1980 and 2007. All countries, except Luxembourg, have at least one banking crisis event. The frequency of the crisis events in the sample is of 15.4%, meaning that the distribution of the events is skewed towards tranquil periods (one quarter of crisis every 6.5 quarters). Since in building the EWSs we use the pre-crisis periods as events, and delete the crisis periods from the sample (together with three quarters before the crisis and 4 the four quarters after the crisis), we also compute the frequency of the pre-crisis periods, in order to have the exact distribution of zeros and ones in our dependent variable. Although it is slightly higher, the message remains the same, the frequency of the ones in the dependent variable is about 20% (one pre-crisis quarter every five quarters). This should already hint to the fact that, once we estimate a critical threshold in the probability of being in a pre-crisis period, this probability should be relatively lower than the *naïve* 0.5.

Using such a qualitative crisis events variable has a drawback in the arbitrariness of the choice of the events, but, on the other hand, it gives more clarity in the type of events under study, as opposed to continuous financial distress variables.

As we are building an early warning model, we are not interested in defining whether we are in a crisis, or to forecast when a crisis is coming. An early warning model is rather meant to issue a warning signal (or a series of signals) when, according to the past behaviour of economic, financial and other variables, imbalances are building up in the economy in such a way that a banking system becomes more likely to experience a stress/crisis event (i.e., it is more vulnerable). This means that a good early warning model has to be able to deliver both a relative and absolute measure of vulnerability of banking systems in different countries. In order to do so, we rearrange the binary dependent variable as follows:

- We assign a one to the periods that go from 20 quarters to 4 quarters before the beginning of the crisis;

- We delete the three quarters before a crisis, the crisis periods themselves and 4 quarters after a crisis, to avoid biased results due to the deteriorating (right ahead of a crisis) or deteriorated (crisis and post-crisis) dynamics of the relevant variables;[5]

- Finally, we assign a zero to every other non-crisis period.

Early warning indicators

In order to understand what happens in advance of a banking crisis, and in order to extract an accurate warning signal, consistent with past events, we select as explanatory variables a set of economic and financial variables. For each variable

[5] Removing the post-crisis periods is suggested also in Bussiere and Fratzscher (2006).

we have downloaded the longest possible series from different sources (Eurostat OECD, BIS, ECB and IMF) starting in 1980 Q1 to 2016 Q4.

We downloaded the series by giving priority to their length (getting the longest possible series) and to their comparability amongst different countries (for the same variable we downloaded the data if that variable was available from the same source for all the countries). We have in total 30 explanatory variables, which we use as early warning indicators and that we group in six different areas for clarity. The areas are Credit, Real estate, Macro, Global, Financial and Contagion.

BIS data on credit to the private non-financial sector are employed to build nine credit variables. We include in the model credit from all sectors of the economy to the private non-financial sector (broad credit henceforth).[6] Moreover, we also include its breakdown in the form of broad credit to non-financial corporations and broad credit to households and non-profit institutions serving households (NPISH). We take these three measures both as a percentage of GDP and in growth rates year-over-year. Another two variables are built by using bank credit to the private non-financial sector, as a percentage of GDP and year-over-year growth rate. Finally, we also employ as an indicator also the so-called Basel gap, which is the distance of the broad measure of credit to the private non-financial economy from its long-term trend, estimated using a one-sided recursive Hodrick-Prescott filter.[7]

In general, we expect that all these credit variables become "higher than normal" in the pre-crisis periods, signalling that the exposures of the financial sector, as well as of other sectors, are becoming dangerous.

Real estate sector dynamics are very important in explaining banking crises (Aldasoro et al., 2018) and, for this purpose, we use four variables that should help us capture some of these dynamics. Using OECD data, we build the real house price growth rate (deflated by HICP inflation), the house price to income ratio and house price to rent ratio.

Excessive house price inflation creates distortions in the balance-sheet and in the behaviour of both banks and private sector agents, especially because of the collateral value effect. A sudden shock may reverse the house prices dynamics, creating the conditions for a banking crisis.

The macroeconomic variables we included in the model are the real GDP growth rate, the inflation rate, the unemployment rate, the current account as a percentage of GDP, the real effective exchange rate deflated by CPI, the growth

[6] This data includes the credit to the private non-financial sector (essentially firms and households) from all sectors of the economy (i.e., banks, but also non-bank financial institutions, general government, trade credit, and so on).

[7] We acknowledge that gap variables estimated using the HP filter suffer from a lack of reliability of end-of-sample estimates of the series' trend due to a methodology drawback. Edge and Meisenzahl (2011), using U.S. data, find that ex post revisions to the credit-to-GDP ratio gap are sizable and as large as the gap itself. We think that these issues are important, but given that, once we have estimated the model, we use the last 16 quarters of data to issue signals, only the last ones could be affected. On the other hand, this variable turns out to be very important in splitting pre-crisis from normal times, as it is also used by the BIS, the ESRB, the ECB and the national authorities. For these reasons we prefer to keep it.

in gross fixed capital formation. We also include here the financial sector employment growth, the debt to GDP ratio (in level and the variation) and the overall population.

The performance of the economy where banking systems reside is important, both because they determine the profitability of banks and the creditworthiness of their customers, and because they contribute to determine the level of risk-taking by banks. We expect that, when an economy is performing well and moves towards overheating, there is a higher likelihood that the conditions for a future banking crisis could appear.

In order to track financial markets dynamics, we use OECD data for the 3-month interbank offer rate as the short-term interest rate, the 10-year government bond yield[8], and the difference between these two as an approximation of the slope of the yield curve[9]. Furthermore, we include the annual growth rate of equity prices (OECD) and the growth rate in the money aggregate M3 (national sources).

These variables, in principle, are important given that they can proxy for the banks' cost of funding, for the sovereign risk (which is always important in European banks given the home-bias in purchasing sovereign bonds), and for the health of the national corporates (the equity price growth).

We also consider the spillover effect coming from other countries' banking systems. Using BIS data on cross-boarder exposures, more specifically the foreign claims of national banks and foreign banks' claims on national banks, we build four variables to proxy for spillover. They are foreign exposures of national banks in percent of total assets of the banking system of the country as well as its annual growth rate, and the exposure of foreign banks to national banks as a percentage of total assets of the banking system as well as its annual growth rate.[10] Another spillover variable that we use is the trade openness, the sum of import and export in percent of GDP.

We expect that these variables signal the build-up of vulnerabilities involving excessive cross-border exposures, especially ahead of widespread banking crises, rather than around isolated episodes.

Finally, we also account for some global determinants by including in the set of early warning indicators the growth rate of global broad credit, the Baa-Aaa spread and the VIX.[11] In particular, we expect that the VIX indicator signals the presence of vulnerabilities when it is too low, given that, historically, a persistently low VIX has corresponded to excessive risk taking by financial institutions (and by

[8] We source these data from the ECB, Eurostat and IMF depending on which is the longest series, provided that they have the same definition.

[9] Long series of short-term interest rates on government bonds were not available for all the countries in the sample and we used the interbank rate, as we preferred to be consistent across the sample instead of using different measures to compute it.

[10] We use this measure on immediate counterpart basis instead of ultimate risk basis in order to have longer series.

[11] We also had an indicator of the global economy, i.e. the global GDP growth, but as it was never relevant we dropped it out.

economic agents in general). Table 3 in the annex shows the list of indicators used in the analysis.

[Table 3]

4. Methodology

Early warning models are typically based on the analysis of signals from single variables, or on empirical binary dependent variable regression. Both of these methodologies have some desirable features which make them the most utilised in this field. The signalling approach is very simple to apply and to explain to a policy audience. It delivers a number of critical thresholds for a set of indicators, but in this case, all the thresholds are unrelated to each other and it is not possible to effectively judge whether (in our case) a banking system is becoming more vulnerable or not. In fact, it is possible that, using the signalling approach, all the indicators are below their critical thresholds, signalling that there is no concern, but it could be that jointly, these indicators would be able to give a different signal if they could interact.

The binary regression tree helps to overcome this issue, as every prediction is conditional on the interaction of a series of variables and their thresholds. Moreover, when applying the signalling approach to a set of indicator, it does not provide any ordering of importance of the various indicators, whereas the binary regression trees endogenously decides which is the first indicator to look at, then the second, and so on.[12]

Unlike the signalling approach, logit/probit regression allows a multivariate framework. Using this model it is possible to estimate the contribution of each indicator to the increase/decrease in the probability of being in a pre-crisis period. However, it does not (easily) allow the estimation of critical thresholds for the explanatory variables (which is a desirable feature for an early warning system) and their ordering. Furthermore, this framework potentially allows for interactions amongst explanatory variables, albeit there is a limit to the amount of interactions that can be introduced in the equation. However, the researcher would arbitrarily decide these interactions.

We apply the Binary Regression Tree (BRT) methodology to build an early warning system for systemic banking crises. It is a technique developed by Breiman (1984) and widely used in genetics, engineering, marketing, biology, chemometrics and many other scientific fields. We improve upon this methodology by integrating the binary regression tree within the CRAGGING algorithm, developed by Vezzoli and Stone (2007). Given that this methodology is not widely used in economics, we

[12] However, with the decision tree, every time that the algorithm does a split and grows the tree further down, the new variable and relevant threshold are found over a sub-sample compared to the previous one.

use a more standard logistic regression as a benchmark in order to compare its prediction performances, as it is easier to grasp given its wide use in this field.

Classification and regression Trees (CaRT)

Using Binary Regression Trees, we construct a prediction model that offers a non-parametric framework for uncovering non-linear and interactive structures in the data. It is a partitioning algorithm which recursively and endogenously identifies the variables and the respective thresholds, which are able to split the sample into homogeneous subsamples from the perspective of the dependent variable. Within each partition, a simple prediction model is fitted. As a result, the partitioning can be represented graphically as a decision tree. This boosts interpretation and provides greater insights to policymakers.

Regression trees allow for the possibility that different relationships may hold between predictors at different times and under different cross-sectional conditions without having to inflate the set of indicators with hundreds of interaction terms. The algorithm starts by finding the one binary split that delivers the best homogeneity of the dependent variable in the resulting two sub-samples. This split delivers our root node and two child nodes, one where the probability of a crisis increases and another where it decreases as compared to the parent node. The algorithm goes through every possible explanatory variable available in the dataset, and, for each of them, it assesses every value as a possible threshold value to split the dependent variable. The algorithm associates a value to each possible threshold. This value is the weighted average of the mean square errors of the dependent variable in the two subsamples stemming from the use of the assessed threshold. The indicator/threshold pair that minimises the mean square error will characterise the first split (root node), which separates all the observations in the sample in two child nodes. Once the first split is done, the algorithm proceeds recursively to further split the resulting subsamples using the same method, creating more child nodes, and it will continue until a stopping rule becomes binding or a full tree is grown (i.e., until each final leaf contains one observation).

Figure 2 in the annex shows a sample of regression tree, which should help with the terminology and with understanding it better. The tree starts with a root node, which is given by the first split done by the algorithm. The root note splits the sample in two according to the estimated critical threshold of one of the early warning indicators. After the first split, we are working on two separate sub-samples on which the algorithm repeats the same procedure, creating two child nodes. At this point, separate parts of the tree (for instance the root node and the right child note with its predictions are called branches). The final prediction is also called a leaf. If we keep splitting, the last child node at the end of each branch is called a terminal node. Once we have a full tree (it is not possible to proceed to further splits, as there is only one observation in the terminal nodes), we can prune (cut) nodes or entire branches according to an "optimal pruning" rule. Alternatively, we can stop growing the tree before it reaches its maximum degree of deepness (i.e., when all the terminal nodes include one observation).

[Figure 2]

Starting from the root node, each node of the tree has a question and a rule. The early warning indicator and its threshold represent the question. According to our answer to the question, we will have a rule. If we answer 'yes' to the question, we go right, otherwise we go left, to the next nodes (or to the predictions, if we are in the terminal node). For instance, in the sample tree, if we are in the first child node on the left, we are asked the question whether the credit gap is larger or smaller than 3.2. If it is larger, we go to the right, ending up in another node where we are asked about the VIX. Otherwise, if the credit gap is smaller than 3.2, we end up in the child node on the left, where another question will be asked. In this case, starting from the credit gap, both the child nodes are also terminal nodes and the split inside them will end up with a prediction.

A tree should never reach its maximum deepness, and for this we have already mentioned above possible stopping rules in growing the trees, or an alternative methodology to decrease its size. Possible stopping rules could be a minimum number of observations in the final leaves, a minimum number of branch node observations, a maximum number of levels of splits. When growing a full tree, we can reduce its dimension by merging the final leaves when the sum of the errors (weighted by the relative frequency of 'zeroes' and 'ones' in the leaves) is larger or equal to the error in the parent node or by pruning the tree by following an optimal pruning sequence.

In this paper, we grow a full tree and then we prune it back by using an optimal pruning rule. The optimal pruning rule helps to balance the trade-off between a good in-sample fit of the model and avoiding the over-fitting of the data. It consists of minimising the following error-complexity measure:

$$EC(\alpha) = Err(T) + \alpha \times \#T \quad \text{(1), where}$$

the Err(T) is the re-substitution error estimate of the tree (the larger the tree the smaller the error), #T is the number of leaves of the tree and α is the complexity parameter and defines the cost of having a larger or a smaller tree. If we would only consider minimising the in-sample error of the model, we would build a very large and complex tree that would not result credible when applied to new data.

The output is a decision tree structure displaying various sequences of conditions to hold in order to obtain different predictions. In our case, the conditions are sequences of thresholds in the relevant explanatory variables, which we use to understand what happens ahead of a banking crisis, whereas the predictions are given by averaging the dependent variable values within each final leaf and represent the probability of being in a pre-crisis period (or the degree of vulnerability of the banking system).

This methodology is particularly useful for building early warning systems for banking crises, as it recognises combinations of vulnerabilities that can trigger crises rather than identifying them in the deterioration of a single indicator. Moreover, the methodology allows us to recognise that economic indicators may have a nonlinear impact on the vulnerability of a country's banking system. Unlike other statistical methods, the binary regression trees method does not need any distributional assumptions on data and it does not assume any underlying functional form. This means that we are not assuming, for example, normality of errors, and we do not assume linearity, additivity or other functional forms. Furthermore, this methodology allows the use of a high number of explanatory variables (even if collinear), can deal with missing values and it is not affected by monotonic transformations of the variables. Last, but not least, the advantage that stems from using this method is the interpretability of the final output. In fact, the tree structure (with its simple rules of thumb) offers a visual outline that significantly simplifies the interpretation of results for policymakers and non-technical audience.

As previously shown, the tree structure is very easy to read. In fact, starting from the root node, what we have is a set of simple rules (i.e., if the indicator is above the threshold value go right, otherwise go left) that end up in a prediction. Hence, conditional to the path, the prediction can be different, and this allows for the fact that not all crises are alike. On the other hand, using this methodology also carries some disadvantages. Indeed, it is not possible to use it to find any causality or to establish any relationships that hold true through the entire dataset. In fact, every time we split the sample (starting from the root node) the relationships become more localised. Another disadvantage is that, being a non-parametric method with no distributional assumptions, there is no possibility to conduct statistical tests on the results or to compute marginal effects. However, it is possible to compute at each node the variation in the probability linked to breaching the threshold.

Another issue is related to the so-called masking problem, where one of two explanatory variables (both very good in explaining the dependent variable) does not show in the final tree because it is very much collinear with the other one. This is comparable to the standard regression analysis, when one of two collinear variables drops out. Regression tree technique tends to be "too" good in-sample as they tend to over-fit the data, at the expense of the goodness of the out-of-sample predictions. As already mentioned, a proper pruning helps to overcome this drawback.

Finally, the main two drawbacks of this methodology are that the model tends to be instable because of the way it partitions the data space and that the variable selected by the algorithm in the first split takes a disproportionate effect on the following choice of predictors and thresholds. In fact, even a relatively small change in the data set (new observations becoming available, or the addition of new indicators to the dataset) can lead to very different trees. As the model is not particularly robust when adding new predictors or observations, other models outperform its out-of-sample prediction ability.

Many contributions in this literature have attempted to overcome these drawbacks in the regression (and classification) tree models by using the "perturbation and combination" approach. The main idea underlying this approach is to create many artificial samples from the original dataset and then perturbing them, in order to estimate multiple models and to generate multiple pseudo out-of-sample predictions, which we then average over time. Breiman (1996) in this direction has introduced the Bagging (Bootstrap and AGGregatING) algorithm, which generates a number of new datasets by bootstrapping the original dataset.

In the second step, we estimate a model for each bootstrapped dataset and aggregate all the predictions by averaging. The aim of this procedure is to reduce potential instability of forecasts and to address the over-fitting problem. Another contribution to this literature comes from Breiman (2001) with the Random Forest algorithm, which is similar to the bagging with the difference that in the random forest algorithm, every bootstrapped dataset has a different dimension and for each of them, only a subset of explanatory variables is selected for prediction. Freund and Schapire (1996), proposed the "Boosting" approach. The idea is to generate multiple simple models for a random portion of the data and then to combine them. As noted, all these statistical approaches generate thousands of different models, and from these models, they generate predictions in the form of a probability. The predictions coming from these models are quite accurate, but they come at the cost of transparency and economic interpretability. Indeed, these approaches are sort of black boxes, which allow for no "economic intuition" and are impossible to interpret.

The CRAGGING algorithm exploits the idea of "perturbation and combination" for achieving predictive accuracy without sacrificing economic intuition. Furthermore, the CRAGGING algorithm can preserve the structure of the data by using in an efficient way the panel data structure of the data. The CRAGGING algorithm improves the stability and the out-of-sample performances of the regression tree model without sacrificing the easy interpretability deriving from the tree structure of the final output.

CRAGGING

In this section we describe the state of the art of the CRAGGING algorithm as in Savona and Vezzoli (2012), and we add the description of an improvement to the empirical strategy.

Let (Y, X) be an unbalanced panel data with N observations. Each unit[13] j of the panel, with $j = 1, \dots, J$, has a number of periods t, with $t = 1, \dots, T_j$ and $J \, x \, T_j = N$. Denote with $L = \{1, 2, \dots, J\}$ the set of units and with $x_{j,t-1} = (x_{1,t-1}, x_{2,t-1}, \dots, x_{r,t-1}, \dots, x_{R,t-1})$ the vector of predictors of unit j observed at time $t - 1$ where $j \in L$ and R the number of predictors for each country. As the name CRAGGING suggests, using the V-fold cross-validation, L is randomly partitioned into V subsets[14] denoted by L_v, with $v = 1, \dots, V$, each containing J_v units and N_v observations[15]. Denote with L_v^c the complementary set of L_v containing J_v^c units and $L_{v/l}^c$ the set where the *l*-th unit is removed by L_v^c ($l \in L_v^c \; and \; L_{v/l}^c \cup l = L_v^c$).

The cost complexity parameter $\alpha \geq 0$, is the tuning parameter of the cross-validation. Hence, for a fixed α, for each L_v and for each $l \in L_v^c$ let

[13] In our case, units are countries.

[14] In the partition, it is necessary that the number of subsets V is smaller than the number of units J.

[15] The dimension of each subset is of as nearly equal size as possible.

$$\hat{f}_{\alpha,L^c_{v/l}}(\cdot) \quad (2)$$

be the prediction function of a single tree trained on data $\{y_{j,t}, x_{j,t-1}\}_{j\in L^c_{v/l}}, t = 1,2,\cdots, T_j$ and pruned with cost-complexity parameter α. The corresponding prediction in the test set is

$$\hat{\mathrm{y}}_{\mathrm{jt},\alpha\mathrm{l}} = \hat{\mathrm{f}}_{\alpha,L^c_{v/l}}(x_{j,t-1}) \text{ with j} \in \mathrm{L_v} \text{, and t} = 1,2,\cdots,\mathrm{T_j}. \quad (3)$$

Therefore, at each step, we exclude one unit (country) from the training set. Then, we use the training set without the excluded country, to grow a tree. If this perturbation causes significant changes in the obtained J^c_v trees, the accuracy of the predictors improves by running the following equation:

$$\hat{\mathrm{y}}_{\mathrm{jt},\alpha} = \frac{1}{J^c_v}\sum\nolimits_{\mathrm{l}\in L^c_v} \hat{\mathrm{f}}_{\alpha,L^c_{v/l}}(x_{j,t-1}) \text{ with j} \in \mathrm{L_v} \text{, and t} = 1,2,\cdots,\mathrm{T_j} \quad (4)$$

which is the average[16] of the functions (3) fitted over the units contained within the test set $\{y_{j,t}, x_{j,t-1}\}_{j\in \mathrm{L_v}}, t = 1,2,\cdots, T_j$. The objective of the CRAGGING is to improve accuracy, reduce the variance of out-of-sample prediction errors, and reduce the variance in the model selection process.

The first step of the CRAGGING algorithm, called leave-one-unit-out cross-validation, is used for perturbing the training set by removing one unit per time. Furthermore, we have to note that such cross-validation does not destroy the structure of the data, unlike the common cross-validation that partitions the observations randomly. Hence, the CRAGGING algorithm tries to solve the sampling of the observations in the case of dataset with time-varying predictors.

The second step of CRAGGING, called v-fold cross-validation, is implemented on the test sets with $v = 1, \ldots, V$, with the purpose to find the optimal tuning parameter, α^*, that minimizes the estimate of the prediction error on all the test sets. Formally,

$$\alpha^* = \mathrm{argmin}_\alpha \, \mathrm{LF(y,\hat{y})} \text{ with j} \in \mathrm{L} \text{, and t} = 1,2,\cdots,\sum\nolimits_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{T_j} \quad (5)$$

where LF(.) is a generic loss function.

The entire procedure described above is run M times to minimize the generalization error, which is the prediction error over an independent test sample, then averaging the results in order to get the CRAGGING predictions to use in the second step. Using the Strong Law of Large Numbers, Breiman (2001a) has indeed shown that, as the number of trees grows larger (M → ∞), the generalization error has a limiting value and the algorithm does not over-fit the data. As a result, the CRAGGING predictions are given by:

$$\tilde{\mathrm{y}}^{\mathrm{crag}}_{\mathrm{jt}} = \mathrm{M}^{-1} \sum\nolimits_{\mathrm{m=1}}^{\mathrm{M}} \hat{\mathrm{y}}_{\mathrm{jt},\alpha^*} \text{ with j} \in \mathrm{L} \text{, and t} = 1,2,\cdots,\sum\nolimits_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{T_j}. \quad (6)$$

[16] The base learners $\hat{\mathrm{f}}_{\alpha,L^c_{v/l}}(.)$ are linearly combined so that the $\hat{\mathrm{y}}_{\mathrm{jt},\alpha}$ will act as a good predictor for future (y|x) in the test set.

In the third step, a single tree, which we name as Final Tree, is fitted on $(\tilde{y}^{\mathrm{crag}}, \boldsymbol{X})$ with cost complexity parameter $\alpha^{**} = \mathrm{M}^{-1} \sum_{\mathrm{m}=1}^{\mathrm{M}} \alpha^{*}$. Here, through the replacement of Y with CRAGGING predictions we do four things:

1) Mitigate the effects of noisy data on the estimation process that affect both the predictors and the dependent variable itself;
2) Give the tool a better grasp of how to predict out of sample;
3) Avoid the cliff effect due to the binary dependent variable Y;
4) Grow a final RT that encompasses the overall forecasting ability arising from multiple trees.

Using this process, we obtain a parsimonious model, with good predictions (accuracy), good interpretability and minimal instability. In other words, the second step of our procedure is conceived to deliver a single tree to better understand the complex CRAGGING predictions. This is in line with the idea of assigning the simplest representations to the most accurate models suggested by Catlett (1991) and others.

In order to be able to issue a warning signal, we need to add a further step in the methodology. Indeed, while classification trees predicts whether a new observation classifies as pre-crisis or normal, regression trees return a probability of being in a pre-crisis period. However, a probability does not tell us enough in absolute terms, as for instance 0.4 could be either small or big. For this reason, we apply the signalling approach to estimate a critical threshold value for the dependent variable coming from the CRAGGING routine.

The signalling approach allows us to identify the threshold for predictions coming from the CRAGGING dependent variable as well as from the final tree. The thresholds that we identify are the ones that, in each of the two cases, best separate normal periods from pre-crisis periods in the initial binary dependent variable. The algorithm that implements this method recursively sets each of the predicted values (for both CRAGGING and final tree prediction) as possible critical thresholds, and then classifies the observations: above the threshold, it signals pre-crisis and below it does not.

For each threshold that the algorithm tries, there are four possible outcomes. The final predictions may issue a warning signal and it is correct (A), fail to issue a warning signal when it should have signalled it (i.e., missed pre-crisis, B), issue a warning signal but it is wrong (i.e., false alarm C), not issue a warning signal and it is right (D). These four outcomes fill the confusion matrix (Figure 1).

[Figure 1]

This means that the algorithm will classify the observations in a number of confusion matrices, which is the same as the number of possible thresholds it attempts to do the splits. Then, for each "filled-in" confusion matrix, it computes the value of a loss function. It will eventually use the threshold associated with the lowest value of the loss function. The loss function we use is called the policymaker loss function (PMLF), which is simply a weighted average of the two types of errors, and it is defined as follows:

$$PMLF = \theta P_1 \left(\frac{B}{A+B}\right) + (1-\theta) P_2 \left(\frac{\mathrm{C}}{\mathrm{C+D}}\right) \qquad (6)$$

where θ is the parameter that defines the policymaker's aversion for the two types of error. If $\theta > 0.5$, we give more weight to the missed alarm rates, with a policymaker that cares more about the potential loss from missing a stress event. If $\theta < 0.5$, we give more weight to the false alarm rate. In this case, the policymaker is more concerned about not issuing too many false alarms, as they are more averse to the potential loss of output given by taking too many pre-emptive measures following the false alarms.

P_1 and P_2represent the relative frequencies of having either of the two final outcomes as a percentage of the total number of observations. This way the weight is also distributed according to which of the two events (pre-crisis and normal times) is more frequent in the sample. This is the most generic version of the PMLF. However, we omit P_1 and P_2 as also in Alessi and Detken (2014). We want to avoid ending up overweighing the false alarms error rate, and consequently come up with thresholds that are too high. In fact, by using these weights in the case of banking crises, we would attribute a larger weight to the false alarm rate because the distribution of the events is skewed towards the tranquil periods. Finally, in the exercises that we carry out in this paper, we set the policymaker's preference $\theta = 0.5$ (the policymaker is equally conservative regarding the two types of errors).

5. Results

The outcome of the described routine is a tree structure which, through a series of rules of thumb, provides a prediction about whether we are in a pre-systemic banking crisis period or not, conditional on some of the mentioned rules. The prediction is a value between zero and one, which is interpretable as the probability of being in a pre-crisis period. It can also be interpreted as a systemic risk indicator. Figure 3 shows the resulting final tree.

[Figure 3]

Using the signal extraction method, we can then identify a threshold in this probability by relating it to the binary crisis variable, such that we will be able to say whether an outcome, in the form of a probability, is high or low, and therefore we would be able to issue a warning signal or not. [17]

The outcome could be used either in absolute terms, or in relative terms. In the first case, once we compute the relevant threshold, as said, we could issue a warning signal in case the model suggests it. In the second case, for the panel of countries that we use in the estimation, we could sort the banking systems from the least to the most risky. Apart from the results, the interesting thing is that the tree structure allows us to understand where in the economy the vulnerabilities are building up.

[17] The relevant threshold computed using the signalling approach is 0.3.

The Final Tree

The estimated final tree shows that, according to the past data, the most relevant variable in splitting normal periods from pre-crisis ones is the bank credit to GDP ratio, with a threshold of about 94%. After this step, we could go either right or left, and find another rule. For instance, if the bank credit to GDP ratio is larger than 94%, we look at the rule on the right, which has as splitting rule the VIX with a critical threshold of 16.7. We need to take the same steps, until we eventually reach a final leaf that shows a prediction in the form of a number between zero and one. We can interpret it in various ways, as, for instance, the probability of being in a pre-crisis period, the degree of vulnerability of a banking system, or as an indicator of systemic risk.

The final tree (Figure 3) shows that, using this methodology, and according to the data we are using, the most important variables to determine whether the banking system is vulnerable are the credit, and the global variables. Regarding the former, the bank credit as a percentage of GDP appears in the root node of the tree, resulting in the most relevant variable when using the entire sample. Year-over-year growth rate of bank credit appears four times, but it is less relevant because for three times it delivers two predictions which are both below the relevant threshold estimated using the signalling approach, as described at the end of the methodology section.

The broad credit measure appears both in distance from its long-term trend (Basel gap) and as a percentage of GDP. The first one, consistently with the early warning literature, is quite relevant as it appears at the first child node, on the left of the root node, which includes a large part of the observations (almost 70%). Finally, amongst the credit measures, also household credit as a percentage of GDP appears once.

The global variables also seem to be important, as they appear three times (the VIX twice and the Baa-Aaa spread once) and, in the nodes where they are the relevant indicator, they include half of the sample.

Other indicators seem to be less relevant, unless they are considered in the broader context of the entire tree.

The final tree mainly shows three different states of the world. One in which the bank credit to the private non-financial sector is high. The other two states of the world have bank credit lower than 94% of GDP, but in one case we are likely to be in a credit boom (Basel gap > 3.2), while in the other case we are not.

In the first case (high level of credit from banks), we end up with four predictions, all of them showing a high probability of being in a pre-crisis period. This signals that banks may be accumulating too much risk in their balance-sheets. In case there is low global market uncertainty, then the model judges the banking system riskier, since the risk perception by economic agents is low. Otherwise, if the VIX is higher than the critical threshold, the model would still issue a warning signal, but there would be a lower probability associated to it.

When bank credit is below its relevant threshold, as mentioned above, we need to look at the Basel gap. If we are in a credit boom, then a low perception of risk (low VIX), and low unemployment (e.g., overheating economy), would trigger a warning signal from our model. However, if the VIX is higher than its critical threshold, the model issues a warning signal only if the year-over-year growth rate of bank credit is above 8.8%.

Finally, when the Basel gap is lower than its critical threshold, our model issues a warning signal only in two cases. One is when broad credit is higher than 153%, signalling that credit to the private non-financial sector is high, despite not coming from banks. In the second case the broad measure of credit is low. However, there is low house affordability (house price to income ratio higher than the critical threshold), combined with high household debt and low perception of risk (low Baa-Aaa spread, which signals that investors are more prone to risks and search for yield by purchasing lower rated corporate bonds).

In order to understand whether the CRAGGING methodology could be of some use, we compare the accuracy of its results with the results coming from other models.[18] Specifically, we compare the in-sample fit of the final tree estimated using the "CRAGGED" dependent variable, where the grouping for the cross-validation was "by country" (Final tree-country henceforth), to the fit of:

- The same model, where the grouping for the cross-validation was "by time" rather than "by country" (Final tree-time henceforth);
- The prediction coming from the CRAGGING (with both types of grouping; CRAGGED-time and CRAGGED-country henceforth);
- A standard regression tree;
- A stepwise Logit;
- A stepwise Logit augmented with interaction terms.[19]

More importantly, we compare the accuracy of the early warning signals, by carrying out an out-of-sample prediction evaluation exercise for the following models:

- The final tree estimated using the "CRAGGED" dependent variable (and where the grouping for the cross-validation where "by country");
- A standard regression tree;
- A stepwise Logit;

In-sample accuracy comparison

In Table 4, we show the results of the in-sample comparison. In the first column, the table shows the value of the PMLF, which is the average of the two error rates shown in the second and third columns (the lower the better). The accuracy rate in the fourth column is the number of times the model issues the right signal, independent of whether it is a warning or not, as a percentage of the total number of observations in the sample. In the fifth column of the Table, we report the Area Under the Receiving Operator Curve (AUROC), which is a performance

[18] The comparison between models will take into account that the tree-based models can deal with missing data, so the datasets used will not be exactly the same than when we use the Logit model.

[19] We also attempted to exploit the partitioning of the data space provided by regression tree methodology in order to choose the interaction terms for the Logit model, however the in-sample improvement was marginal.

measurement for classification problems at various thresholds settings.[20] Finally, in column 6 there is the threshold associated with the model used, which is important in the assessment, as it shows that different models, while using the same dataset, could set different thresholds.

[Table 4]

It is clear that there is a trade-off between the missed crisis and false alarm rate, and that, for instance, in order to reduce the missed crises error rate, we would need to give it more weight, and this would lower the critical threshold. However, by lowering the threshold, we automatically increase the false alarm rate, given that now we issue a warning signal more often.

The results show that the standard regression tree and the CRAGGED-time have the lowest value for the loss function and the highest accuracy rate. However, they both achieve this result particularly for the very low rate of false alarms. If we focus on both types of errors separately, we notice that the only model that has a missed crisis error rate lower than false alarm error rate, is the Final tree-country, which has also the lowest accuracy rate and the second lowest AUROC. The AUROC and the accuracy rate do not distinguish between the two errors, and since the distribution of events is skewed towards the normal periods, the false alarm errors will weigh far more when looking at these two measures. For this reason, we also look at the two error rates separately and at the PMLF in order to compare the performance of different methodologies.

According to the PMLF, the regression tree, the final tree-time and the CRAGGED-time are the models with the best in-sample prediction. The final tree-country is the only model that achieves a missed crises rate lower than the false alarm rate (which, for some monitoring institutions, is desirable). This means that all other models are better (at least in-sample) in predicting when we are not in a pre-crisis period. The Logit has a similar overall result, compared to the final tree-country, but with the figures for the error rates inverted. Apart from the standard regression tree, all the AUROCs are comparable. When we use the tree-based models, we find lower thresholds than when we use the Logit model.

Out-of-sample accuracy comparison: full forecasting horizon

To compare the out-of-sample predictions, we narrow down the set of models for various reasons. We implemented the out-of-sample predictions also for the CRAGGED-country, but the results were similar to the ones from the final tree-country model (slightly worse), and we decided not to report them.[21] Finally, we do not implement the Augmented Logit as the in-sample results were very close to the Logit results without interaction terms. When performing the out-of-sample prediction exercise, having the interaction terms adds complexity to the model, which would have likely deteriorated the forecasting performances.

We implement the exercise by stopping the sample at some point in the past, and predicting whether the upcoming 16 quarters are pre-crisis or not, and repeating

[20] The AUROC has to be larger than 0.5 in order for the classifier to be of some use.

[21] Results available upon request.

it recursively every time a new data point is added. We start the exercise by stopping the dataset in 2000 Q1, and repeating it up to 2008 Q1 each time that we add a new quarter of data. We stop the exercise in 2008 Q1 because with the beginning of the global financial crisis, it becomes more difficult to assess the prediction since most of the "actual" values would be a crisis, but we are predicting-pre-crisis. This means that we would end up with too few observations to compute the relevant statistics and, as we assessed, they become unstable and unreliable. Because of the way the dependent variable is structured, we can only update it up to four years ahead of the current date, because we assign the ones to the pre-crisis periods. This means that, for instance, we may assign a zero to the last three years now, but if a crisis happens within six months from now, then we had assigned zeros erroneously, given that now all the previous 16 quarters ahead of the mentioned crisis will have to be a one. For this reason, we implement the out-of-sample by estimating the models using the data up to when we have the dependent variable available, and use the successive 16 quarters of data to predict out-of-sample. We compute some measures to assess and compare the model for the 16 quarters ahead all together, and the we do the same looking only at the first 8 quarters ahead, and then at the next 8 quarters (from the 9th to the 16th).

Table 5, 6, and 7 show the results. Looking at the 16-quarters-ahead predictions assessment (Table 5), we can say that the final tree-country and the stepwise logit model perform overall similarly, given the similar values of the PMLF. The simple tree out-of-sample prediction performances become (as expected) the poorest of the three models. Looking deeper in the performances of the three models, a few points should be noted. The standard tree is not viable as it makes a missed crisis error rate larger than 0.5. It means that from that perspective, tossing a coin is more reliable.

[Table 5]

Comparing the final tree to the logit, we notice that they achieve a similar PMLF with opposite error rates. As compared to the logit, a gain of 12 p.p. in the missed crises rate (for the final tree) costs 16 p.p. more in the false alarm rate. This means that, overall they perform similarly, according to our PMLF. In absolute terms, if we do not distinguish between the two types of errors, the logit would be more accurate. If, instead, we only look at the missed crisis error (while still getting acceptable false alarm rates, at least below the naïve threshold of 0.5), then the final tree overperforms the logit.

Although the thresholds of the final tree and the logit are not comparable, it is worth mentioning that they have, on average, the same threshold over the various iterations of the out-of-sample predictions.[22] However, this threshold splits the events very differently as it effectively minimises different types of error rates in different cases.

Figure 4 shows the evolution of the missed crisis rate for the three models over the evaluation period. The standard tree is constantly higher than the others, while

[22] As already mentioned, the tree-based models can deal with missing data while the logit ones cannot.

the final tree has the smaller error most of the time. The missed crisis error rate for all the models, starts decreasing when the global financial crisis approaches, as it becomes easier for the models to issue a right warning signal. The reasons are mostly that the models learn every time that a data point is added, and that there are, at each iteration, more pre-crisis events (ones) in the evaluation sample.

[Figure 4]

Figure 5 shows the evolution of the false alarm rate for the three models over the evaluation period. The logit and the standard tree have the best performance from this perspective over time, but this comes at the cost of an excessive missed crisis rate. The spike in error rates between 2003 and 2004 probably come from the fact that most countries were in pre-crisis period and the models, especially because of global and credit variables, started to issue warning signals for every country (also the ones that were not – yet – in a pre-crisis periods), creating a spike in the false alarm rate.

[Figure 5]

Figure 6 shows the evolution of the estimated thresholds over the evaluation period. Apart from the standard tree, which has unstable thresholds, the other two models show quite stable thresholds over time.

[Figure 6]

Out-of-sample accuracy comparison: shorter and longer horizons

Looking only at the first 8 quarters out of sample (Table 6), the results remain unchanged from the perspective of false alarms, while the missed crisis rates become smaller. Qualitatively the results remain similar, with the final tree-country outperforming the other models on the missed crises rates, and the logit outperforming the final tree-country on the false alarm rate. The standard tree would be even better than the logit on the false alarm rates, but its missed crises rate is 0.5, and it is not acceptable as an early warning indicator.

[Table 6]

Looking at the forecasting horizon that goes from 9 quarters to 16 quarters ahead (Table 7), the performances deteriorate on average for all models. The final tree-country is the only model that has both error rates below 0.5, but it has the highest false alarm rate.

[Table 7]

We observe that the false alarm rate is quite stable over the forecasting horizon, and even improves (marginally) for longer horizons. It means that these models are at least as good at not issuing false alarms in the shorter term, than in the long term when they are not needed. On the other hand, all the models, when passing from shorter horizons to longer ones, become less precise in issuing warning signals when they are actually needed.

Figures 7 through 10 show that the evolution of the error rates for all the models over the shorter and the longer term, is similar to the ones computed over the entire forecasting horizon (16 quarters ahead).

[Figures 7-10]

We believe that our method to build an early warning system represents a good complement to the mainstream early warning system tools. As introduced at the beginning of this paper, it has some desirable features while maintaining a good prediction performance, sometimes outperforming the classic models.

However, it is not a perfect methodology. Apart from the already mentioned drawbacks, we have to point out that the final predictions are not a global optimum. This means that, once the decision tree algorithm finds the first variable and threshold for the first split, it keeps going, until a binding rule stops it (or until all the final leaves include only one observation). However, it could be that another variable, or another threshold within the same variable, could be a worse splitter at that point, but could end up in a better final prediction. However, this is something that would be so "expensive" from the intensity of calculation that to the best of our knowledge, nobody has tried to overcome this issue.

6. Further work for implementation and future research

In order to make this methodology implementable within a monitoring institution, there would be additional work to do. One should be able to update independently the banking crises dataset, such that every quarter the model could be run again. The set of early warning indicators needs to be expanded, in order to capture more features (e.g., capital flows, bank specific variables, etc.). At the same time, increasing the countries in the panel, keeping into account possible data constraints, would be of help for practitioners.

We could introduce another layer to the methodology by using (as mentioned earlier in the paper) the random forest technique in order to choose only the most important variables from a larger amount, before going through the CRAGGING procedure.

Finally, this methodology could be transferred also to other types of crises, as it is not bound to work only for banking crises.

Future research would involve a comparison of the results of the CRAGGING where the grouping units are countries with the results of the CRAGGING where the groupings are done according to time and randomly. We believe that, given that we use mainly macro data, with strong interdependence amongst countries and time persistence, when grouping according to time, the results could improve.

7. Conclusions

In this paper, we build a multivariate early warning model for systemic banking crises combining a statistical algorithm (CRAGGING) with the regression tree technology. The combination of the two methodologies helps to improve the accuracy of the standard regression trees. The resulting early warning model has a set of desirable features for this class of models. It provides endogenously-determined critical thresholds for a set of indicators, related amongst each other, presenting them in the form of an intuitive decision tree structure. In fact, it takes into account the conditional relations between various indicators when setting early warning thresholds. In doing so, it produces accurate early warning signals as it results from a comparison with the signals of a logit model and of a standard regression tree.

Early warning models are not models that predict crises. We use them to understand whether imbalances are building up in the economy in such a way that the system becomes more vulnerable and therefore more prone to a crisis. A good early warning model is able to issue accurate warning signals, and, at the same time, it is able to show us where the vulnerability is likely to come from in an intuitive way.

An early warning system represents one of the tools that monitoring institutions should have in their toolkit, as a complement to other tools. It should issue warning signals and provide insights on the roots of the vulnerability, so that the practitioners can start a more judgemental analysis in an informed way.

Different institutions might have a different preference/aversion towards either of the two error types. Unfortunately, in classification problems, we cannot minimise both errors at the same time. Therefore, depending on the type of institution, one should give more weight to either of the two error types. The result is that in some cases, the same model applied to the same data could be preferable for one institution but not for another one.

References

Alessi, L., Detken, C., (2014). "Identifying excessive credit growth and leverage". *Working Paper Series* 1723, European Central Bank.

Aldasoro, I., Borio, C., and M. Drehmann, (2018). "Early warning indicators of banking crises: expanding the family". *BIS Quarterly Review,* 2018.

Babecký, J., Havránek, T., Matějů, J., Rusnák, M., Šmídková, K. and Vašíček, B., (2012). "Banking, Debt and Currency Crises: Early Warning Indicators for Developed Countries". *Working Paper Series 1485*, European Central Bank.

Borio, C. and Drehmann M. (2009): "Towards an Operational Framework for Financial Stability: 'Fuzzy' Measurement and Its Consequences.", BIS Working Paper No. 284.

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

Breiman, L., Friedman, J., Olshen, R. and Stone,C., (1984). "Classification and Regression Trees". *Wadsworth and Brooks, Monterey, CA.*

Breiman, L., (2001). "Random Forests." *Machine Learning* 45 (1): 5–32.

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

J. Catlett, Megainduction, (1991): "Machine learning on very large databases", PhD Thesis, School of Computer Science, University of Technology, Sydney, Australia, 1991.

Davis, E.P., Karim, D., (2008). "Comparing early warning systems for banking crises". *Journal of Financial Stability* 4, 89–120.

Davis, E. P., Karim D. and Liadze I. (2011): "Should Multivariate Early Warning Systems for Banking Crises Pool Across Regions?" Review of World Economics 147, pp. 693–716.

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., Detragiache, E., (2005). "Cross-country empirical studies of systemic bank distress: a survey". *National Institute of Economic Review* 192, 68–83.

Duttagupta, R., Cashin, P., (2011). "Anatomy of banking crises in developing and emerging market countries". *Journal of International Money and Finance* 30, 354–376.

Frankel, J. A. and Wei S.-J. (2004): "Managing Macroeconomic Crises." NBER Working Paper No. 10907.

Freund, Y., and Robert E. Schapire. "Experiments with a new boosting algorithm." In Machine Learning: Proceedings of the Thirteenth International Conference, pages 148–156, 1996

Ghosh, S. R. and Ghosh A. R. (2003): "Structural Vulnerabilities and Currency Crises." IMF Staff Papers 50(3), pp. 481–506.

Kaminsky, G.L., Reinhart, C.M., (1999). "The twin crises: the causes of banking and balance-of-payments problems". *American Economic Review* 89, 473–500.

Manasse, P. and Roubini,N., (2009). ""Rules of thumb" for sovereign debt crises". Journal of International Economics, 78: 192-205.

Manasse, P., Roubini N., and Schimmelpfennig A. (2003): "Predicting Sovereign Debt Crises." IMF Working Paper No. 03/221.

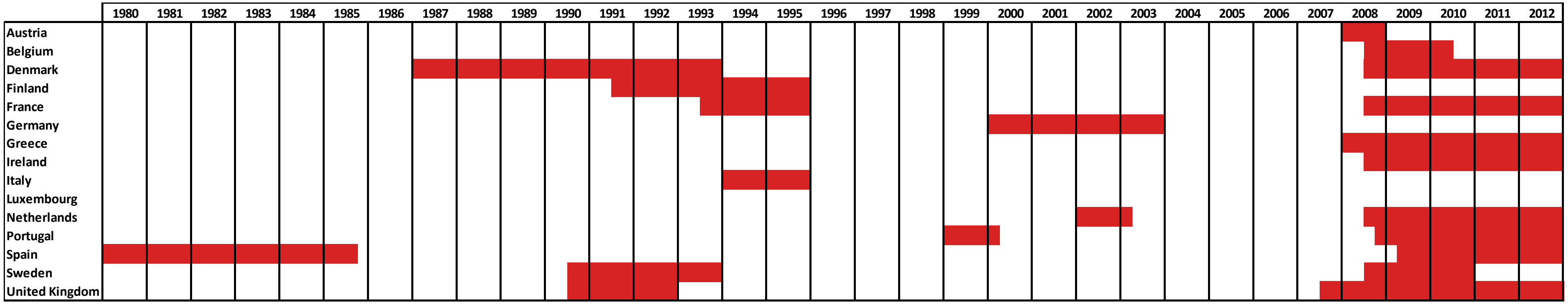
Savona, R. and Vezzoli M. (2012): "Multidimensional Distance-to-Collapse Point and Sovereign Default Prediction." Intelligent Systems in Accounting, Finance and Management 19(4), pp. 205–228.

Savona, R. and Vezzoli M. (2015): "Fitting and Forecasting Sovereign Defaults using Multiple Risk Signals." Oxford Bulletin of Economics and Statistics 77(1), pp. 66–92.

Vezzoli, M. and C. Stone (2007). "Cragging", manuscript, Department of Statistics, University of California, Berkeley also published in Book of Short Papers CLADAG 2007, University of Macerata, September 12-14, 2007.

Annex

Table 1: Dating of the crisis events

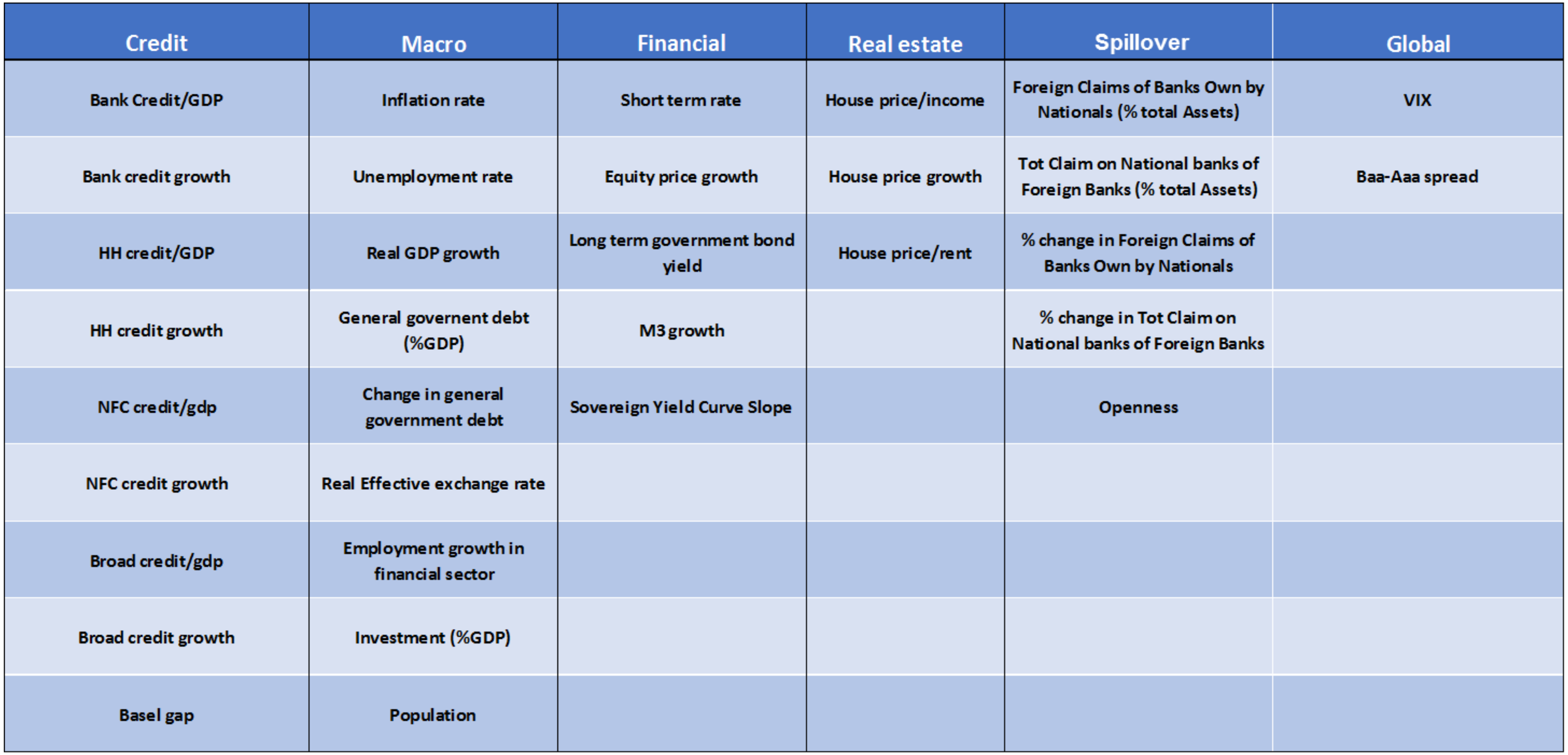


	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Austria																													🟥				
Belgium																													🟥	🟥	🟥		
Denmark								🟥	🟥	🟥	🟥	🟥	🟥	🟥															🟥	🟥	🟥	🟥	🟥
Finland												🟥	🟥	🟥	🟥	🟥																	
France														🟥	🟥	🟥													🟥	🟥	🟥	🟥	🟥
Germany																					🟥	🟥	🟥	🟥									
Greece																													🟥	🟥	🟥	🟥	🟥
Ireland																													🟥	🟥	🟥	🟥	🟥
Italy															🟥	🟥																	
Luxembourg																																	
Netherlands																							🟥	🟥					🟥	🟥	🟥	🟥	🟥
Portugal																				🟥	🟥								🟥	🟥	🟥	🟥	🟥
Spain	🟥	🟥	🟥	🟥	🟥	🟥																								🟥	🟥	🟥	🟥
Sweden											🟥	🟥	🟥	🟥															🟥	🟥	🟥		
United Kingdom											🟥	🟥	🟥															🟥	🟥	🟥	🟥	🟥	🟥

Table 2: Descriptive statistics of the dependent variable

Frequency of crisis events	15.4%
Frequency of pre-crisis periods	20.4%
Share of pre-crisis before 2008	45.6%
Share of pre-crisis from 2008 onwards	54.4%

Table 3: Early warning indicators



Credit	Macro	Financial	Real estate	Spillover	Global
Bank Credit/GDP	Inflation rate	Short term rate	House price/income	Foreign Claims of Banks Own by Nationals (% total Assets)	VIX
Bank credit growth	Unemployment rate	Equity price growth	House price growth	Tot Claim on National banks of Foreign Banks (% total Assets)	Baa-Aaa spread
HH credit/GDP	Real GDP growth	Long term government bond yield	House price/rent	% change in Foreign Claims of Banks Own by Nationals	
HH credit growth	General governent debt (%GDP)	M3 growth		% change in Tot Claim on National banks of Foreign Banks	
NFC credit/gdp	Change in general government debt	Sovereign Yield Curve Slope		Openness	
NFC credit growth	Real Effective exchange rate				
Broad credit/gdp	Employment growth in financial sector				
Broad credit growth	Investment (%GDP)				
Basel gap	Population				

Table 4: In-sample prediction comparison

	PMLF	Missed crisis rate	False alarm rate	Accuracy rate	AUROC	Threshold
Final tree (Time)	0.20	0.27	0.12	0.84	0.86	0.26
Final tree (Country)	0.24	0.16	0.33	0.71	0.79	0.27
Final tree_crag (Time)	0.16	0.22	0.11	0.87	0.91	0.30
Final tree_crag (Country)	0.28	0.29	0.27	0.73	0.75	0.26
Regression tree	0.15	0.18	0.11	0.87	0.95	0.23
Logit	0.25	0.35	0.15	0.80	0.83	0.43
Augmented Logit	0.24	0.35	0.14	0.81	0.83	0.43

PMLF is the simple average between missed crisis and false alarm rates (there could be some approximation error due to rounding). Missed crisis and false alarm rates are as a percentage of the number of observations in their respective actual events (pre-crisis or tranquil). Accuracy rate is the complement to 1 of the total error rate, irrespective of the type of error. The Area Under the Receiving Operator Curve (AUROC) is also an accuracy measure. The ROC curve is created by plotting the true positive rate (A / A+B from the confusion matrix) against the false positive rate (C/C+D from the confusion matrix, false alarm rate) at various threshold settings. The threshold column shows the splitting criteria for the probability of being in a pre-crisis period, computed using the signalling approach as described in the paper.

Table 5: Out-of-sample prediction comparison (16 quarters ahead)

	Missed crises	False alarms	Avg threshold	PMLF
Simple tree	0.61	0.23	0.30	0.42
Final tree (country)	0.34	0.40	0.20	0.37
Logit	0.46	0.24	0.22	0.35

PMLF is the simple average between missed crisis and false alarm rates (there could be some approximation error due to rounding). Missed crisis and false alarm rates are as a percentage of the number of observations in their respective actual events (pre-crisis or tranquil). The average threshold (over the evaluation period) column shows the splitting criteria for the probability of being in a pre-crisis period, computed using the signalling approach as described in the paper.

Table 6: Out-of-sample prediction comparison (1-8 quarters ahead)

	Missed crises	False alarms	PMLF
Simple tree	0.50	0.21	0.35
Final tree (country)	0.28	0.41	0.34
Logit	0.39	0.23	0.31

PMLF is the simple average between missed crisis and false alarm rates (there could be some approximation error due to rounding). Missed crisis and false alarm rates are as a percentage of the number of observations in their respective actual events (pre-crisis or tranquil).

Table 7: Out-of-sample prediction comparison (9-16 quarters ahead)

	Missed crises	False alarms	PMLF
Simple tree	0.64	0.26	0.45
Final tree (country)	0.42	0.39	0.40
Logit	0.52	0.17	0.34

PMLF is the simple average between missed crisis and false alarm rates (there could be some approximation error due to rounding). Missed crisis and false alarm rates are as a percentage of the number of observations in their respective actual events (pre-crisis or tranquil).

Figure 1: Confusion matrix

		Actual	
		Pre-crisis	Normal
Signal	Pre-crisis	A	C
	Normal	B	D

Type I error rate = $\frac{B}{A+B}$; Type II error rate = $\frac{C}{C+D}$

The EWS can have final predictions that:

1) Issues a warning signal and it is correct (A);
2) Fails to issue a warning signal when it should have signalled it (i.e., missed pre-crisis, B);
3) Issues a warning signal but it is wrong (i.e., false alarm C);
4) Does not issue a warning signal and it is right (D). These four outcomes fill the confusion matrix.

Figure 2: Sample regression tree

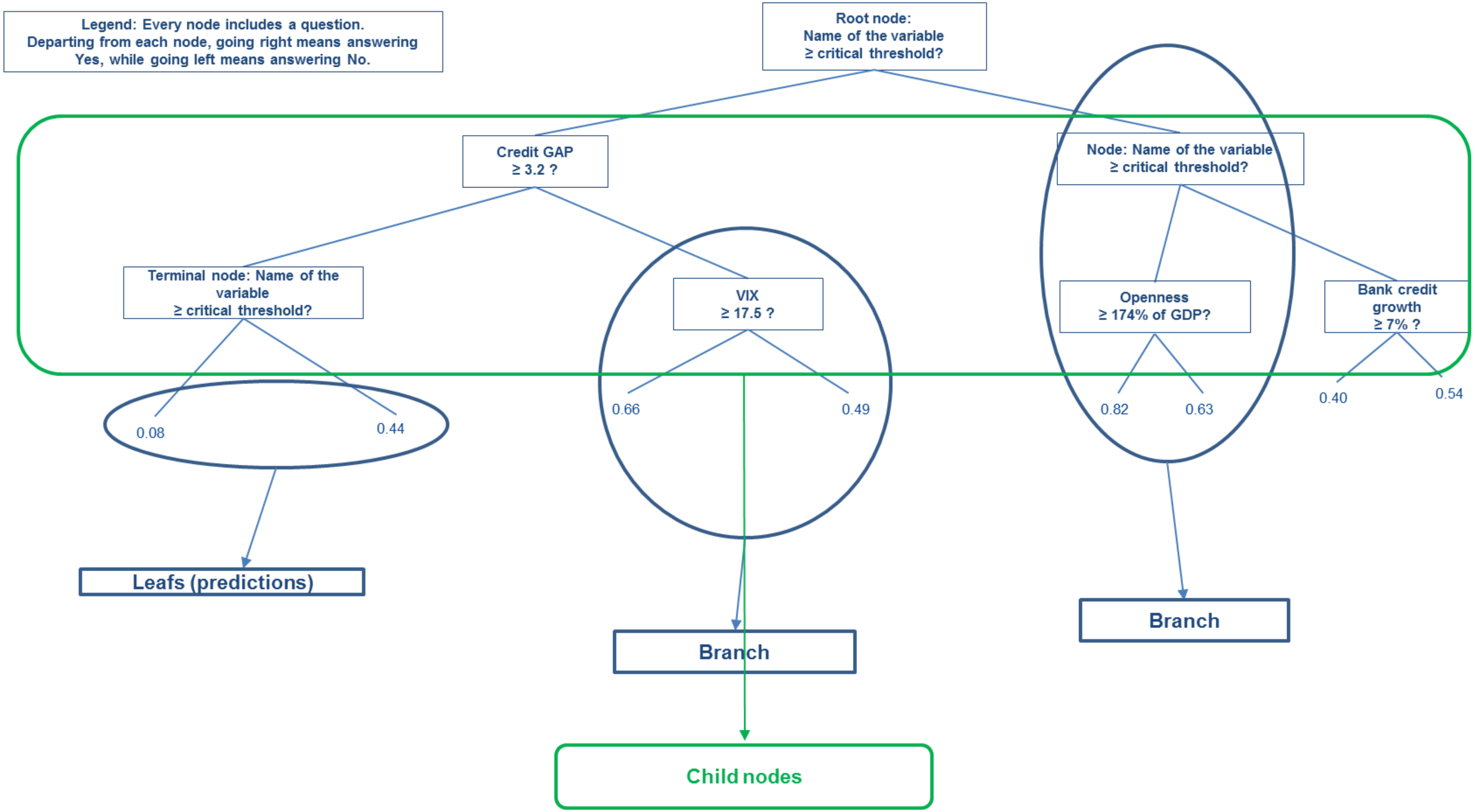

Legend: Every node includes a question. Departing from each node, going right means answering Yes, while going left means answering No.
Root node: Name of the variable ≥ critical threshold?
Credit GAP ≥ 3.2 ?
Node: Name of the variable ≥ critical threshold?
Terminal node: Name of the variable ≥ critical threshold?
VIX ≥ 17.5 ?
Openness ≥ 174% of GDP?
Bank credit growth ≥ 7% ?
0.08
0.44
0.66
0.49
0.82
0.63
0.40
0.54
Leafs (predictions)
Branch
Branch
Child nodes

Figure 3: The final tree

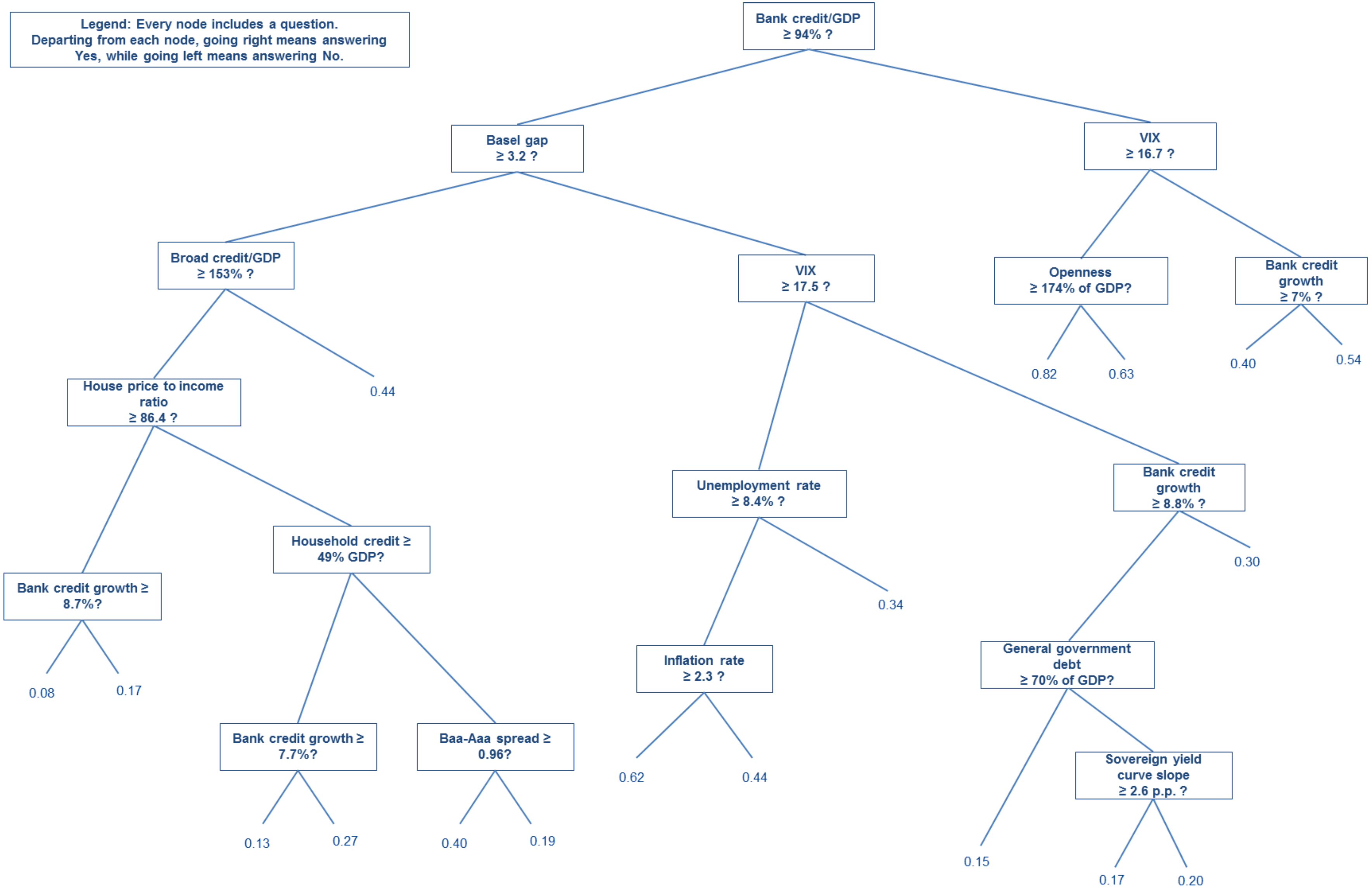
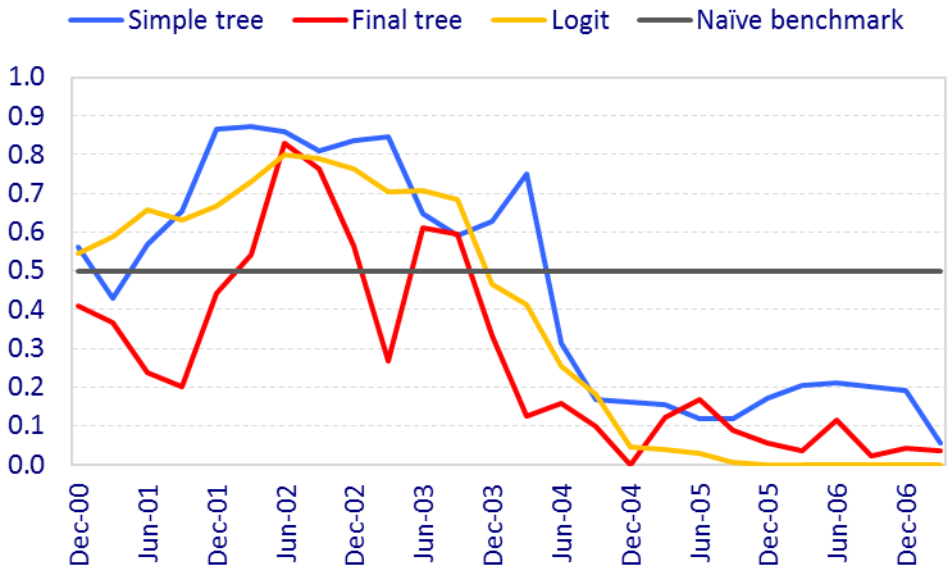
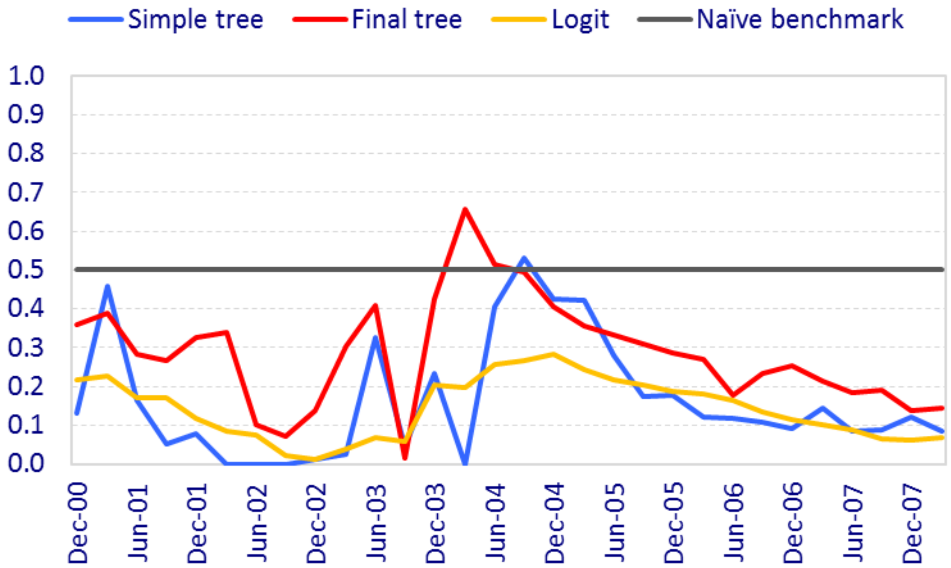

Legend: Every node includes a question. Departing from each node, going right means answering Yes, while going left means answering No.
Bank credit/GDP ≥ 94% ?
Basel gap ≥ 3.2 ?
VIX ≥ 16.7 ?
Broad credit/GDP ≥ 153% ?
VIX ≥ 17.5 ?
Openness ≥ 174% of GDP?
Bank credit growth ≥ 7% ?
0.82
0.63
0.40
0.54
House price to income ratio ≥ 86.4 ?
0.44
Unemployment rate ≥ 8.4% ?
Bank credit growth ≥ 8.8% ?
0.30
Household credit ≥ 49% GDP?
Bank credit growth ≥ 8.7%?
0.34
0.08
0.17
Inflation rate ≥ 2.3 ?
General government debt ≥ 70% of GDP?
Bank credit growth ≥ 7.7%?
Baa-Aaa spread ≥ 0.96?
0.62
0.44
Sovereign yield curve slope ≥ 2.6 p.p. ?
0.13
0.27
0.40
0.19
0.15
0.17
0.20

Figure 4: Out-of-sample missed crises rate



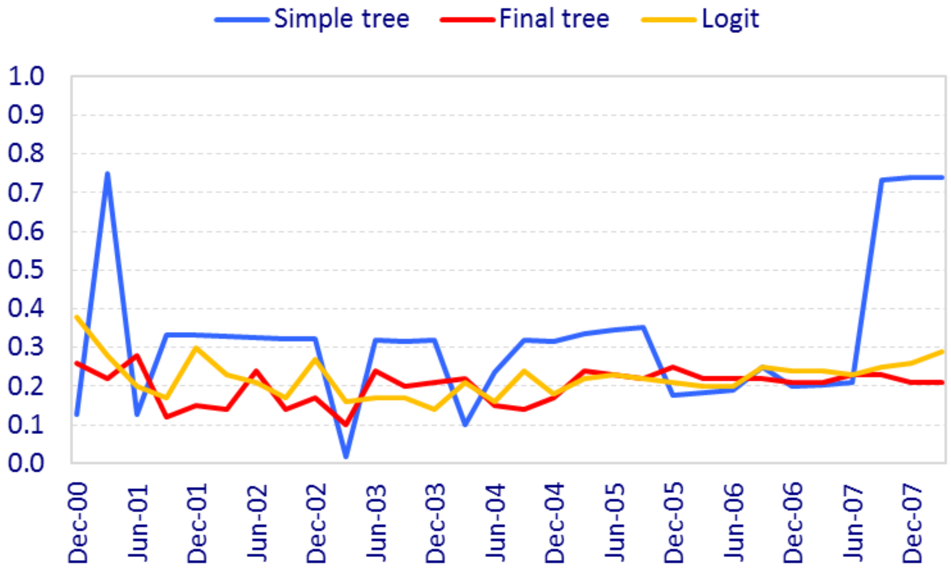
Vertical axis is in percent.

Figure 5: Out-of-sample false alarm rate



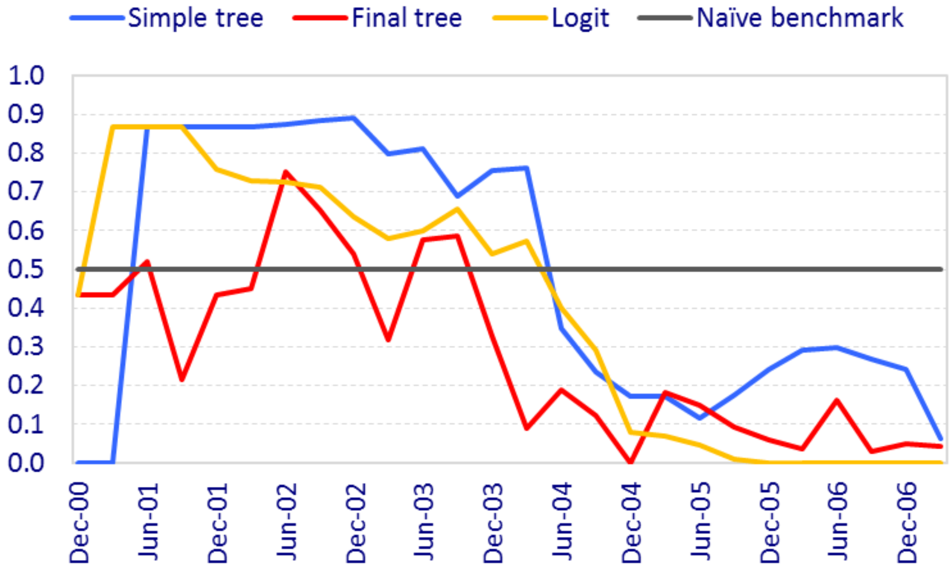
Vertical axis is in percent.

Figure 6: Out-of-sample critical thresholds



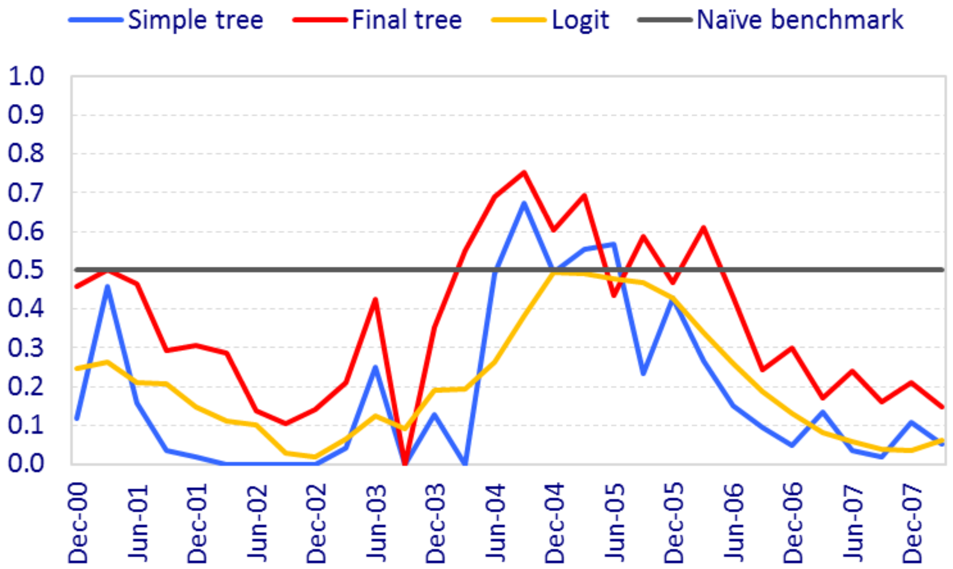
Vertical axis represents a probability.

Figure 7: Out-of-sample missed crises rate (1-8 quarters ahead)



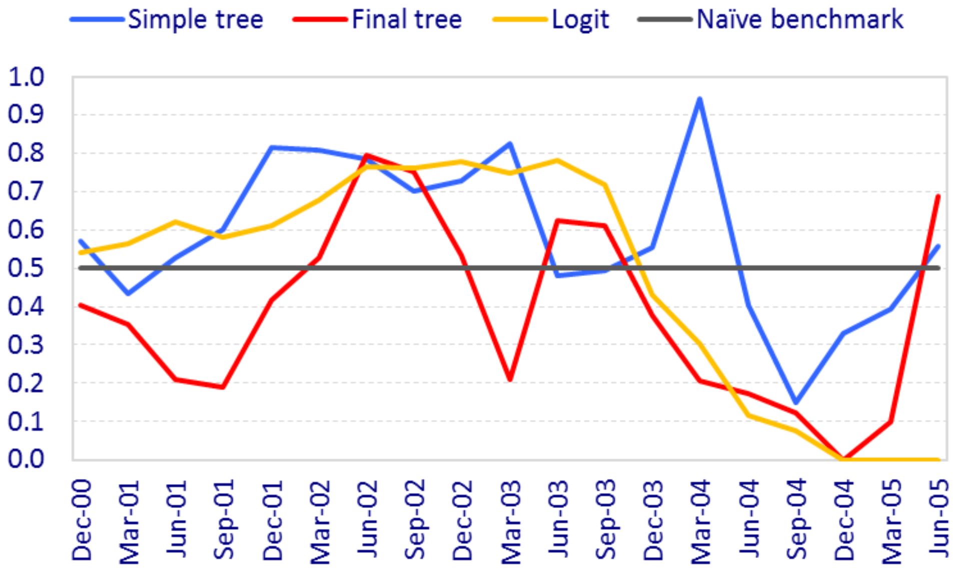
Vertical axis is in percent.

Figure 8: Out-of-sample false alarm rate (1-8 quarters ahead)



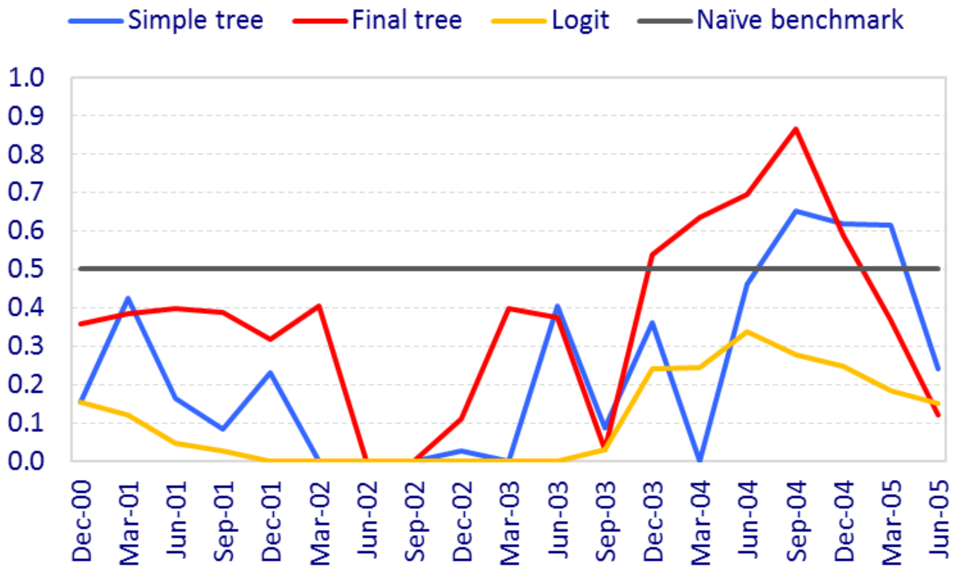
Vertical axis is in percent.

Figure 9: Out-of-sample missed crises rate (9-16 quarters ahead)



Vertical axis is in percent.

Figure 10: Out-of-sample false alarm rate (9-16 quarters ahead)



Vertical axis is in percent.

European Stability Mechanism

6a Circuit de la Foire Internationale
L-1347 Luxembourg

Tel: +352 260 292 0

www.esm.europa.eu

info@esm.europa.eu