Ward, Felix

Working Paper

Spotting the Danger Zone - Forecasting Financial Crises with Classification Tree Ensembles and Many Predictors

Bonn Econ Discussion Papers, No. 01/2014

Provided in Cooperation with:
Bonn Graduate School of Economics (BGSE), University of Bonn

Suggested Citation: Ward, Felix (2014) : Spotting the Danger Zone - Forecasting Financial Crises with Classification Tree Ensembles and Many Predictors, Bonn Econ Discussion Papers, No. 01/2014, University of Bonn, Bonn Graduate School of Economics (BGSE), Bonn

This Version is available at:
https://hdl.handle.net/10419/128613

Bonn Econ Discussion Papers

Discussion Paper 01/2014

Spotting the Danger Zone – Forecasting Financial Crises with Classification Tree Ensembles and Many Predictors
by

Felix Ward

November 2014

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Kaiserstrasse 1
D-53113 Bonn

Financial support by the
Deutsche Forschungsgemeinschaft (DFG)
through the
Bonn Graduate School of Economics (BGSE)
is gratefully acknowledged.

Deutsche Post Stiftung is a sponsor of the BGSE.

Spotting the Danger Zone

Forecasting Financial Crises with Classification Tree Ensembles and Many Predictors

Felix Ward

University of Bonn
Department of Macroeconomics and Econometrics

Version: October 24th, 2014

Abstract

To improve the detection of the economic "danger zones" from which severe banking crises emanate, this paper introduces classification tree ensembles to the banking crisis forecasting literature. I show that their out-of-sample performance in forecasting binary banking crisis indicators surpasses current best-practice early warning systems based on logit models by a substantial margin. I obtain this result on the basis of one long-run- (1870-2011), as well as two broad post-1970 macroeconomic panel datasets. I particularly show that two marked improvements in forecasting performance result from the combination of many classification trees into an ensemble, and the use of many predictors.

1 Introduction

In his 1878 publication on *Commercial Crises and Sun-Spots* William Stanley Jevons (Jevons, 1878) reached out to redraw the boundary between economic order and chaos: He doubted that economic crises are unpredictable anomalies. Instead, he proposed, they occur rather reliably about every 11 years, caused by a cycle in solar activity impacting agricultural production. Mr Jevons was careful to back up his hypothesis with quantitative evidence on solar and economic activity. Based on this evidence he was confident

enough to forecast a crisis for the year 1879. The crisis did not occur and his monocausal account was soon "made the subject of inconsiderate ridicule" (Jevons, 1879). Crisis forecasting clearly is a challenging task.

In this paper I will introduce classification tree ensembles and analyze their out-of-sample performance in forecasting the binary banking crisis indicators defined by Schularick and Taylor (2012) and Laeven and Valencia (2013) on the basis of three datasets: One long run annual dataset (1870-2011), which covers 17 developed countries, and two post-1970 datasets – one annual, one quarterly – covering 162 countries. The results suggest that the out-of-sample forecasting performance of classification tree ensembles substantially surpasses current best-practice logit specifications. To give a concrete example of the trade-offs involved, the favorite classification tree ensemble allows policy makers to correctly forecast about 50% of banking crises, at the cost of a 5% chance of wrongly forecasting a crisis when none is actually occuring. The best-practice logit specification can achieve the same 50% rate of correct crisis forecasts only at the substantially higher cost of a 25% chance of making a wrong crisis call. If policy makers should have a higher preference for making correct crisis forecasts, different bargains can be struck on the basis of the same two models: The classification tree ensemble can correctly forecast about 90% of banking crises, at the cost of making wrong crisis calls with a 25% probability. The best-practice logit specification can achieve the same 90% rate of correct crisis forecasts only at the far higher cost of an 80% chance of making a wrong crisis call. In both cases the classification tree ensemble offers the better trade-off.

My work relates to the existent literature in the following ways: First, this article adds to

the modern literature on early warning systems for banking crises, which was pioneered by Kaminsky (1998) and Kaminsky and Reinhart (1999) in the wake of the 1997 Southeast Asian crises.[1] More recent contributions analyzing the predictability of banking crises in developed economies in the long-run since 1870 are due to Schularick and Taylor (2012) and Jordà (2013), while others rely on post-1970 samples covering many countries (see Drehmann and Juselius (2012) and Drehmann (2013)). This literature has shown that already relatively simple model structures based on few predictors - most notably credit aggregates - can convey valuable information on the imminence of a banking crisis. The main contribution of this paper will consist in the exploration of whether somewhat more complex model structures based on many predictors can improve financial crisis forecasts.

This article is thus also related to the literature on economic forecasts based on many predictors (see Stock and Watson, 2002, 2006). This literature has stressed the possibility of improving economic forecasts by basing them on a larger set of economic indicators. Here I apply this logic to the banking crisis forecasting task. In particular, I show that increasing the number of predictors from the 7-10, typically applied in current best-practice early warning systems, to about 70-80, markedly improves banking crisis forecasts. Whereas the literature on economic forecasting based on many predictors has focused largely on factor modelling and prestep-dimensionality reduction techniques, such approaches do not easily lend themselves to banking crisis forecasting. For once, most banking

[1]Kaminsky (1998) and Kaminsky and Reinhart (1999) also analyze the predictability of currency crises. Here, I will focus exclusively on severe, systemic banking crises.

crisis indicators are binary 0-1 dummies which require discrete classification techniques.[2] Furthermore, widely held beliefs on the genesis of banking crises, namely that they are characterized by discontinuous threshold effects and nonlinear interaction effects between several risk factors (see Duttagupta and Cashin, 2011), are more naturally accommodated by methods which dispense with linearity assumptions from the outset. I therefore turn to classification tree structures which naturally accommodate discontinuous threshold effects as well as nonlinear interactions between several predictors. Their ability to thus precisely delineate several "danger zones", *and* their ability to harness many predictors in doing so, has already made them a mainstay in other research areas, such as genetics, where often thousands of genetic markers are analyzed with respect to their contributions to particular diseases (e.g. Díaz-Uriarte and De Andres, 2006).

Methodologically this article draws from a genre of statistical techniques, which runs under the various headings of "machine learning", "artificial intelligence" or "nonparametric statistics", reflecting their diverse origins. While diverse in origin, these approaches share certain generic similarities. They tend to sacrifice ease of model interpretability for predictive accuracy. This is achieved through rather flexible model structures, which put increased demands on computation power and data amounts. Given the former is getting ever cheaper, and the latter ever more abundant, these approaches have become increasingly applicable – from internet search engines and trading algorithms to bioinformatics and astrophysics (e.g. Albert et al., 2008). Beyond their generic similarities however, methodologies in this field differ. This renders careful model selection necessary

[2]Exceptions are continuous crisis indices such as the exchange market pressure index pioneered by Eichengreen, Rose, and Wyplosz (1994). Such indices are available for fewer countries and cover shorter time-spans than their binary counterparts.

– particularly so since severe banking crises are rare events: The datasets I use contain almost all known systemic banking crises of the last 150 years – their absolute number however still lies only somewhat above 200. Simply comparing various forecasting methods according to out-of-sample forecasting performance would thus quickly run into the multiple testing problem. I thus restrict my analysis to classification trees (Breiman et al., 1984) and their ensembles (Breiman, 1996, 2001), which ex ante appear to be a particularly promising candidate for banking crisis forecasting for several reasons:

First, classification tree ensembles are able to combine the information from many predictors into an overall crisis risk assessment, while circumvening the curse of dimensionality. This nicely conforms to the multicausality of banking crises, which may announce themselves by developments on the real- or nominal side of the economy – they may originate from within the domestic financial system or spill over from abroad. Furthermore, critical information may be contained in different frequency bands of the same predictor. For example, credit levels, their growth rates, as well as their deviations from trend may indicate an increase in banking crisis risk. Thus, the number of predictors which cannot be excluded from analysis on convincing a priori grounds is quite large. This renders an classification tree ensemble's ability to accommodate many predictors valuable for banking crisis forecasting.

Second, classification tree structures are built to capture nonlinear interaction effects between several predictors. Whether a string of unexpected inflation events announces a banking crisis or not for example, may depend on the level of nominally fixed debt contracts on the banks' balance sheets.

Third, classification tree structures are also built to accommodate discontinuous threshold effects between any single predictor and crisis risk, while maintaining a sufficient degree of flexibility to also approximate continuous nonlinear- as well as linear relationships. Nonlinear effects have often been associated with the onset of banking crises. In this way, many accounts of severe banking crises don't begin with the description of an equally severe shock – often they begin with an economically comparatively minor incidence, such as an increase in the default rate on US subprime mortages in 2006 and 2007.

Finally, despite their structural suitability, and their ability to accommodate many predictors, the forecasting performance of single classification trees is held back by their high variability. This high variability however can be countered by combining many trees into an ensemble – a technique which will be described in more detail in section 3.

This article is structured as follows: Section 2 will point out the main constraints which hold back current early warning systems (EWS), including the signals approach, logit models and single classification trees. Section 3 provides a description of how classification tree ($\mathcal{CT}$)-ensembles work around these constraints. Section 4 introduces the datasets. These datasets form the basis for the out-of-sample forecasting contest between $\mathcal{CT}$-ensembles and representative logit specifications in section 5. Section 6 concludes by showing how a particular $\mathcal{CT}$-ensemble, random forest, would have fared in forecasting the 2007/2008 financial crisis.

2 Constraints of Current EWS

The EWS literature is dominated by two methodologies: the *signals* approach and *probit/logit* models. Both have been shown to offer valuable information on the imminence of banking crises. More recently, Davis and Karim (2008*a*) and Duttagupta and Cashin (2011) have begun to explore the potential of a third method – *classification trees*. The forecasting performance of these methods is held back by several constraints, which I will discuss in this section. In the next section I will then show how $\mathcal{CT}$-ensembles improve upon this. In general, current EWS face at least one of the following four constraints:

(i) First, the amount of predictors they can handle.

(ii) Second, the ease with which they accommodate nonlinear interaction effects.

(iii) Third, the ease with which they accommodate nonlinear effects between any single predictor and crisis risk.

(iv) Fourth, the degree of estimator variability.

It will be helpful to shortly discuss how currently applied EWS are held back by these.

2.1 The Signals Approach

Signals models were first introduced by Kaminsky and Reinhart (1999). These models issue a warning signal whenever a predictor passes a certain threshold. Signals models can be expressed as

$$y_i = I(x_i > t) + \epsilon_i, \tag{1}$$

where for forecasting purposes y_i is a horizon dummy, which takes the value 1 in the two year horizon preceding a crisis (from now on also referred to 'crisis observations' for ease of exposition), and 0 otherwise (from now on also referred to as 'no crisis observations'). $I(\cdot)$ denotes an indicator function, x_i is a predictor, t a threshold and ϵ_i an error term

$$\epsilon_i = \begin{cases} 0, & \text{for correct signals: } I(x_i > t) = y_i \\ 1, & \text{otherwise,} \end{cases}$$

for observations $i = 1, ..., N$. A threshold $\hat{t}$, also called *split point*, is estimated via minimization of a loss function – most commonly the *adjusted noise to signal ratio* ($aNtS$): $\hat{t} = \arg\min_t\{aNtS(t)\}$, where $aNtS = \frac{FPR(t)}{TPR(t)}$, and $FPR(t)$ and $TPR(t)$ denote the false- and true positive rate, depending on threshold t. Based on $\hat{t}$ it is possible to make crisis predictions $\widehat{y}_i = I(x_i > \hat{t})$.

While contributions based on the signals approach have made clear that even single predictors contain valuable information on the imminence of a banking crisis, the simplicity of the approach constrains forecasting accuracy in several ways:

First, the indicator function in (1) rules out potential non-discontinuities in the build-up of banking crisis risk. Davis and Karim (2008*a*) for example have shown that continous (though also nonlinear) logit specifications do a better job in forecasting crises.

Second, in terms of variable selection, the approach focuses on only one predictor at a time. To the extent that models of the form (1) are estimated and evaluated for several predictors the *multiple testing* problem kicks in. Later contributions have aimed at combining two to three predictors in a slightly adjusted model of the form

$$y_i = min\{I(x_i > t_x), I(y_i > t_y)\} + \epsilon_i, \tag{2}$$

with predictor specific thresholds t_x and t_y (e.g. Borio and Lowe (2002)). However, the inclusion of two to three predictors still falls short of reflecting the multicausal nature of crises.

Third, in equation (2) a warning signal is only issued when both predictors x_i and y_i exceed their thresholds. This allows for the analysis of only one first-order interaction effect, which has to be selectively specified by the researcher through the selection of predictors x and y.

2.2 Probit and Logit Specifications

Probit and *logit* specifications constitute the main workhorses when it comes to binary classification or probability modelling across many disciplines. This certainly holds true for banking crisis forecasting, where logit specifications dominate. The first contribution in this vein is usually attributed to Demirgüç-Kunt and Detragiache (1998). More recent contributions relying on logit specifications come from Frankel and Saravelos (2012) and Schularick and Taylor (2012). The logit model's structure can be given as

$$ln\left(\frac{P(y_i = 1)}{1 - P(y_i = 1)}\right) = \beta X_i, \tag{3}$$

where y_i is a horizon dummy, for observations $i = 1, ..., N$. β is a $1 \times J$ vector of coefficients and X_i is a $J \times 1$ vector of predictors. Once β has been estimated by numerical maximum likelihood (ML) procedures, the probability of being in the pre-crisis horizon can be prediced as $\widehat{P}(y_i = 1) = \Lambda(\widehat{\beta} X_i)$, where $\Lambda(\cdot)$ denotes the logistic cumulative distribution function.

Multivariate logit models have been shown to slightly outperform signals models in

banking crisis forecasting (Davis and Karim, 2008*a*). Nevertheless their performance is constrained by several factors:

First, ML-estimation of the coefficient vector β runs into the curse of dimensionality as the number of predictors increases. Thus restricted variable selection and/or dimensionality reduction is a necessary prestep for logit models.

Second, though equation (3) captures a particular kind of nonlinearity its structure is rigid. More flexible functional forms and variable interactions have to be explicitly added by the researcher through, for example, higher order- and interaction terms, which again runs into the curse of dimensionality.

2.3 Classification Tree Analysis

More recent contributions which explore the potential of *classification trees* for the analysis of banking crises are due to Davis and Karim (2008*a*) and Duttagupta and Cashin (2011).[3] Formally, a classification tree predicts the probability of being within the pre-crisis horizon as

$$\widehat{\mathcal{T}}(X_i) = \sum_{m=1}^{M} \widehat{p}_m I(X_i \in \widehat{R}_m), \tag{4}$$

where $\widehat{\mathcal{T}}(X_i)$ is the probability estimate of being in the pre-crisis horizon (subsequently also termed 'crisis probability'), depending on a $J \times 1$ vector of predictor values X_i, for observations $i = 1, ..., N$. $I(\cdot)$ is an indicator function and $\{\widehat{R}_m\}_{m=1}^{M}$ is a set of M region estimates in J-dimensional predictor space, to each of which a crisis probability

[3]See Kaminsky (2006) for an application of classification tree analysis to currency crisis classification. Also see Marais, Patell, and Wolfson (1984) and Frydman, Altman, and Kao (1985) for early applications of classification tree techniques to financial distress detection on a micro-level.

estimate $\widehat{p}_m$ is attached. Given the region estimates $\{\widehat{R}_m\}_{m=1}^M$, probabilities $\widehat{p}_m$ are simply estimated as the proportion of crisis observations in region $\widehat{R}_m$, for $m = 1, ..., M$:

$$\widehat{p}_m = \frac{\sum_{i \in \widehat{R}_m} y_i}{\sum_{i \in \widehat{R}_m} 1} \tag{5}$$

The difficulty however lies in obtaining region estimates $\{\widehat{R}_m\}_{m=1}^M$. Globally optimal estimation of all parameters $\Theta = \{R_m, p_m\}_{m=1}^M$ constitutes a NP-complete problem (Hyafil and Rivest, 1976). Thus classification trees are typically estimated through *recursive partitioning* – a greedy search algorithm which engages in step-wise, locally optimal estimation.

Figure 1 illustrates how recursive partitioning works. To simplify matters a four step – two predictor example is illustrated. The two-dimensional predictor space spanned by x_1 and x_2 is populated by observations. The filled circles stand for crisis observations, the empty circles stand for no crisis observations. In several steps the predictor space is partitioned into five terminal regions $R_{1,2,3,4,5}$, separated by the four splits $t_j^{1,2,3,4}$ (dashed lines), where $j \in \{1, 2\}$ indicates the split predictor. The exact rule according to which these splits are conducted will be introduced in the next paragraph. In the first step, illustrated in the upper left panel, the split point t_1^1 partitions the sample along the range of the splitting predictor x_1. In the second step (upper right panel) t_1^2 partitions the sample once more along the range of the splitting predictor x_1. In a third and fourth step, illustrated in the lower left panel, the same procedure is repeated for two of the subsamples obtained from the first two steps. In this vein at each recursive partitioning step a 'parent'-region is divided into two 'child'-regions. The splitting stops in one of two ways: Either the terminal regions contain only crisis or no crisis observations, as is

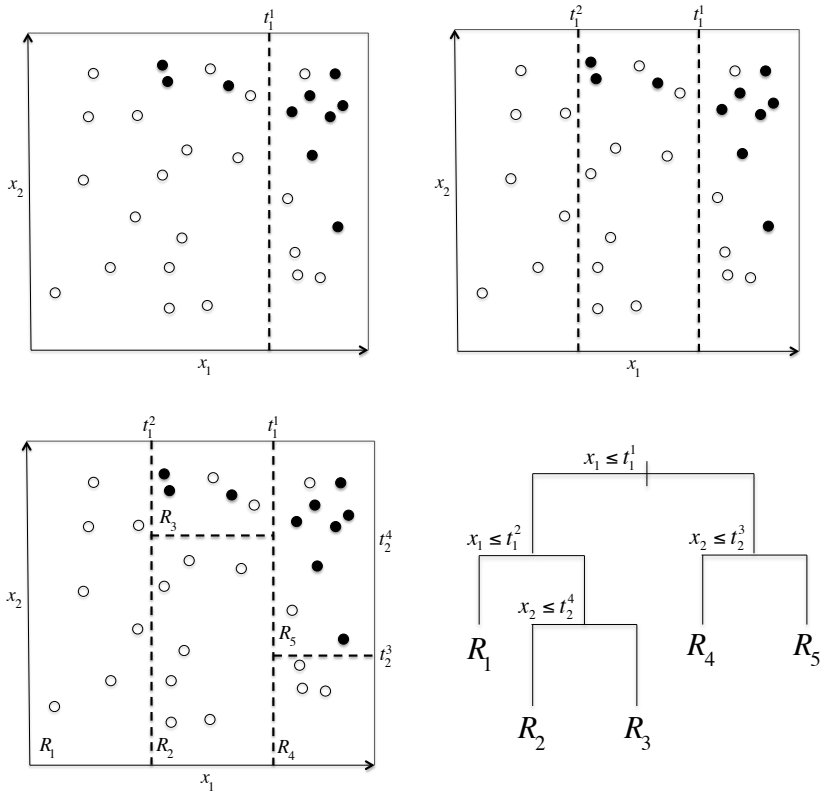

Figure 1: Recursive Partitioning: An Illustration

Notes: Upper left panel shows the first recursive partitioning step. Upper right panel shows the second recursive partitioning step. Lower left panel shows a third and fourth partitioning step. Lower right panel shows the tree corresponding to the partition in the lower left panel. Filled circles – $y_i = 1$ (Crisis); empty circles – $y_i = 0$ (No crisis). x_j – predictors. t_j^s – splits. R_m – terminal regions.

the case in regions $R_{1,2,4}$, or an ad hoc stopping rule bites, such as a given maximum number of splits. In this vein the lower left panel in figure 1 could be the result of a partitioning where the maximum number of splits is set to four. The final partition can also be represented as a tree structure (lower right panel).

To determine a splitting predictor and a split point at each recursive partitioning step a splitting rule has to be applied. The *gini impurity* criterion is such a splitting rule. The gini impurity of region R_a is defined as

$$\mathcal{GI}(p_a) = -2p_a^2 + 2p_a,$$

where p_a denotes the proportion of crisis observations in region R_a. $\mathcal{GI}(R_a)$ reaches minima of 0 when p_a equals 1 or 0, and a maximum of 0.5 when p_a equals 0.5. The gini impurity criterion thus takes high values for regions where crisis and no crisis observations mix, and low values in 'pure' regions. In the end this property allows for a separation of crisis and no crisis observations. Based on gini impurity and given a certain split, which determines the child regions, the *information gain* $\mathcal{IG}$ can be calculated as the difference between the parent region-gini impurity and the average child region-gini impurity:

$$\mathcal{IG}(R_a, R_b) = \mathcal{GI}(R_a \cup R_b) - 0.5\big[\mathcal{GI}(R_a) + \mathcal{GI}(R_b)\big],$$

where $\mathcal{GI}(R_a \cup R_b)$ stands for the gini impurity of the parent region. $\mathcal{GI}(R_a)$ and $\mathcal{GI}(R_b)$ denote the gini impurities of the child regions.

At each recursive partitioning step $s = 2, ..., S$ a splitting predictor j and a split point t along the range of this splitting predictor are selected such as to maximize the resulting information gain:

$$\hat{t}_j^s = \arg\max_{j,t} \mathcal{IG}\Big(R_a^s(t_j|\hat{t}_j^1, ..., \hat{t}_j^{s-1}), R_b^s(t_j|\hat{t}_j^1, ..., \hat{t}_j^{s-1})\Big) \tag{6}$$

Only the first split $s = 1$ is an unconditional one; all others depend on all previously estimated splits $\hat{t}_j^1, ..., \hat{t}_j^{s-1}$. The idea behind (6) is to partition a parent region into two child regions in such a way that crisis and no crisis observations get separated into different regions. Note that recursive partitioning is robust to outliers as extreme values do not influence the internally optimal split point. This property is especially convenient given the occasional extreme-value observations in samples covering many countries and long time-spans. Recursive partitioning can end in two ways: It can either run its course

until the classification tree has been "fully grown" and only 'pure' regions of exclusively crisis or no crisis observations are left. Alternatively, recursive partitioning can be ended through an ad hoc stopping rule, such as a minimum number of observations per region. Usually the application of such an ad hoc stopping rule is necessary to avoid severe overfitting. In the following I constrain the terminal region size of single classification trees to 10 observations. The single tree results are not unduely affected by this decision and hold up under different stopping rules.

In any case, the final partitioning constitutes an estimate of the M terminal regions $\{\widehat{R}_m\}_{m=1}^{M}$, on the basis of which the classification tree (4) can be completed by estimating crisis probabilities according to (5).

In many respects classification trees appear to be suitable candidates for banking crisis forecasting: Firstly, they can be estimated on the basis of many predictors without running into the curse of dimensionality – a positive side-effect of the step-wise estimation through recursive partitioning. Secondly, they naturally accommodate nonlinear interaction effects. Thirdly, beyond discontinuous threshold effects, classification trees can accommodate more and more functional forms as they grow larger. Despite all these favourable characteristics classification trees have been associated with poor out-of-sample forecasts for banking crises (see Davis and Karim, 2008*a*). What is the reason for this?

The overriding constraint which holds back the forecasting performance of single classification trees is their high variability – an unwelcome side-effect of recursive partitioning: Small changes in the sample under analysis can easily change splitting predictors and split points in the early partitions. This change then reverberates throughout all subsequent

steps, as each further partition is conditional on all previous ones. In this way only small sample changes can have profound effects on the final model. As section 5 will show, this instability deals a severe blow to the forecasting performance of single classification trees. Fortunately, as will be explained in the next section, combining many classification trees into a $\mathcal{CT}$-ensemble constitutes a solution to the problem of high variability.

In sum, current signals- and logit EWS are held back by their inability to accommodate (i) many predictors and (ii) nonlinear predictor interactions, as well as (iii) an overly rigid functional forms. While classification trees seem to be suitable candidates for banking crisis forecasting with respect to (i) - (iii) they are plagued by (iv) high estimator variability. The next section will introduce $\mathcal{CT}$-ensembles and explain how they constitute an improvement over single classification trees in terms of (iv) estimator variability, but also (iii) the degree of functional flexibiliy.

3 Why Classification Tree-Ensembles?

As the name suggests a *forest* $\mathcal{F}$ consists of many classification trees $\mathcal{T}_b$, $b = 1, ..., B$. Each individual tree 'grows' on a bootstrap-sample X^b from the original data X. Such bootstrapping with subsequent aggregation is referred to as *bagging* (Breiman, 1996). If each tree is given the same weight a forest predicts the probability of being in a pre-crisis horizon as

$$\widehat{\mathcal{F}}(X_i) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mathcal{T}}_b(X_i), \tag{7}$$

where X_i is the $J \times 1$ vector of predictor values for observations $i = 1, ..., N$. Thus the forest prediction is simply the average prediction of the B single trees.

The linear combination of B trees into a forest preserves the beneficial properties of an individual tree: (i) A forest can be estimated on the basis of many predictors, as the same step-wise local estimation procedure is applied. (ii) Like any single tree, a forest can also accommodate nonlinear interaction effects. In terms of (iii) functional flexibility and (iv) estimator variability however a forest is more than the sum of its parts:

As regards functional flexibility an average of several models has a *representational advantage* over any single model (Dietterich, 2000). A forest for example can better approximate continuous relationships between crisis risk and its predictors as the averaging over many trees smoothes the jump discontinuities of any single tree. Thus to the extent that classification tree structures are in principle fit for banking crisis forecasting, forests are even more so.

Next, how does bagging address estimator variability? The intuition is conveyed by the *wisdom of crowds*-mechanism: While most people are probably rather ignorant on the number of beans contained in one can, some individuals happen to be more knowledgeable – be it because of a special talent in spatial visualization, or because they work at the factory filling the cans. As long as the guesses of the ignorant individuals are randomly distributed around the true value, the more knowledgable individuals' guesses determine the mean of the overall guess-distribution. Under such circumstances averaging is a simple aggregation-mechanism, which provides a good estimate of the true number of beans in a can.[4] The guess of any randomly drawn individual however ranges from 'very close' to 'far off' the true value, i.e. has high variability. In a similar vein each tree in a

[4]Note that the wisdom of crowds is no unconditional law. It can easily give way to what might be termed the *madness of crowds*, where individual guesses are not independent of each other and may veer further and further away from the true value.

forest grows on a different bootstrap-sample and thus emboddies a different experience. Through aggregating these experiences a forest is stripped of the high variability of its constituent trees.

A stylized formal argument is helpful in further clarifying the variability-reducing effect of bagging. Under certain assumptions it can be shown that individual estimators $\widehat{\mathcal{T}}_b(x)$ have higher *mean squared error* (MSE) than bagged ones $\mathcal{F}(x)$ due to their variability (see Hastie, Tibshirani, and Friedman, 2013, p. 285):

$$\begin{aligned} E(y - \widehat{\mathcal{T}}_b(x))^2 =& E(y - \mathcal{F}(x))^2 + E(\widehat{\mathcal{T}}_b(x) - \mathcal{F}(x))^2 \\ \geq& E(y - \mathcal{F}(x))^2, \end{aligned} \tag{8}$$

where y is an output variable and x a fixed input. $\widehat{\mathcal{T}}_b(x)$ is estimated on the basis of a bootstrap dataset X^b, which is sampled from the actual population. Though stylized, this formulation clarifies how for an ideal bagging estimator $\mathcal{F}(x) = E(\widehat{\mathcal{T}}_b(x))$ bagging reduces the MSE simply by eliminating the variance of $\widehat{\mathcal{T}}_b(x)$ around its expected value $\mathcal{F}(x)$.[5]

Note how the inequality in (8) turns into an equality if all trees in an ensemble are identical, i.e. $\widehat{\mathcal{T}}_b(x) = \mathcal{F}(x)$, for $b = 1, ..., B$. Thus, to reap the benefits of bagging the individual trees have to differ from each other. This is usually ensured through the random generation of the B bootstrap samples. Above and beyond that however, weakening the similarities between individual classification trees further, can enhance the variability-reducing effect of bagging. This is the idea behind a particular variant of

[5]In positing that $\mathcal{F}(x) = E(\widehat{\mathcal{T}}_b(x))$ the argument abstracts from the unresolved problem of bootstrap consistency in bagging classification trees (see Bühlmann and Yu (2002) and Bühlmann (2012)). Although a formal proof does not yet exist, evidence from simulation analyses as well as practical applications suggest the argument is broadly applicable.

$\mathcal{CT}$-ensemble, *random forest* (Breiman, 2001).

As the name suggests *Random* Forest ($\mathcal{RF}$) introduces some extra randomness into the tree growth process such as to increase the diversity of the forest. In particular, $\mathcal{RF}$ aims at increasing tree diversity through the following randomization mechanism: At each recursive partitioning step s in (6) only a random subset $try^s \subset \{1, ..., J\}$ containing J_{try} of the total of J predictors in X^b is drawn (without replacement) and checked for their split potential:

$$\tilde{t}_j^s = \arg \max_{j \in try^s, t} \mathcal{IG}\left(R_a^s(t_j|\tilde{t}_j^1, ..., \tilde{t}_j^{s-1}), R_b^s(t_j|\tilde{t}_j^1, ..., \tilde{t}_j^{s-1})\right) \tag{9}$$

The size of the random subset J_{try} will be set to the recommended default of $\lfloor\sqrt{J}\rfloor$ (Breiman, 2002) throughout the following analysis. Consequently the similarity between the individual trees constituting a random forest, $\{\widehat{\mathcal{T}_b^{RF}}(X_i)\}_{b=1}^B$, is lower than the similarity between the previously discussed non-randomized trees, $\{\widehat{\mathcal{T}}_b(X_i)\}_{b=1}^B$. This increases the effectiveness of the bagging mechanism, which has been termed the *decorrelation effect* (Hastie, Tibshirani, and Friedman, 2013, p. 597).

A short formal argument helps to further clarify the decorrelation effect: The variance of an average of B identically, but not independently distributed trees (trees are not independent of each other due to overlapping bootstrap samples) with variability σ^2 is

$$\sigma_{bag}^2 = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2,$$

where ρ is the pairwise correlation between two trees. While the second term is brought to zero through bagging as the number of trees B increases $\mathcal{RF}$ aims at directly decreasing ρ through adding some randomness to the tree growth process. Randomization thus reduces the lower bound in variability, $\rho\sigma^2$, which is obtainable through bagging.

Except for the additional randomization, the random forest estimator equals the previously discussed forest in (7). By analogy the random forest estimator can be written as

$$\widehat{\mathcal{RF}}(X_i) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathcal{T}_b^{RF}}(X_i). \tag{10}$$

A concise overview of all the steps involved in generating the $\mathcal{F}$- and $\mathcal{RF}$-estimators (7) and (10) can be found in algorithm A1 shown in the appendix.

With respect to the stopping rules for recursive partitioning, there exists a notable difference between single classification trees and $\mathcal{CT}$-ensembles: Fully growing a single classification tree usually results in severe overfitting, making ad hoc stopping rules preferable. This is not the case for an ensemble of trees, for which stopping rules appear to be rather uninfluential with respect to overfitting (see Segal, 2004; Hastie, Tibshirani, and Friedman, 2013, p. 596). Fully growing each tree in an ensemble thus has established itsself as a standard. In the following analysis I will adhere to this standard for another reason. Given the rarity of severe banking crises and the fact that a dummy variable is a rather noisy indicator of them, I deem setting aside part of the data for the estimation of an optimal stopping rule to be an excessive strain on the samples.

To sum up, how do $\mathcal{CT}$-ensembles constitute an improvement over single classification trees? First, the averaging over many classification trees allows $\mathcal{CT}$-ensembles to better approximate smooth relationships. Second, $\mathcal{CT}$-ensembles are more stable than single classification trees, as bagging and randomization counter estimator variability.

4 Data

In this section I introduce three datasets on the basis of which I will evaluate the forecasting performance of logit models, single classification trees and $\mathcal{CT}$-ensembles. Systemic banking crises are rare. Their statistical analysis thus necessitates datasets, which cover large time spans or many countries – one usually comes at the cost of the other. I therefore make use of one long-run sample spanning from 1870 to 2011, as well as two post-1970 samples with broader country coverage.

4.1 The Long-Run Sample, 1870-2011

As regards the long-run sample I use the dataset introduced by Schularick and Taylor (2012). After further extensions this dataset now ranges from 1870 to 2011 and covers 17 countries (Jordà, Schularick, and Taylor, 2013). Most of the time from 1870 to 2011 these 17 countries together make up more than half of world GDP (according to Maddison GDP estimates).

The dataset features macroeconomic indicators (GDP, consumption, investment, consumer prices, the current account and exchange rates) as well as financial indicators (bank loans, total bank assets, stock prices, interest rates, public debt and monetary aggregates). These are the base indicators from which I then derive 77 predictors. Schularick and Taylor (2012) also provide a binary banking crisis indicator, the definition of which follows Laeven and Valencia (2008): The indicator takes a value of 1 for years characterized by bank runs, a jump in default rates and large capital losses associated with public interventions as well as bankruptcies or forced mergers of major financial institutions –

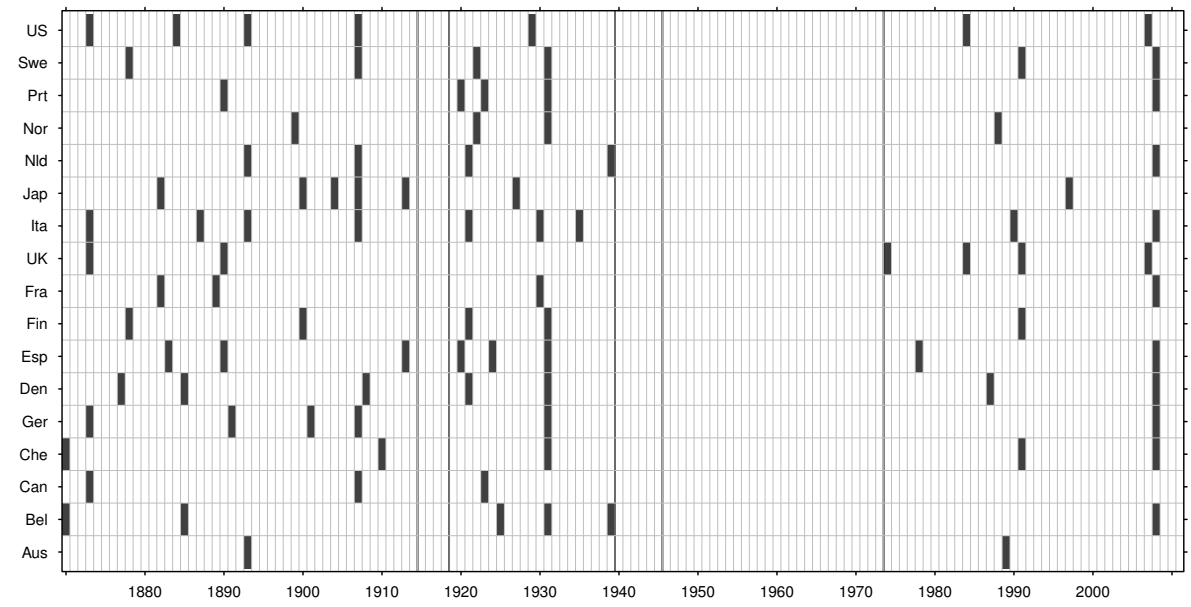

Figure 2: Crisis Map: Long-Run Sample, 1870-2011

Notes: Systemic banking crises (black). Vertical bars separate the pre-WW1 period, WW1, interwar period, WW2, the Bretton Woods period and the post-Bretton Woods period. Source: Systemic banking crisis dummies from Schularick and Taylor (2012) and Jordà, Schularick, and Taylor (2013) (extended dataset).

otherwise the indicator takes a value of 0. Figure 2 gives an overview of the 93 systemic banking crises contained in the dataset.

4.2 The Broad Post-1970 Samples

As regards the post-1970 period I make use of the binary banking crisis indicator provided by Laeven and Valencia (2013). This indicator covers 162 countries and spans the years 1970 to 2011. Its definition equals the one by Schularick and Taylor (2012). Figure 3 gives an overview of the 147 systemic banking crises contained in this dataset.

Next, I obtain annual and quarterly base indicators from the IMF IFS-database and match them to the banking crisis indicator. In my selection of base indicators the availabiltiy of

a series across many countries is paramount, as many missing values would endanger the already small number of financial crises even further. The annual indicators cover consumer prices, net exports, exchange rates, bank loans, stock prices, interest rates and public debt (provided by Abbas et al., 2013). The quarterly indicators cover GDP, consumer prices, exchange rates, bank loans, stock prices, house prices, interest rates, foreign liabilities and reserves. Furthermore, for the annual post-1970 sample I make use of the GDP, consumption and investment estimates from the Penn World Tables (Feenstra, Inklaar, and Timmer, 2013).

Together, the long-run and the post-1970s sample account for almost the entire population of systemic banking crises in modern history.

4.3 The Predictors

For each sample I derive about 70 predictors from the base indicators (see tables A1, A2 and A3 in the appendix). I make use of the bare nominal series (n) where they are of interest (e.g. nominal interest rates), but also obtain CPI-deflated quantities, growth rates (gr), trend deviations (gap), "to GDP" ratios (/GDP), global (GDP-weighted) averages (glo), real exchange rates and interest rate differentials (see Alessi and Detken, 2011, for a similar approach). Furthermore, I combine several of these transformations where deemed appropriate – for example, to obtain the gap of the loans to GDP ratio ("Loans/GDP (gap)"). In this way I obtain a detailled snapshot of economic conditions.

I now give a more precise account of the transformations I conduct: First, as regards the gap measure I use deviations from a slowly adjusting HP-trend ($\lambda = 1600$), which

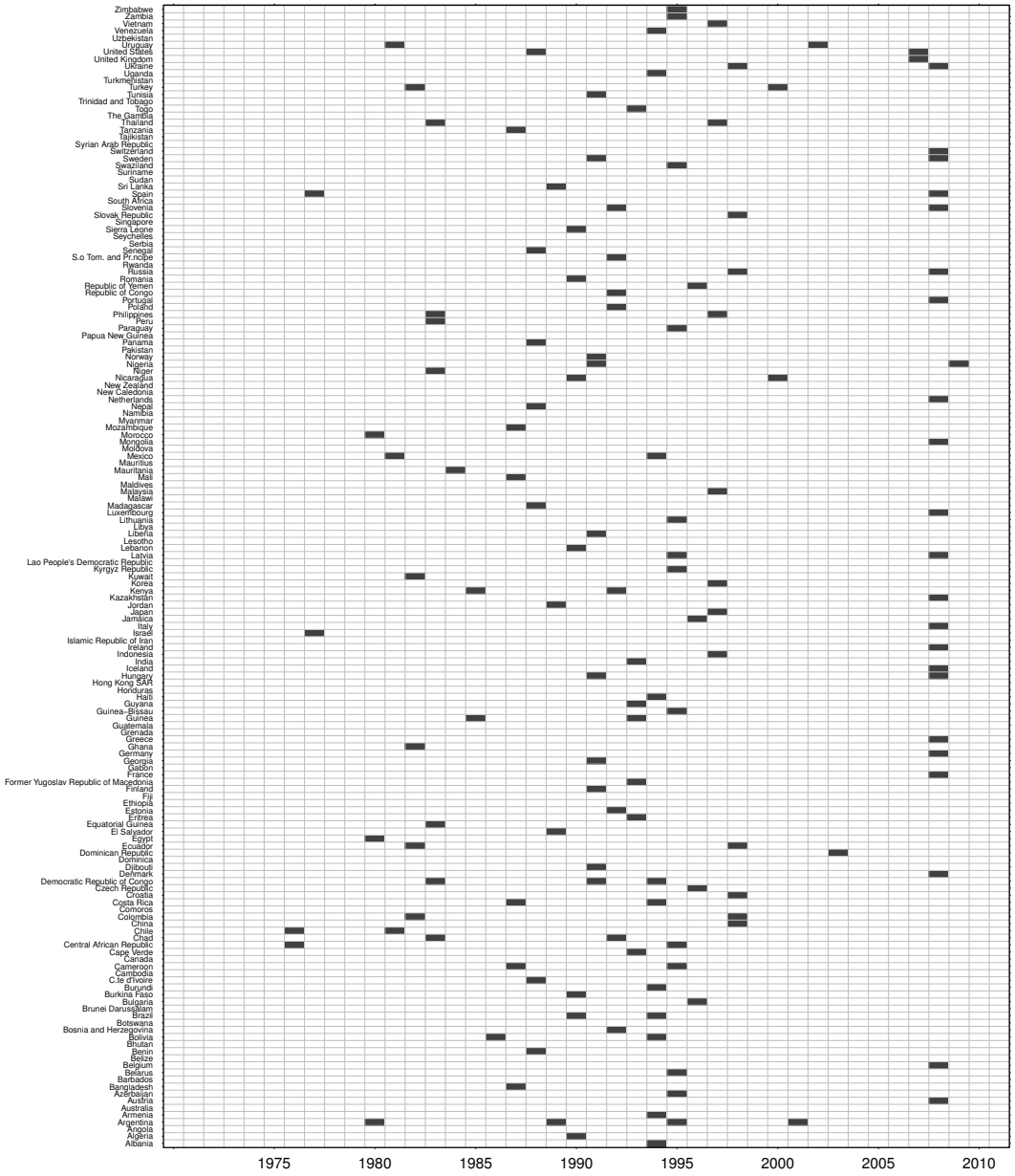

Figure 3: Crisis Map: Post-1970 Samples

Notes: Systemic banking crises (black). Source: Systemic banking crisis dummies from Laeven and Valencia (2013).

captures the slow build-up of financial imbalances (see Borio and Drehmann, 2009). As the following is an excercise in forecasting I use an one-sided HP-filter (Mehra, 2004).
Second, I standardize the "to GDP" ratios when there exists a lack of cross-country comparability between series (e.g. the loan aggregates (see Schularick and Taylor, 2012)). Therefore these predictors contain only information on the relative predictor-level within each country.

Third, I calculate the global averages at time t as the GDP-weighted average of all countries with nonmissing values at time t. For the quarterly post-1970 sample I calculate the non-weighted global average, due to the limited availability of quarterly PPP GDP data.

Fourth, I calculate the real exchange rate as: $RER_{i,t} = NER_{i,t} \cdot P_{i,t} \cdot P_t^{*-1}$, where $NER_{i,t}$ denotes the nominal USD-exchange rate in price notation of country i at time t, $P_{i,t}$ is the domestic CPI and P_t^* is the GDP-weighted average CPI of all countries with nonmissing values at time t.

Fifth, I obtain the interest rate differential-predictors through subtracting global average interest rates from country interest rates.

Sixth, specifically for the logit analysis I generate six interaction terms. I don't use these interaction terms in the classification tree-based analysis, as classification trees automatically identify important predictor interactions. I proxy private debt servicing costs by the interaction of the Loans/GDP (gap) with long-term interest rates. In the same way I obtain public debt servicing costs as the interaction of the Public Debt/GDP (gap) with long-term interest rates (see Drehmann and Juselius, 2012; Jordà, 2013). I define

Table 1: Datasets

Dataset	Long-run 1870-2011 sample	Broad post-1970 sample I	Broad post-1970 sample II
Source	Schularick and Taylor (2012)	IMF IFS, PWT, Abbas et al. (2013)	IMF IFS
Crisis dummy	Schularick and Taylor (2012)	Laeven and Valencia (2013)	Laeven and Valencia (2013)
Frequency	annual	annual	quarterly
Time Span	1870 - 2011	1970 - 2011	1970 - 2011
# of countries	17	162	162
# of predictors	77 (Table A1)	70 (Table A2)	73 (Table A3)
# of crises	93	147	147
N	2414	7081	30967

Notes: *N* number observations. IFS International Financial Statistics. PWT Penn World Tables. The Schularick and Taylor (2012) dataset has subsequently been extended and updated (see Jordà, Schularick, and Taylor, 2013). All three datasets are unbalanced. Thus the number of observations and crises will vary across applications.

the joint debt burden as the interaction of long-term rates with the Loans/GDP (gap) and the Public Debt/GDP (gap). The remaining three interaction terms are: Loans/GDP x GDP (gr) and Public Debt/GDP x GDP (gr), aimed at capturing debt sustainability consideration in the face of low GDP realizations, and Loans/GDP (gap) x Exchange Rate (gap), aimed at capturing effects from exchange rate devaluations on the banking system.

Furthermore, I only make use of predictors which are broadly available across countries and time, because the limited number of crisis observations does not allow for large data losses due to missing values. For each of the three samples I eventually opt for an eclectic, though to some extent ad hoc, selection of somewhat more than 70 predictors with an eye on erring on the side of inclusion. Readers with a background in nonparametric statistics may wonder why I don't tune the precise number of predictors according to optimal

out-of-sample performance. I don't endogenize the number of predictors in such a way for the same reason I don't tune the stopping rule: Given the rarity of severe banking crises I deem setting aside part of the data for tuning considerations to be an excessive strain on the samples.

An exhaustive list of all predictors can be found in tables A1, A2 and A3 in the appendix. These can be compared to table A4, which provides a summary of variable selections in empirical publications on banking crises.

An overview of the characteristics of the three datasets is given in table 1.

5 Performance Comparison

I now stage the competition between logit models, single classification trees and $\mathcal{CT}$-ensembles. The rules of the competition are simple: The method whose crisis-probability predictions achieve the highest out-of-sample *area under the receiver operating characteristic curve* (AUC) wins. The AUC ranges from 0.5 to 1. An AUC of 1 indicates the perfect EWS, which correctly forecasts all crises as crises and all non-crises as non-crises. An AUC of 0.5 indicates an entirely uninformative EWS (For a comprehensive introduction to the AUC measure see Jordà, 2013).

First, I report baseline results based on the long-run sample for logit EWS (see subsections 5.1) and classification tree-based EWS (see subsection 5.2). In subsection 5.3 I go into more detail in directly comparing the best logit and $\mathcal{CT}$-ensemble EWS on the basis of their *receiver operating characteristic* (ROC) curves. I then make use of the two post-1970 samples to test the robustness of my results in subsection 5.4. In subsection 5.5 I check

whether the forecasting performance of $\mathcal{CT}$-ensembles holds up for 1- and 3-year crisis horizons. Finally, in subsection 5.6 I shortly investigate the performance of tree boosting algorithms, which share much of their properties with $\mathcal{CT}$-ensembles and thus also suggest themselves for crisis forecasting. Nevertheless I will recommend against their use in crisis forecasting.

5.1 Logit EWS

To obtain a yardstick against which to measure the performance of $\mathcal{CT}$-ensembles I first carry out a logistic regression-based analysis. I estimate bi- and multivariate logit models based on a selection of predictors which is comparable to those found in the literature. Among the single predictors the usual suspects make their appearance. The largest AUCs come from the private burden (AUC=0.64) and the loans/GDP gap (AUC=0.63). They are statistically significantly different from 0.5 at the 99% level. The public debt/GDP gap (AUC=0.59) and the public burden (AUC=0.58) achieve significance at lower levels. Most of the other AUC estimates hover closely above 0.5; a rather poor result. Also, notice the multiple testing problem involved. All ten single predictors loose their significance level when I conduct the appropriate *Bonferroni adjustment* ($\frac{\alpha}{10}$). Generally, these results are similar to the ones obtained by Jordà (2013) who, based on comparable specifications, reports AUCs ranging from 0.52 to 0.67.[6]

[6]Drehmann and Juselius (2013) report mean AUC estimates between 0.8 and 0.9 for their loans/GDP gap and their debt servicing ratio. These estimates are substantially higher than the ones obtained here and come close to the soon to be introduced $\mathcal{CT}$-ensembles. The most important factor behind this discrepancy is their rather homogenous post-1980 sample. Their sample contains only 19 systemic crises, 11 of which are associated with the most recent global financial crisis. This compares to the more than 70 systemic crises under analysis here.

Next are mulitvariate specifications. The variable selections are displayed on the right half of table 2. They are inspired by similar specifications in Schularick and Taylor (2012) and Jordà, Schularick, and Taylor (2011). AUCs of all three multivariate models are significantly different from 0.5 at the 99% level. They range from 0.62 to 0.65. Compared to the ”baseline” specification the ”IA” specification with interaction terms is successful in conveying extra information on the imminence of a banking crisis (AUC=0.65). The AUC remains the same after the additional inclusion of country-fixed effects. These results are similar to the out-of-sample results reported by Schularick and Taylor (2012) (AUC=0.646). Note however, that even 0.65 does not lie outside of the 95% confidence interval of the best single predictors.

Table 2: Logit-EWS

	Results			Specification		
	AUC	95%-CI	N	Baseline	IA	FE & IA
Bivariate				**Variables**		
Loans/GDP (gap)	0.63 * *	[0.55,0.71]	1295	✓	✓	✓
Public Debt/GDP (gap)	0.59 *	[0.51,0.66]	1348	✓	✓	✓
Narrow Money/GDP (gap)	0.55	[0.47,0.63]	1310	✓	✓	✓
LT Interest Rate	0.52	[0.44,0.59]	1437	✓	✓	✓
GDP (gr)	0.52	[0.44,0.6]	1414	✓	✓	✓
Inflation	0.54	[0.46,0.61]	1503	✓	✓	✓
Exchange Rate (gap)	0.51	[0.44,0.59]	1503	✓	✓	✓
Public Burden	0.58 †	[0.5,0.65]	1322			
Private Burden	0.64 * *	[0.56,0.72]	1237			
Joint Burden	0.53	[0.44,0.61]	1182			
Multivariate				**Interaction Terms**		
Baseline	0.62 * *	[0.55,0.7]	1146		✓	✓
IA	0.65 * *	[0.57,0.73]	1146		✓	✓
FE & IA	0.65 * *	[0.57,0.73]	1146		✓	✓
					✓	✓
					✓	✓
					✓	✓
					✓	✓
				Fixed Effects		
				Country-FE		✓

Notes: Dependent variable: two-year horizon before crisis. Out-of-sample mean AUC- and confidence band estimates are based on Monte Carlo Cross-Validation (see Picard and Cook, 1984; Arlot, Celisse et al., 2010): 5000 MC-draws of training (63,2%) - test (36,8%) data partitions. IA - interaction terms; FE - country fixed effects; N - number of training observations ($= 0.632 \times$ total number of observations). † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

5.2 Classification Tree-based EWS

$\mathcal{CT}$-ensembles are not just an ensemble of trees but also an ensemble of techniques. To obtain an impression of the relative efficacy of bagging, randomization and the inclusion of many predictors the following analysis will build up to the final $\mathcal{RF}$ model one step at a time. First a single classification tree based on the same restricted selection of ten predictors as the IA logit model will take the stage, before bagging and randomization will be added to the recipe. After that, the same three steps, single tree - bagging - randomization, will be followed through on the basis of all 76 predictors.

5.2.1 Single Tree

The left hand side of table 3 displays results for the restricted predictor selection. A single tree performs badly (AUC = 0.55). A similar finding has led Jagtiani et al. (2003) to suggest, prematurely, that "simple is beautiful" when it comes to EWS. They find that a logit EWS performs better than a more complex nonparametric approach: trait recognition analysis (TRA). Compared to a logit model TRA buys an increase in functional flexibility at the cost of higher model variability - similar to a classification tree. In both cases the net effect on predictive performance is unfavourable. Concluding from this however that the increase in functional flexibility is not worth having is premature. In general I argue that "simple" is not adequate for forecasting banking crises. To accommodate complex predictor interactions and various nonlinearities, functional flexibility is well worth having, while the associated problem of model instability can be resolved through

bagging.[7]

5.2.2 Bagging

The second row in the upper left quadrant of table 3 displays the effect of bagging in the ten-variable setting. The AUC leaps by more than 0.2 to a value of 0.77. This AUC is significantly higher than that displayed by any of the logit models. Bagging thus constitutes a large step forward towards a good forecasting performance. Furthermore, the confidence interval indicates that the AUC estimate is rather precise, especially if compared to the multivariate logit specifications.

5.2.3 Random Forest

The third model in the upper left quadrant of table 3 is the $\mathcal{RF}$-estimator. The additional randomization in form of randomly analyzing only three out of the ten predictors at each recursive partitioning step, leads to a slightly higher mean AUC estimate of 0.79.
$\mathcal{CT}$-ensembles have thus already left behind their logistic competitors without yet having capitalized on their second fundamental advantage – their ability to make forecasts on the basis of many predictors.

5.2.4 Many Predictors

I now turn to the more predictor-intensive contenders. The results are displayed in the upper right quadrant of table 3. The extension of the list of predictors to a total of 76 results

[7] Liu, Bowyer, and Hall (2004) present a set of conditions which the crisis forecasting task fulfills, and under which Artificial Neural Networks are outperformed by classification trees, although the structure of the latter is arguably less flexible. The right sort and degree of functional flexibility of course depends on the nature of the problem. Also see Peltonen (2006) for evidence that Neural Networks perform unfavourably in out-of-sample forecasts of emerging market currency crises.

Table 3: $\mathcal{CT}$-EWS: Long-run 1870-2012 Sample

	Results					
	Restricted Selection			**Many Predictors**		
Model	AUC	95%-CI	N	AUC	95%-CI	N
Single Tree	0.55	[0.5,0.6]	1816	0.63 * *	[0.57,0.7]	1742
Bagging	0.77 * *	[0.73,0.81]	1816	0.87 * *	[0.84,0.9]	1742
Random Forest	0.79 * *	[0.75,0.83]	1816	0.88 * *	[0.85,0.91]	1742

	Specification					
	Restricted Selection			**Many Predictors**		
Parameter	Single	Bagging	RF	Single	Bagging	RF
B	1	5000	5000	1	5000	5000
J_{try}	10	10	3	76	76	9
J		10			76	
# of crises		72			70	

Notes: Dependent variable: two-year horizon before crisis. Restricted Selection: Loans/GDP (gap), Public Debt/GDP (gap), Narrow Money/GDP (gap), LT Interest Rate, GDP (gr), Inflation, Exchange Rate (gap), Loans/GDP, Public Debt/GDP, LT Interest Rate (n). Many Predictors: see table A1. Out-of-sample AUC-estimates (and confidence intervals) based on out-of-bag (OOB)-data. N number observations. J number of predictors under analysis. J_{try} number of predictors randomly selected and considered as a splitting variable at each recursive partitioning step. B number of trees. Specification table: If only the bagging column has an entry this means all models share the same specification. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

in a second leap in forecasting performance, by about 0.1 for the $\mathcal{F}$- (AUC=0.87) and $\mathcal{RF}$ (AUC=0.88) estimators. Even the single classification tree (AUC=0.63) now performs similarly to the multivariate logit EWS. Overall a kitchen sink approach significantly beats one based on gnostic variable selection. This is not to say that any amount of predictors should be considered, regardless of any theoretical plausibility as at some point increases in model variability will doubtlessly outweight any benefits coming from further bias

reduction. But given the rather broad bounds set by theory, large datasets and a method capable of dealing with sparce predictor spaces, these results suggest that extending the number of predictors is a worthwhile excercise.

To sum up, bagged and randomized classification trees win the banking crisis EWS contest for the restricted as well as unrestricted set of predictors. Substantial increases in forecasting performance result from bagging and the inclusion of many predictors.

5.3 ROC-Comparison of Logit- and $\mathcal{RF}$ EWS

I will now directly compare the IA logit- and the $\mathcal{RF}$ EWS in some more detail. The AUC is an aggregate measure of predictive performance derived from the more informative *receiver operating characteristic* (ROC) curve. The ROC curve is a graphical representation of all true positive rate - false positive rate (TPR-FPR) combinations a model is capable of (see Schularick and Taylor (2012) and Jordà (2013) for introductions to the ROC curve for evaluating banking crisis EWS). Thus despite the significantly higher AUC of the $\mathcal{RF}$ EWS compared to the IA logit EWS it might be the case that the logit EWS still has some TPR-FPR trade-off on offer which cannot be replicated by the $\mathcal{RF}$ EWS. Thus, depending on how much weight the policy maker's preferences put on making correct crisis calls as opposed to correct no crisis calls, the logit EWS might be preferable after all. To see whether this is the case figure 4 displays the ROC curves of both EWS. As can be seen, the $\mathcal{RF}$ ROC-curve lies north-west of the logit ROC-curve throughout. Thus, regardless of policy-makers' preferences, the $\mathcal{RF}$ EWS has the better TPR-FPR trade-off on offer.

To get a better intuition for the performance of the $\mathcal{RF}$ EWS let's look at some

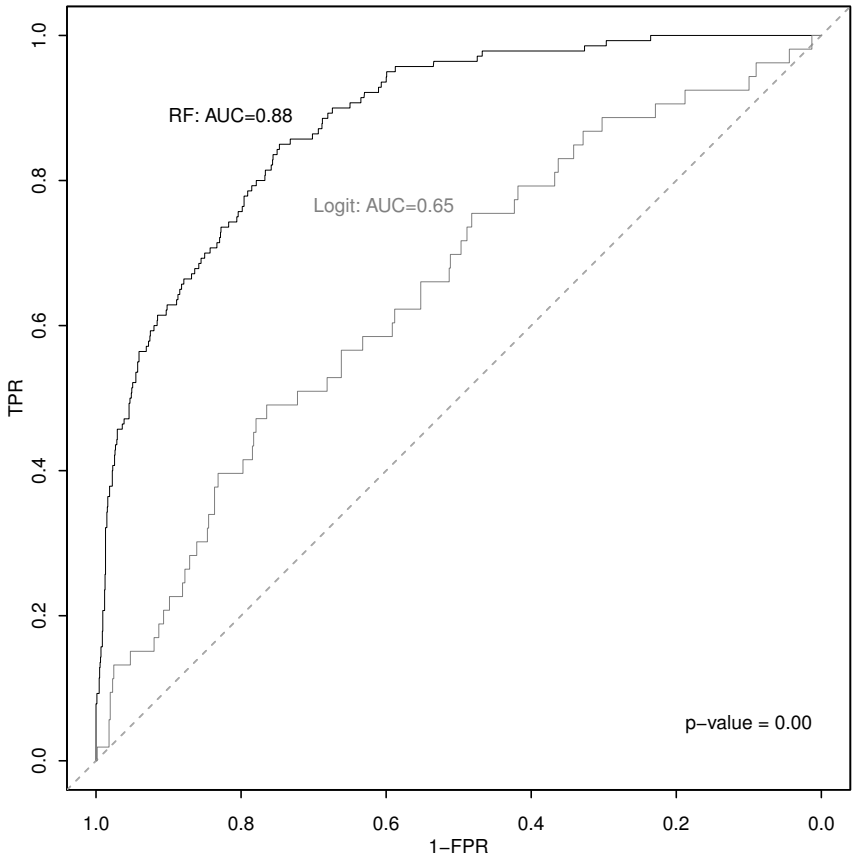

Figure 4: ROC-Comparison

Notes: Receiver Operating Characteristic curves of the IA logit (grey) and $\mathcal{RF}$ model (black). P-value for test of equality of AUCs according to (DeLong, DeLong, and Clarke-Pearson, 1988). TPR true positive rate. FPR false positive rate.

exemplary TPR-FPR combinations. The $\mathcal{RF}$ EWS offers a balanced TPR-FPR trade-off at about $TPR = FPR = 0.75$, i.e. it enables policy makers to correctly forecast 75% of crises and 75% of no crises. If a 25% probability of mistakenly forecasting a crisis is deemed too high by policy makers the $\mathcal{RF}$ EWS allows for a reduction of the probabilty of mistakenly forecasting a crisis to 5%, while still correctly forecasting about 50% of banking crises. At the other extreme, policy makers eager not to miss any crisis could use the $\mathcal{RF}$ estimator to correctly forecast 95% of crises, while still getting about 50% of the no crisis forecasts

right. Any of these trade-offs leaves policy makers substantially better off than the logit EWS would.

5.4 Post-1970 Samples

I now turn to the two post-1970 samples. Up to now, all results were based on only one dataset. Replicability on others would be reassuring. I thus repeat the analysis on the basis of the annual and quarterly post-1970 samples. Except for a slightly different set of predictors the analysis is identical to the one I conducted for the long-run sample. The results I obtain take the same line as before: Bagging and the inclusion of many predictors substantially improve forecasting performance. I now discuss the results for the annual and quarterly post-1970 samples in turn.

The results for the annual post-1970 sample are reported in table 4. As was the case for the long-run sample, a single classification tree estimated on the basis of the restricted set of seven predictors performs poorly (AUC = 0.52). This poor performance is again remedied through bagging: Compared to the single classification tree the mean AUC estimate for the $\mathcal{F}$ estimator jumps by more than 0.25 to 0.79. The additional randomization in the $\mathcal{RF}$ estimator (AUC=0.81) is again associated with a 0.02 increase in the mean AUC estimate.

Estimation on the basis of 70 predictors results in a second jump in forecasting performance for the two $\mathcal{CT}$-ensembles. For both I obtain a mean AUC estimate of 0.85. The single classification tree also benefits from the inclusion of many predictors; its mean AUC estimate increases from an insignificant 0.52 to a significant 0.56.

Overall, the results for the annual long-run- and annual post-1970 sample are thus

Table 4: $\mathcal{CT}$-EWS: Annual Post-1970 Sample

	Results					
	Restricted Selection			**Many Predictors**		
Model	AUC	95%-CI	N	AUC	95%-CI	N
Single Tree	0.52	[0.49,0.55]	4274	0.56 * *	[0.52,0.6]	4189
Bagging	0.79 * *	[0.76,0.82]	4274	0.85 * *	[0.83,0.87]	4189
Random Forest	0.81 * *	[0.78,0.84]	4274	0.85 * *	[0.83,0.88]	4189

	Specification					
	Restricted Selection			**Many Predictors**		
Parameter	Single	Bagging	RF	Single	Bagging	RF
B	1	5000	5000	1	5000	5000
J_{try}	7	7	3	70	70	8
J		7			70	
# of crises		100			100	

Notes: Dependent variable: two-year horizon before crisis. Restricted Selection: Loans/GDP (gap), Public Debt/GDP (gap), Inflation, Real Exchange Rate (gap), Loans/GDP, Public Debt/GDP, Net Exports/GDP (gap). Many Predictors: see table A2. Out-of-sample *AUC*-estimates (and confidence intervals) based on out-of-bag (OOB)-data. N number observations. J number of predictors under analysis. J_{try} number of predictors randomly selected and considered as a splitting variable at each recursive partitioning step. B number of trees. Specification table: If only the bagging column has an entry this means all models share the same specification. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

very similar. The mean AUC estimates on one sample are usually contained within the 95% confidence band estimates on the other sample. Such a degree of replicability is reassuring.

I now turn to the quarterly post-1970 sample. The results are reported in table 5. The mean AUC estimates are generally higher regardless of specification. However, the mean AUC estimate for the single classification tree based on the restricted set of predictors

Table 5: $\mathcal{CT}$-EWS: Quarterly Post-1970 Sample

	Results					
	Restricted Selection			**Many Predictors**		
Model	AUC	95%-CI	N	AUC	95%-CI	N
Single Tree	0.58 * *	[0.55,0.6]	19126	0.7 * *	[0.67,0.73]	19061
Bagging	0.85 * *	[0.83,0.86]	19126	0.97 * *	[0.97,0.98]	19061
Random Forest	0.85 * *	[0.84,0.86]	19126	0.95 * *	[0.95,0.96]	19061

	Specification					
	Restricted Selection			**Many Predictors**		
Parameter	Single	Bagging	RF	Single	Bagging	RF
B	1	5000	5000	1	5000	5000
J_{try}	9	9	3	73	73	9
J		9			73	
# of crises		102			102	

Notes: Dependent variable: two-year horizon before crisis. Restricted Selection: Loans (gap), Loans (gr), Foreign Liabilities (gap)(glo), LT Interest Rate (gap)(glo), GDP (gap)(glo), Inflation, Exchange Rate (gap), Reserves (gap), GDP (gr)(glo). Many Predictors: see table A3. Out-of-sample AUC-estimates (and confidence intervals) based on out-of-bag (OOB)-data. N number observations. J number of predictors under analysis. J_{try} number of predictors randomly selected and considered as a splitting variable at each recursive partitioning step. B number of trees. Specification table: If only the bagging column has an entry this means all models share the same specification. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

(AUC=0.58) still is rather close to the uninformative 0.5. Once again it is their aggregation into an ensemble which renders classification trees fit for forecasting; the mean AUC estimate for the two $\mathcal{CT}$-ensembles based on the restricted set of predictors is 0.83. As opposed to the other two samples the $\mathcal{RF}$ estimator does not improve upon the $\mathcal{F}$ estimator.

Estimation on the basis of the eclectic set of 73 predictors again improves forecasts.

Even the single classification tree now achieves an AUC of 0.7. For the two $\mathcal{CT}-ensembles$ I obtain high AUC estimates of 0.97 and 0.95. This time the additional randomization puts the $\mathcal{RF}$ estimator at a small disadvantage against the plain $\mathcal{F}$ estimator.

These results suggest that EWS based on quarterly data will do a better job than comparable EWS based on annual data. More generally the strong performance of $\mathcal{CT}$-ensembles now extends to at least three datasets. Thus, their success appears to be solidly grounded in their suitability for banking crisis forecasting.

5.5 Different Crisis Horizons

Up to now I have trained all models in correctly identifying the 2-year horizon before a banking crisis event. If however particularly next year's crisis risk is of concern, then a 1-year crisis horizon better serves the purpose. In other cases a 3-year crisis horizon might be preferred. I thus obtain estimates of the $\mathcal{RF}$ EWS for the 1- and 3-year crisis horizons.

Table 6 shows the results. The 1- and 3-year horizon AUC estimates for the long-run sample are displayed in the first two columns. For the 2-year horizon case the mean AUC estimate was 0.88 (see table 3). The mean AUC estimate for the 1-year horizon (AUC=0.78) is substantially lower than that. The 3-year horizon mean AUC estimate (AUC=0.88) on the other hand is identical to the 2-year horizon one. Thus it is harder to assess whether there will be a crisis next year, than to assess whether there will be a crisis within the next two or three years.

The results for the post-1970 sample suggest the same conclusion. As regards the annual post-1970 sample the 1-year horizon AUC estimate (AUC=0.75) again falls by 0.06

compared to the 2-year horizon estimate (AUC=0.81)(see table 4). The 3-year horizon AUC estimate of 0.88 on the other hand exceeds the corresponding 2-year horizon estimate by 0.07.

Finally, for the quarterly post-1970 sample the 1-year horizon AUC estimate (AUC=0.93) is somewhat lower than the 2-year one (AUC=0.95)(see table 5). The 3-year horizon AUC estimate is 0.96 – close to the 2-year horizon one.

To sum up, while the AUC estimates for the 3-year horizon generally are very close to those for the 2-year horizon, the AUC estimates for the 1-year horizon are substantially lower. In general it is thus harder to assess whether there will be a crisis next year, than to assess whether there will be a crisis within the next few years. On a more general note, the fact that the data allow the $\mathcal{RF}$ algorithm to better discern the 3-year crisis horizon from all other observations than the 1-year horizon conforms to accounts which picture banking crisis risks as building up slowly over time, while the actual crisis realization is less determinate – triggered by a shock which may or may not occur in any particular year.

Table 6: Different Crisis Horizons

	1-year horizon Long-run sample yearly	3-year horizon Long-run sample yearly	1-year horizon Post-1970 sample yearly	3-year horizon Post-1970 sample yearly	1-year horizon Post-1970 sample quarterly	3-year horizon Post-1970 sample quarterly
AUC	0.78 * *	0.88 * *	0.75 * *	0.88 * *	0.93 * *	0.96 * *
95%-CI	[0.72,0.84]	[0.86,0.91]	[0.7,0.79]	[0.86,0.9]	[0.92,0.94]	[0.95,0.96]
N	1742	1742	4189	4189	19061	19061
B	5000	5000	5000	5000	5000	5000
J_{try}	9	9	8	8	9	9
J	76	76	70	70	73	73
# of crises	70	70	100	100	102	102

Notes: Estimator: Random forest. Predictors: see tables A1, A2 and A3. Out-of-sample *AUC*-estimates (and confidence intervals) based on out-of-bag (OOB)-data. N number observations. J number of predictors under analysis. J_{try} number of predictors randomly selected and considered as a splitting variable at each recursive partitioning step. B number of trees. Specification table: If only the bagging column has an entry this means all models share the same specification. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

5.6 Boosting

Before turning to the 2007/2008 case study I take another look at a loose relative of the $\mathcal{CT}$-ensemble family: *Boosting*. Boosted classification trees can be described as a forest with directed tree growth: In the estimation of each new tree particular weight is put on the correct classification of those observations which have been misclassified by the aggregate of all previously estimated trees (for a comprehensive introduction to boosting see Hastie, Tibshirani, and Friedman, 2013, chapter 10). In principle boosting features many of the properties which make $\mathcal{CT}$-ensembles fit for banking crisis forecasting. This, and their exceptional track record is reason enough to check up on their crisis forecasting performance here. In particular I make use of a *stochastic gradient boosting machine* based on many predictors. As regards the parameter specification I follow standard recommendations from the literature (see Friedman, 2002; Buehlmann, 2006).

Columns one to three in table 7 display the results for the three samples. The AUC estimates range from 0.75 to 0.84. This places boosting somewhere between a single classification tree and bagging. Why does boosting stay behind the $\mathcal{F}$- and $\mathcal{RF}$ EWS? A likely explanation is the vulnerability of boosting algorithms to noisy data (Long and Servedio, 2010). In putting extra weight on the correct classification of those observations which are hard to classify boosting often ends up giving undue weight to datapoints which are merely noisy. To the extent that macroeconomic data, and in particular the crisis dummies are noisy this should be expected to constrain the performance of boosting algorithms. When it comes to forecasting banking crises the $\mathcal{F}$- and $\mathcal{F}$ EWS are preferable.

Table 7: Tree Boosting

	Long-run sample yearly	Post-1970 sample yearly	Post-1970 sample quarterly
AUC	0.78 * *	0.75 * *	0.84 * *
95%-CI	[0.71,0.85]	[0.7,0.8]	[0.82,0.86]
N	1724	4189	19061
B	5000	5000	5000
J	77	70	73
η	0.5	0.5	0.5
ν	0.1	0.1	0.1
# of crises	70	100	102

Notes: Dependent variable: two-year horizon before crisis. Predictors: see tables A1, A2 and A3. Out-of-sample AUC-estimates (and confidence intervals) based on Monte Carlo Cross-Validation (see Picard and Cook, 1984; Arlot, Celisse et al., 2010) for Boosting: 5000 MC-draws of training (63,2%) - test (36,8%) data partitions. N number observations. J number of predictors under analysis. B number of trees. η random fraction of observations in training data I use to estimate each tree. ν shrinkage parameter indicating the weight given to each new tree. † $p < 0.10$, * $p < 0.05$, ** $p < 0.01$

6 Case Study: 2007/2008

To conclude the preceeding analysis I contrast the performance of the $\mathcal{RF}$ EWS based on many predictors and the IA-logit EWS in forecasting the the 2007/2008 global financial crisis. I estimate both EWS on the basis of the long-run sample. I use data up to 1997 for estimation and put aside the rest as test data. The resulting crisis probability estimates for the test data (1998 to 2011) are reported in figure 5.[8]

[8]By selecting a probability threshold the reported crisis probability estimates can be translated into correct and false warning signals and thus different TPR-FPR combinations. For example for a threshold of 0.4 the $\mathcal{RF}$ EWS would have correctly forecasted a banking crisis for Spain within the next two years in 2006 and 2007. Policy makers less sensitive to making wrong crisis calls could chose a lower threshold, while those more sensitive could chose a higher one. For expository reasons however the discussion of the results will develop along less formal lines – comparing trends and levels in probability estimates of the two EWS.

On first sight it is clear that the $\mathcal{RF}$ crisis risk evaluation exhibits considerably more variation than the logit model: For most countries it would have signalled a build-up in crisis probability in the late 1990s, and again in the mid-2000s. Thus the $\mathcal{RF}$ model would have signalled rather clearly that the developed world as a whole was embarking upon a path which historically has often ended in crisis. The evidence for the logit model is less flattering: while for some countries it signals (slightly) higher crisis risk, for others it signals no big changes, or even increasing resilience over the 2000s. Thus, prior to the global financial crisis policy makers would have been better served by the $\mathcal{RF}$-EWS.

With respect to the country-specific incidence of the crisis, the record of the $\mathcal{RF}$ model is more mixed. Let's take a closer look: For some countries crisis risk went up and a crisis did indeed occur: Belgium, Switzerland, Denmark, Spain, France, UK, Italy, Netherlands, Portugal, Sweden and the USA. Though for all of these countries $\mathcal{RF}$ crisis risk is upward trending, its level is relatively low for some. Switzerland and the USA belong into this category. Germany, for which crisis risk does not even trend upwards, also exhibits a very low risk level. How can these cases be explained? What brought down German and Swiss banks was their exposure to foreign assets. As regards the USA, nonbank intermediation was at the heart of its banking crisis. Neither exposure to foreign assets, nor nonbank intermediation are well reflected by any of the base indicators in the long-run sample. Extending the list of base indicators may thus help to improve forecasts. Unfortunately detailed data on international financial linkages for example simply have not been collected for that many years yet. This is one reason why current formal EWS should only be considered part of a more comprehensive risk assessment.

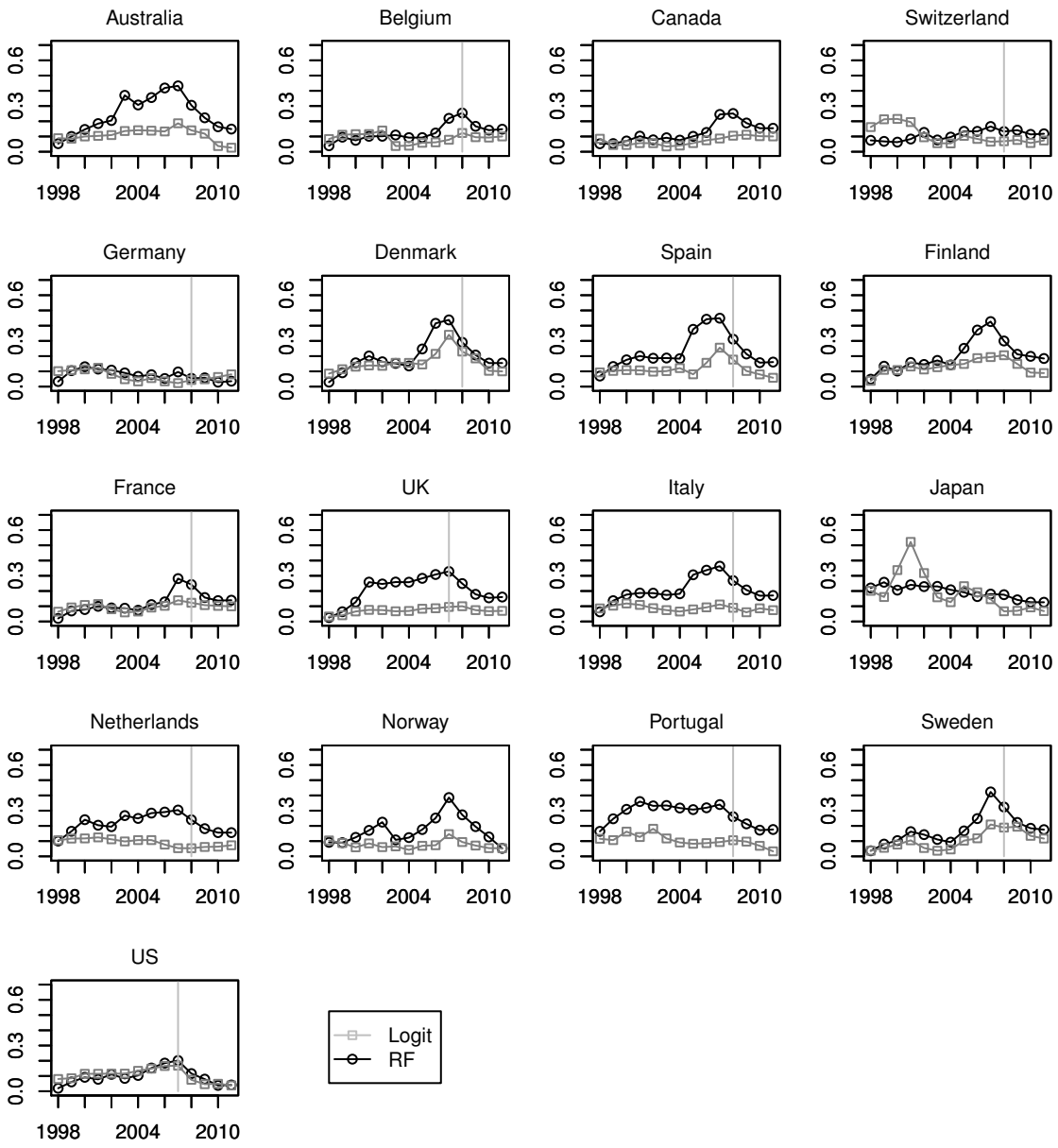

Figure 5: The 2007/2008 Global Financial Crisis

Notes: 1998-2011 out-of-sample probability estimates of being in the two year horizon before a banking crisis for 17 countries. 10 country-year observations exhibit missing values, which were replaced by the respective variable's mean to obtain a probability estimate. Vertical gray bars indicate year of systemic banking crisis.

Several countries show clear signs of being in a danger zone prior to 2007/2008, but did not experience a systemic banking crisis according to the binary indicator: Australia,

Canada, Finland, Norway. There is a notable concentration of Scandinavian countries and primary good exporters in this group. Hardy and Pazarbasioglu (1998) show that primary-product exporting countries possess a dinstinct set of early warning indicators, which might explain the bad performance of the $\mathcal{RF}$-EWS in this case. What is also interesting is that, except for Canada, all of these countries had experienced a banking crisis in the late 1980s or early 1990s. The ensuing institutional changes might have rendered their banking systems more resilient 20 years later. Giannone, Lenza, and Reichlin (2010) present related evidence for the importance of regulatory quality in credit markets in explaining cross-country differences in weathering the global recession. Also note, that in Australia and Norway banking systems did in fact come under considerable stress during the relevant period – they are knife-edge cases with respect to the dummy categorization applied.

The last group consists of countries which did not see their risk profiles rise and indeed did not experience a systemic event: Japan is the only country in this category.

In sum, the record of the $\mathcal{RF}$ model on the most recent crisis is mixed. While the model would not have performed as convincingly with respect to the country-specific incidence of the crisis, it would have clearly signalled that the developed world as a whole was on a dangerous path from the early 2000s on. The first part of this conclusion rings nicely with results reported by Claessens et al. (2010), Rose and Spiegel (2010*a*), Rose and Spiegel (2010*b*) and Rose and Spiegel (2012) who find that prior to the global financial crisis hardly any predictor conveyed reliable information about the crisis' subsequent cross-country severity. While Rose and Spiegel (2012) go on arguing that their results

warrant skepticism towards the potential of EWS to do a good job my analysis suggests a different conclusion. Given the historical track record, the proposed $\mathcal{CT}$-ensemble EWS are very promising. Also note that the evaluation of the $\mathcal{RF}$ EWS's performance in 2007/2008 depends on the categorization of two knife-edge cases. Given a more lenient evaluation of these cases (Australia and Norway) figure 5 shows that even in terms of cross-country incidence for 2007/2008 the $\mathcal{RF}$ predictor did not perform too badly. Especially if combined with country-specific knowledge, as exemplified above, the proposed $\mathcal{RF}$-EWS would have supplied policy makers with a valuable heads-up on the vulnerability of the world financial system prior to the event.

7 Conclusion

This article has explored the potential of classification tree ensembles for forecasting narrative banking crisis indicators. Their out-of-sample performance surpasses current best-practice early warning systems based on logit models by a substantial margin. I obtain this result on the basis of one long-run- (1870-2011), as well as two broad post-1970 macroeconomic panel datasets.

The good forecasting performance of classification tree ensembles contrasts with the poor performance of single classification trees. Single classification trees are held back by their high variability. Bagging many classification trees into an ensemble counters this variability. This article has shown that such bagging is associated with a substantial improvement in out-of-sample forecasts for banking crises.

Another driver behind the good forecasting performance of classification tree ensem-

bles is their ability to make forecasts on the basis of many predictors; classification tree ensembles are typically estimated through recursive partitioning, which circumvenes the curse of dimensionality. This article has shown that an increase in the number of predictors from around ten to several dozens is associated with another substantial improvement in out-of-sample forecasts of banking crises. Together, bagging and the use of many predictors allow classification tree ensembles to substantially surpass the forecasting performance of current alternatives.

Several further results have been obtained: As regards forecasting horizons, it is in general harder to assess whether there will be a crisis next year, than to assess whether there will be a crisis within the next few years. This result is indicative of banking crisis risk building up slowly over time, while the actual crisis realization is less determinate – triggered by a shock which may or may not occur in any particular year.

This article furthermore suggests that plain classification tree ensembles are the preferable choice over classification tree boosting. To the extent that macroeconomic data, and in particular binary banking crisis indicators are noisy measures, boosting is not suited to banking crisis forecasting, as it ends up attributing too much weight to noisy observations (Long and Servedio, 2010).

Finally, the 2007/2008 case study shows that by the mid-2000s a classification tree ensemble would have clearly signalled that the developed world as a whole was embarking upon a path which historically has often ended in crisis. However, with respect to the country-specific incidence of banking crises in 2007 and 2008, the performance of the classification tree ensemble is more ambiguous. This acts as a reminder that for most

practical purposes, formal early warning systems are best thought of as part of a more comprehensive risk assessment, which can take additional country- and time specific knowledge into account.

Acknowledgements

This article has profited from clarifying comments by several colleagues. In particular I am grateful to Thilo Albers, Daniel Becker, Jörg Breitung, Yao Chen, Alois Kneipp, Keith Küster, Alexander Kriwoluzky, Gernot Müller, Moritz Schularick whose helpful advice has improved this article. All remaining errors are mine.

References

Abbas, SM, N. Belhocine, A. El-Ganainy, and M. Horton. 2013. "A historical public debt database."

Acemoglu, Daron, Simon Johnson, James Robinson, and Yunyong Thaicharoen. 2003. "Institutional causes, macroeconomic symptoms: volatility, crises and growth." *Journal of monetary economics*.

Albert, J., E. Aliu, H. Anderhub, P. Antoranz, A. Armada, M. Asensio, C. Baixeras, JA Barrio, H. Bartko, D. Bastieri, et al. 2008. "Implementation of the Random Forest method for the imaging atmospheric Cherenkov telescope MAGIC." *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*.

Alessi, L., and C. Detken. 2011. "Quasi real time early warning indicators for costly asset price boom/bust cycles: A role for global liquidity." *European Journal of Political Economy*.

Arlot, Sylvain, Alain Celisse, et al. 2010. "A survey of cross-validation procedures for model selection." *Statistics surveys*.

Babus, A., E. Carletti, and F. Allen. 2009. "Financial crises: theory and evidence."

Blanchard, O. 2014. "Where Danger Lurks." *Finance and Development*.

Bordo, M., B. Eichengreen, D. Klingebiel, and M.S. Martinez-Peria. 2001. "Is the crisis problem growing more severe?" *Economic policy*.

Borio, C., and M. Drehmann. 2009. "Assessing the risk of banking crises–revisited." *BIS Quarterly Review*.

Borio, C., and P. Lowe. 2002. "Asset prices, financial and monetary stability: exploring the nexus." *BIS Working Paper No. 114*.

Borio, C., and P. Lowe. 2004. "Securing Sustainable Price Stability: Should Credit Come Back from the Wilderness." *BIS Working Paper No. 157*.

Breiman, L. 2001. "Random forests." *Machine learning*.

Breiman, Leo. 1996. "Bagging predictors." *Machine learning*.

Breiman, Leo. 2002. "Manual on setting up, using, and understanding random forests v3. 1." *Statistics Department University of California Berkeley, CA, USA*.

Breiman, L., J. Friedman, C.J. Stone, and R.A. Olshen. 1984. "Classification and regression trees."

Brenda González-Hermosillo, Ceyla Pazarba&scedil, et al. 1997. "Determinants of banking system fragility: A case study of Mexico." *IMF Staff papers*.

Brüggemann, A., and T. Linne. 1999. "How good are leading indicators for currency and banking crises in Central and Eastern Europe?: An Empirical Test." *Halle Institute for Economic Research - Discussion Papers*.

Buehlmann, Peter. 2006. "Boosting for high-dimensional linear models." *The Annals of Statistics*.

Bühlmann, Peter. 2012. "Bagging, boosting and ensemble methods."

Bühlmann, Peter, and Bin Yu. 2002. "Analyzing bagging." *The Annals of Statistics*.

Caprio, Gerard, and Daniela Klingebiel. 1996. "Bank insolvency: bad luck, bad policy, or bad banking?" Vol. 79, The World Bank WashingtonVol. 79, The World Bank Washington. .

Casu, Barbara, Andrew Clare, and Nashwa Saleh. 2011. "Towards a new model for early warning signals for systemic financial fragility and near crises: an application to OECD countries." University Library of Munich, Germany MPRA Paper 37043University Library of Munich, Germany MPRA Paper 37043. .

Claessens, Stijn, Giovanni Dell'Ariccia, Deniz Igan, and Luc Laeven. 2010. "Cross-country experiences and policy implications from the global financial crisis." *Economic Policy*.

Davis, E. Philip, and Dilruba Karim. 2008*a*. "Comparing early warning systems for banking crises." *Journal of Financial Stability*.

Davis, E. Philip, and Dilruba Karim. 2008*b*. "Could Early Warning Systems Have Helped To Predict the Sub-Prime Crisis?" *National Institute Economic Review*.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach." *Biometrics*.

Demirguc, Asli, and Enrica Detragiache. 2000. "Monitoring Banking Sector Fragility: A Multivariate Logit Approach." *World Bank Economic Review*.

Demirgüç-Kunt, A., and E. Detragiache. 1998. "The determinants of banking crises in developing and developed countries." *Staff Papers-International Monetary Fund*.

Demirguc-Kunt, Asli, and Enrica Detragiache. 2005. "Cross-country empirical studies of systemic bank distress : a survey."

Detragiache, Enrica, and Asli Demirgüc-Kunt. 1998. "Financial Liberalization and Financial Fragility (EPub)."

Díaz-Uriarte, R., and S.A. De Andres. 2006. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics*.

Dietterich, Thomas G. 2000. "Ensemble Methods in Machine Learning."

Drehmann, M. 2013. "Total Credit as an Early Warning Indicator for Systemic Banking Crises." *BIS Quarterly Review*.

Drehmann, Mathias, and Mikael Juselius. 2012. "Do debt service costs affect macroeconomic and financial stability?" *BIS Quarterly Review September*.

Drehmann, Mathias, and Mikael Juselius. 2013. "Evaluating Early Warning Indicators of Banking Crises: Satisfying Policy Requirements." *BIS Working Paper No. 421*.

Duttagupta, R., and P. Cashin. 2011. "Anatomy of Banking Crises in Developing and Emerging Market Countries." *Journal of International Money and Finance*.

Eichengreen, Barry, and Andrew K Rose. 1998. "Staying afloat when the wind shifts: External factors and emerging-market banking crises."

Eichengreen, Barry, and Carlos Arteta. 2002. "Banking crises in emerging markets: presumptions and evidence." *Financial policies in emerging markets*.

Eichengreen, Barry, Andrew K Rose, and Charles Wyplosz. 1994. "Speculative attacks on pegged exchange rates: an empirical exploration with special reference to the European Monetary System."

Eicher, Theo S., Charis Christofides, and Chris Papageorgiou. 2012. "Did Established Early Warning Signals Predict the 2008 Crises?"

Feenstra, Robert C, Robert Inklaar, and Marcel Timmer. 2013. "The next generation of the Penn World Table." National Bureau of Economic ResearchNational Bureau of Economic Research. .

Frankel, Jeffrey, and George Saravelos. 2012. "Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis." *Journal of International Economics*.

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational Statistics & Data Analysis*.

Frydman, H., E.I. Altman, and D.L. Kao. 1985. "Introducing recursive partitioning for financial classification: the case of financial distress." *Journal of Finance*.

Gavin, M., and R. Hausmann. 1996. "The Roots of Banking Crises: The Macroeconomic Context." *Inter-American Development Bank Working Paper 318*.

Giannone, Domenico, Michele Lenza, and Lucrezia Reichlin. 2010. "Market freedom and the global recession." *IMF Economic Review*.

Glick, Reuven, and Michael Hutchison. 2000. "Banking and currency crises: how common are the twins?"

Goldstein, Morris Arthur, Graciela Laura Kaminsky, and Carmen Reinhart. 2000. "Assessing Financial Vulnerability: An Early Warning Signals for Emerging Markets."

González-Hermosillo, Brenda. 1996. "Banking sector fragility and systemic sources of fragility."

Gonzalez-Hermosillo, Brenda. 1999. *Determinants of ex-ante banking system distress: A macro-micro empirical exploration of some recent episodes. . .*

Gourinchas, P., and M. Obstfeld. 2011. "Stories of the 20th Century for the 21st." *American Economic Journal: Macroeconomics, forthcoming*.

Gourinchas, Pierre-Olivier, Rodrigo Valdes, and Oscar Landerretche. 1999. "Lending booms: Some stylized facts."

Hahm, Joon-Ho, Hyun Song Shin, and Kwanho Shin. 2012. "Non-core bank liabilities and financial vulnerability."

Hardy, Daniel CL, and Ceyla Pazarbasioglu. 1998. "Leading Indicators of Banking Crises-Was Asia Different?(EPub)."

Hastie, T., R. Tibshirani, and J. Friedman. 2013. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction."

Hawkins, John, and Marc Klau. 2000. "Measuring potential vulnerabilities in emerging market economies."

Honohan, Patrick. 2000. "Banking system failures in developing and transition countries: diagnosis and prediction." *Economic Notes*.

Hutchison, Michael, and Kathleen McDill. 1999. "Are all banking crises alike? The Japanese experience in international comparison." *Journal of the Japanese and International Economies*.

Hutchison, Michael M. 2002. "European banking distress and EMU: institutional and macroeconomic risks." *The Scandinavian Journal of Economics*.

Hyafil, L., and R. L. Rivest. 1976. "Constructing optimal binary decision trees is NP-Complete." *Information Processing Letters*.

Jagtiani, Julapa, James Kolari, Catharine Lemieux, and Hwan Shin. 2003. "Early warning models for bank supervision: Simpler could be better." *ECONOMIC PERSPECTIVES-FEDERAL RESERVE BANK OF CHICAGO*.

Jevons, W. S. 1878. "Commercial Crises and Sun-Spots." *Nature*.

Jevons, W. S. 1879. "Sun-Spots and Commercial Crises." *Nature*.

Jordà, Ò. 2013. "Assessing the Historical Role of Credit: Business Cycles, Financial Crises and the Legacy of Charles S. Peirce." *Federal Reserve Bank of San Francisco Working Paper Series*.

Jordà, Òscar, Moritz Schularick, and Alan M Taylor. 2011. "Financial crises, credit booms, and external imbalances: 140 years of lessons." *IMF Economic Review*.

Jordà, Òscar, Moritz Schularick, and Alan M. Taylor. 2013. "Sovereigns versus Banks: Credit, Crises, and Consequences."

Kaminsky, G.L. 2006. "Currency crises: Are they all the same?" *Journal of International Money and Finance*.

Kaminsky, Graciela L., and Carmen M. Reinhart. 1999. "The Twin Crises: The Causes of Banking and Balance-Of-Payments Problems." *The American Economic Review*.

Kaminsky, Graciela Laura. 1998. "Currency and banking crises: the early warnings of distress."

Laeven, L. 2011. "Banking Crises: A Review." *Annu. Rev. Financ. Econ*.

Laeven, L., and F. Valencia. 2008. "Systemic banking crises: a new database."

Laeven, L., and F. Valencia. 2013. "SYSTEMIC BANKING CRISES DATABASE." *IMF Econ Rev*.

Lindgren, Carl Johan. 1999. "Financial sector crisis and restructuring: lessons from Asia."

Liu, X., K. W. Bowyer, and L. O. Hall. 2004. "Decision Trees Work Better Than Feed-Forward Back-Prop Neural Nets for A Specific Class of Problems âĹŮ."

Long, Philip M., and Rocco A. Servedio. 2010. "Random classification noise defeats all convex potential boosters." *Machine Learning*.

Marais, M.L., J.M. Patell, and M.A. Wolfson. 1984. "The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications." *Journal of accounting Research*.

Mehra, Yash P. 2004. "The output gap, expected future inflation and inflation dynamics: another look."

Mendis, Chandima. 2002. "External Shocks and Banking Crises in Developing Countries: Does the Exchange Rate Regime Matter?"

Peltonen, Tuomas A. 2006. "Are emerging market currency crises predictable?: A test."

Picard, Richard R., and R. Dennis Cook. 1984. "Cross-Validation of Regression Models." *Journal of the American Statistical Association*.

Rose, A.K., and M.M. Spiegel. 2010*a*. "Cross-country causes and consequences of the crisis: An update." *European economic review*.

Rose, Andrew K, and Mark M Spiegel. 2010*b*. "Cross-Country Causes and Consequences of the 2008 Crisis: International Linkages and American Exposure." *Pacific Economic Review*.

Rose, Andrew K., and Mark M. Spiegel. 2012. "Cross-country causes and consequences of the 2008 crisis: Early warning." *Japan and the World Economy*.

Rossi, Marco. 1999. "Financial Fragility and Economic Performance in Developing Economies-Do Capital Controls Prudential Regulation and Supervision Matter?"

Sachs, J., A. Tornell, and Velasco A. 1996. "Financial Crises in Emerging Markets: The Lessons From 1995." *NBER Working Papers*.

Schularick, M., and A.M. Taylor. 2011. "Credit booms gone bust: monetary policy, leverage cycles and financial crises, 1870–2008."

Schularick, Moritz, and Alan M Taylor. 2012. "Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870–2008." *American Economic Review*.

Segal, Mark R. 2004. "Machine learning benchmarks and random forest regression."

Stock, James H, and Mark W Watson. 2002. "Forecasting using principal components from a large number of predictors." *Journal of the American statistical association*.

Stock, James H, and Mark W Watson. 2006. "Forecasting with many predictors." *Handbook of economic forecasting*.

World Economic Outlook: Financial Crises: Causes and Indicators. 1998. "World Economic Outlook: Financial Crises: Causes and Indicators."

Appendix

Algorithm A1 Random Forest Pseudocode

for $b = 1$ to B **do** ▷ Estimate B Classification Trees

1. Draw bootstrap sample X^b of size N (with replacement).

2. Estimate regions $\{R_m^b\}_{m=1}^M$ through recursive partitioning:

repeat ▷ Recursive Partitioning

a) Draw (without replacement) J_{try} predictors from X^b.

b) Select splitting predictor and split point according to (9).

until Stopping rule applies

$\Rightarrow$ Region estimates $\{\widehat{R}_m^b\}_{m=1}^M$:

3. Estimate crisis probabilities $\{p_m^b\}_{m=1}^M$ according to (5).

$\Rightarrow$ Classification Tree $\widehat{\mathcal{T}_b^{RF}}(X_i) = \sum_{m=1}^M \widehat{p}_m^b I(X_i \in \widehat{R}_m^b)$

end for

$\Rightarrow$ B Classification Trees $\{\widehat{\mathcal{T}_b^{RF}}(X_i)\}_{b=1}^B$

$\Rightarrow$ Random Forest $\widehat{\mathcal{RF}}(X_i) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mathcal{T}_b^{RF}}(X_i)$ ▷ Generate Ensemble

Notes: The pseudocode shows the steps involved in generating of the $\mathcal{RF}$-estimator (10). This pseudocode also covers the generation of the $\mathcal{F}$-estimator (7) if the number of randomly drawn predictors J_{try} in step 2. a) is set equal to the total number of predictors J – in this case the expressions for the $\mathcal{F}$- and the $\mathcal{RF}$-estimator become equivalent.

Table A1: Indicators

Indicator	Obs.	Mean	S.D.	Min	Max
Bank Assets (gap)(glo)	2380	2.77	4.58	-14.78	17.29
Bank Assets (gr)(glo)	2397	0.05	0.04	-0.06	0.15
Bank Assets/GDP (gap)(glo)	2380	-0.02	1.02	-3.64	3.27
Bank Assets/GDP (glo)	2414	9.03	2.25	2.56	15.52
Bank Assets/GDP (gr)(glo)	2397	0.00	0.01	-0.02	0.03
Broad Money (gap)(glo)	2380	3.06	5.42	-8.25	33.44
Broad Money (gr)(glo)	2397	0.04	0.03	-0.07	0.16
Broad Money/GDP (gap)(glo)	2380	0.11	2.35	-9.80	13.76
Broad Money/GDP (glo)	2414	10.46	1.68	8.70	21.53
Broad Money/GDP (gr)(glo)	2397	0.00	0.02	-0.05	0.12
C (gap)	2261	1.28	7.53	-45.41	45.42
C (gap)(glo)	2380	1.19	4.17	-7.88	11.32
C (gr)	2278	0.02	0.06	-0.43	0.56
C (gr)(glo)	2397	0.02	0.02	-0.07	0.14
C/GDP	2140	8.64	19.78	0.00	137.95
C/GDP (gap)	2079	0.96	6.40	-27.68	76.03
C/GDP (gap)(glo)	2380	0.65	5.31	-26.15	20.54
C/GDP (glo)	2414	10.04	4.93	4.99	19.92
C/GDP (gr)	2109	-0.01	0.06	-0.40	1.26
C/GDP (gr)(glo)	2397	-0.01	0.04	-0.17	0.22
Exchange Rate (gap)	2380	3.21	25.98	-55.63	398.60
Exchange Rate (gr)	2397	249,929.16	12236321.00	-0.70	599080064.00
Exchange Rate (n)	2414	45.79	196.22	0.00	2,172.90
GDP (gap)	2192	1.22	7.17	-62.04	52.57
GDP (gap)(glo)	2380	1.49	4.97	-8.65	26.28
GDP (gr)	2223	0.02	0.06	-0.58	0.89
GDP (gr)(glo)	2397	0.02	0.03	-0.06	0.15
I (gap)(glo)	2380	4.56	23.95	-66.79	176.65
I (gr)(glo)	2397	0.08	0.34	-0.57	3.71
I/GDP (gap)(glo)	2380	1.67	17.70	-46.66	109.89
I/GDP (glo)	2414	0.18	0.04	0.10	0.25
I/GDP (gr)(glo)	2397	0.03	0.23	-0.36	2.36
Inflation	2363	0.04	0.15	-0.39	3.44
Inflation (glo)	2397	0.04	0.07	-0.08	0.37
LT Interest Rate	2259	0.02	0.12	-3.41	0.47
LT Interest Rate (gap)	2197	-0.00	0.09	-2.58	0.72
LT Interest Rate (gap)(glo)	2363	-0.00	0.04	-0.18	0.19
LT Interest Rate (glo)	2397	0.02	0.06	-0.34	0.14
LT Interest Rate (n)	2300	0.06	0.03	0.01	0.24
LT Interest Rate (n)(gap)	2242	-0.00	0.01	-0.13	0.07
LT Interest Rate (n)(gap)(glo)	2380	-0.00	0.01	-0.03	0.02
LT Interest Rate (n)(glo)	2414	0.05	0.02	0.02	0.13
LT Interest Rate Diff.	2259	-0.00	0.10	-3.14	0.41
LT Interest Rate Diff. (n)	2300	0.00	0.02	-0.09	0.18
Loans (gap)	2151	4.50	15.41	-82.26	95.43
Loans (gap)(glo)	2380	3.34	9.29	-22.86	24.00
Loans (gr)	2177	0.05	0.12	-0.76	2.49
Loans (gr)(glo)	2397	0.05	0.06	-0.11	0.24
Loans/GDP	2101	10.00	1.00	7.42	13.56
Loans/GDP (gap)	2035	0.05	2.63	-13.02	11.20

Continued on next page

Table A1: Indicators (continued)

Indicator	Obs.	Mean	S.D.	Min	Max
Loans/GDP (gap)(glo)	2380	-0.07	2.83	-9.85	8.54
Loans/GDP (glo)	2414	9.68	2.14	2.81	15.04
Loans/GDP (gr)	2068	0.00	0.02	-0.09	0.11
Loans/GDP (gr)(glo)	2397	0.00	0.01	-0.05	0.05
Narrow Money (gap)	2150	3.59	11.86	-42.11	86.63
Narrow Money (gap)(glo)	2380	3.56	7.71	-17.73	30.03
Narrow Money (gr)	2183	0.04	0.10	-0.39	1.08
Narrow Money (gr)(glo)	2397	0.04	0.05	-0.13	0.32
Narrow Money/GDP (gap)	2075	0.68	3.08	-15.16	22.85
Public Debt (gap)	2135	5.75	22.14	-84.51	200.59
Public Debt (gap)(glo)	2380	8.97	16.09	-14.96	93.85
Public Debt (gr)	2169	0.04	0.17	-0.74	3.71
Public Debt (gr)(glo)	2397	0.05	0.12	-0.16	0.88
Public Debt/GDP	2207	0.54	0.39	0.02	2.70
Public Debt/GDP (gap)	2135	13.98	109.21	-86.34	2,136.51
Public Debt/GDP (gap)(glo)	2380	8.97	16.09	-14.96	93.85
Public Debt/GDP (glo)	2414	0.66	0.29	0.30	2.23
Real Exchange Rate	2384	10.00	1.00	7.96	16.54
Real Exchange Rate (gap)	2342	-0.04	3.10	-22.52	30.93
Real Exchange Rate (gr)	2363	0.00	0.02	-0.28	0.25
ST Interest Rate (gap)(glo)	2363	-0.00	0.04	-0.17	0.23
ST Interest Rate (glo)	2397	0.01	0.07	-0.54	0.17
ST Interest Rate (n)(gap)(glo)	2380	0.00	0.01	-0.03	0.03
ST Interest Rate (n)(glo)	2414	0.05	0.02	0.01	0.14
Stock Prices (gap)(glo)	2380	22.10	62.05	-44.65	391.73
Stock Prices (gr)(glo)	2397	0.03	0.15	-0.44	0.65

Notes: (n) nominal; (gr) growth; (glo) global GDP-weighted average; (gap) percentage deviation from (one-sided) HP-trend ($\lambda = 1600$).

Table A2: Indicators, annual post-1970 sample

Indicator	Obs.	Mean	S.D.	Min	Max
C (gap)	5821	3.15	10.69	-68.27	107.37
C (gap)(glo)	6440	3.50	2.95	-3.57	9.71
C (gr)	5979	0.04	0.10	-0.63	2.23
C (gr)(glo)	6601	0.05	0.03	-0.03	0.14
C/GDP	6157	0.66	0.17	0.04	1.76
C/GDP (gap)	5841	2.87	205.36	-55.09	15656.46
C/GDP (gap)(glo)	6440	2.08	3.57	-3.17	9.43
C/GDP (glo)	6762	0.58	0.03	0.51	0.64
C/GDP (gr)	5999	0.00	0.09	-0.64	2.20
C/GDP (gr)(glo)	6601	-0.00	0.04	-0.08	0.09
Exchange Rate (n)	6197	301.52	1341.10	0.00	20509.75
Exchange Rate (n)(gap)	5881	14.69	42.79	-94.35	389.33
Exchange Rate (n)(gr)	6039	3.6e+05	2.8e+07	-0.96	2.2e+09
GDP (gap)	5821	3.35	12.31	-58.18	165.64
GDP (gap)(glo)	6440	2.18	4.31	-6.64	8.99
GDP (gr)	5979	0.02	0.06	-0.62	0.93
GDP (gr)(glo)	6601	0.03	0.03	-0.05	0.08
I (gap)	5821	12.62	278.75	-87.11	20011.98
I (gap)(glo)	6440	4.93	9.61	-24.69	21.98
I (gr)	5979	0.06	0.44	-25.24	12.63
I (gr)(glo)	6601	0.08	0.13	-0.24	0.61
I/GDP	6157	0.23	0.09	-0.01	1.13
I/GDP (gap)	5841	1.99	21.27	-85.42	383.85
I/GDP (gap)(glo)	6440	0.56	7.31	-27.57	10.69
I/GDP (glo)	6762	0.27	0.03	0.21	0.33
I/GDP (gr)	5999	0.02	0.51	-32.52	14.21
I/GDP (gr)(glo)	6601	0.02	0.11	-0.22	0.53
Inflation	5979	3.7e+05	2.9e+07	-0.45	2.2e+09
Inflation (glo)	6601	12671.34	80145.44	0.05	5.2e+05
LT Interest Rate (gap)(glo)	6440	-0.13	0.83	-2.13	1.75
LT Interest Rate (gap)(glo)	6279	-0.12	0.84	-2.10	1.80
LT Interest Rate (glo)	6601	5.85	3.04	1.54	12.73
LT Interest Rate (n)(glo)	6762	5.79	3.02	1.52	12.83
Loans (gap)	4936	14.44	149.33	-100.00	9615.57
Loans (gap)(glo)	6440	4.71	9.75	-20.19	29.17
Loans (gr)	5099	3.8e+06	2.7e+08	-1.00	2.0e+10
Loans (gr)(glo)	6601	326.77	2065.77	-0.17	13390.86
Loans/GDP	5294	10.00	0.99	7.43	15.29
Loans/GDP (gap)	4960	0.66	4.71	-23.28	40.09
Loans/GDP (gap)	6440	-0.36	2.46	-7.74	4.36
Loans/GDP (glo)	6762	10.21	2.06	3.25	13.78
Loans/GDP (gr)	5127	0.01	0.04	-0.31	0.58
Loans/GDP (gr)(glo)	6601	0.01	0.02	-0.07	0.05
Net Exports (gap)	5813	-4.94	1510.24	-5.0e+04	60853.19
Net Exports (gap)(glo)	6440	10.31	91.74	-254.37	260.84
Net Exports (gr)	5964	-0.33	17.49	-1002.06	321.87
Net Exports (gr)(glo)	6601	-0.55	3.63	-20.80	5.47
Net Exports/GDP	6157	-0.06	0.17	-2.34	0.82
Net Exports/GDP (gap)	5833	-7.08	3359.35	-1.6e+05	1.0e+05
Net Exports/GDP (gap)(glo)	6440	-98.12	941.21	-5760.65	1575.54

Continued on next page

Table A2: Indicators, annual post-1970 sample (continued)

Indicator	Obs.	Mean	S.D.	Min	Max
Net Exports/GDP (glo)	6762	-0.00	0.03	-0.05	0.06
Net Exports/GDP (gr)	5984	-0.34	16.79	-956.17	323.74
Net Exports/GDP (gr)(glo)	6601	-0.55	3.45	-19.72	5.38
Public Debt (gap)	4780	4.90	88.01	-100.00	4382.34
Public Debt (gr)	5009	0.07	0.27	-1.00	7.12
Public Debt (gr)(glo)	6601	0.08	0.22	-0.26	1.14
Public Debt/GDP	5388	62.18	65.19	0.00	2092.92
Public Debt/GDP (gap)	4913	7.20	158.25	-100.00	8135.19
Public Debt/GDP (gap)(glo)	6762	40.54	15.50	11.93	68.85
Public Debt/GDP (gap)(glo)	6440	3.00	16.65	-21.91	50.54
Public Debt/GDP (gap)(glo)	6440	3.00	16.65	-21.91	50.54
Real Exchange Rate	5883	130.80	604.90	0.00	14489.79
Real Exchange Rate (gap)	5581	16.66	41.07	-50.26	389.22
Real Exchange Rate (gr)	5732	3.9e+05	2.9e+07	-0.61	2.2e+09
ST Interest Rate (gap)(glo)	6279	-26.53	303.98	-696.23	1529.51
ST Interest Rate (glo)	6601	77.58	351.66	3.36	2278.75
ST Interest Rate (n)(gap)(glo)	6440	-25.78	300.19	-696.12	1529.66
ST Interest Rate (n)(glo)	6762	76.06	347.62	3.40	2278.90
Stock Prices (gap)(glo)	6440	61.01	71.81	-57.72	281.64
Stock Prices (gr)(glo)	6601	0.05	0.38	-1.08	1.01

Notes: (n) nominal; (gr) growth; (glo) global GDP-weighted average; (gap) percentage deviation from (one-sided) HP-trend ($\lambda = 1600$).

Table A3: Indicators, quarterly post-1970 sample

Indicator	Obs.	Mean	S.D.	Min	Max
CPI	22279	65.27	59.65	0.00	2430.25
CPI (glo)	30960	58.60	41.91	9.88	171.58
Exchange Rate (n)(gap)	27866	4.16	51.78	-100.00	6907.62
Exchange Rate (n)(gr)	28051	14.76	2201.33	-1.00	3.7e+05
Foreign Liabilities (gap)(glo)	30600	23.76	43.45	-7.04	461.93
Foreign Liabilities (glo)	30960	215.68	159.83	16.88	807.14
Foreign Liabilities (gr)(glo)	30780	1.3e+06	1.7e+07	-0.07	2.2e+08
Foreign Liabilities (n)(gap)(glo)	30600	18.75	70.67	-5.16	918.87
Foreign Liabilities (n)(glo)	30960	6.6e+10	8.6e+11	61.98	1.1e+13
Foreign Liabilities (n)(gr)(glo)	30780	7.1e+06	7.8e+07	-0.04	1.0e+09
Foreign Liabilities/GDP (gap)(glo)	30600	0.15	1.92	-7.18	16.06
Foreign Liabilities/GDP (glo)	30960	10.00	0.32	9.23	13.09
Foreign Liabilities/GDP (gr)(glo)	30780	-0.06	0.72	-9.26	0.41
GDP (gap)(glo)	30600	114.90	497.82	-2079.58	4659.59
GDP (glo)	30960	7404.10	41252.75	-5.4e+04	2.8e+05
GDP (gr)(glo)	30780	0.26	6.27	-41.26	45.26
GDP (n)(gap)(glo)	30600	57.42	234.64	-603.37	1628.23
GDP (n)(glo)	30960	3.55	1.97	-3.68	11.47
GDP (n)(gr)(glo)	30780	0.33	6.18	-40.44	43.46
House Prices (gap)(glo)	30060	-0.00	2.88	-8.26	4.33
House Prices (glo)	30420	3.75	3.32	1.32	12.27
House Prices (gr)(glo)	30240	0.00	0.01	-0.04	0.05
House Prices (n)(gap)(glo)	30600	0.55	3.02	-6.65	10.21
House Prices (n)(glo)	30960	373.19	495.26	32.72	1603.55
House Prices (n)(gr)(glo)	30780	0.02	0.02	-0.03	0.13
Inflation	21866	0.04	0.35	-1.00	40.58
Inflation (glo)	30780	0.04	0.04	0.00	0.38
LT Interest Rate (glo)	30780	9.76	2.92	5.71	24.97
LT Interest Rate (n)(glo)	30960	10.69	3.38	5.84	32.91
LT Interest Rate Diff. (glo)	30780	-0.00	0.00	-0.00	0.00
LT Interest Rate Diff. (n)(glo)	30960	-0.00	0.00	-0.00	0.00
Loans	20266	1015.99	2947.06	0.00	61949.44
Loans (gap)	19775	20.82	1828.59	-88.54	2.6e+05
Loans (gap)(glo)	30600	16.96	136.05	-2.95	1779.97
Loans (glo)	30960	973.98	571.51	369.29	2594.73
Loans (gr)	19961	0.02	0.11	-1.00	3.74
Loans (gr)(glo)	30780	0.02	0.02	-0.01	0.07
Loans (n)	24234	2.0e+14	3.1e+16	0.00	4.8e+18
Loans (n)(gap)	23729	7.86	165.20	-100.00	24531.77
Loans (n)(gap)(glo)	30600	7.53	13.14	-0.86	167.15
Loans (n)(glo)	30960	1.7e+14	2.2e+15	539.40	2.9e+16
Loans (n)(gr)	23920	2.6e+07	4.1e+09	-1.00	6.3e+11
Loans (n)(gr)(glo)	30780	2.2e+07	2.9e+08	0.01	3.8e+09
Loans/GDP (gap)(glo)	30600	0.08	1.03	-4.61	4.84
Loans/GDP (glo)	30960	9.97	0.20	9.50	11.13
Loans/GDP (gr)(glo)	30780	0.00	0.25	-2.56	1.94
Real Exchange Rate (gap)	21563	0.08	2.25	-20.46	53.54
Real Exchange Rate (gr)	21760	0.00	0.02	-0.25	0.77
Reserves	20982	7.1e+11	2.1e+13	-1.07	9.8e+14
Reserves (gap)	20514	307.29	27492.34	-256.77	3.8e+06

Continued on next page

Table A2: Indicators, quarterly post-1970 sample (continued)

Indicator	Obs.	Mean	S.D.	Min	Max
Reserves (gap)(glo)	30600	310.91	2723.12	-8.43	34990.50
Reserves (glo)	30960	1.0e+12	2.6e+12	21.55	1.2e+13
Reserves (gr)	20719	0.03	0.66	-41.28	39.68
Reserves (gr)(glo)	30780	0.03	0.08	-0.49	0.42
Reserves (n)	25411	14489.74	96910.52	-35.40	3.5e+06
Reserves (n)(gap)	24975	13.42	379.33	-787.19	53967.28
Reserves (n)(gap)(glo)	30600	13.03	28.28	-10.11	320.63
Reserves (n)(glo)	30960	12554.25	15958.76	760.46	64814.01
Reserves (n)(gr)	25178	0.07	0.86	-44.34	59.68
Reserves (n)(gr)(glo)	30780	0.07	0.09	-0.21	0.48
Reserves/GDP (gap)(glo)	30600	0.07	1.17	-3.99	4.47
Reserves/GDP (glo)	30960	9.96	0.18	9.52	10.81
Reserves/GDP (gr)(glo)	30780	0.00	0.10	-1.18	0.45
ST Interest Rate (glo)	30780	5011.86	57413.27	3.61	7.5e+05
ST Interest Rate (n)(glo)	30960	4330.27	49689.00	3.59	6.5e+05
ST Interest Rate Diff. (glo)	30780	0.00	0.00	-0.00	0.01
ST Interest Rate Diff. (n)(glo)	30960	-0.00	0.00	-0.01	0.00
Stock Prices (gap)(glo)	30600	7.48	14.26	-41.12	62.00
Stock Prices (glo)	30960	0.60	0.28	0.25	1.65
Stock Prices (gr)(glo)	30780	0.01	0.07	-0.30	0.16
Stock Prices (n)(gap)(glo)	30600	9.95	60.50	-39.78	785.33
Stock Prices (n)(glo)	30960	610.50	6942.76	5.97	91265.20
Stock Prices (n)(gr)(glo)	30780	0.04	0.07	-0.29	0.29

Notes: (n) nominal; (gr) growth; (glo) global GDP-weighted average; (gap) percentage deviation from (one-sided) HP-trend ($\lambda = 1600$).

Table A4: Banking Crises and Variable Selection

Publication	**Method**	Domestic		External		Financial				Fiscal
		GDP	CPI	CA	ER	Bank Assets	Money	Stock Prices	Interest Rates	Public Debt
Sachs, Tornell, and A. (1996)	OLS	(/GDP gap)	-	(/GDP)	(rer gap)	(/GDP gap)	-	-	-	-
Caprio and Klingebiel (1996)	Frequency	(r gap)	-	-	-	(r gap)	-	-	-	-
Brenda González-Hermosillo et al. (1997)	Logit	-	-	-	(n gr)	(r) (/GDP)	-	-	(r)	-
Demirgüç-Kunt and Detragiache (1998)	Logit	(r gr)	(gr)	-	(n gr)	(/GDP) (r gr)	(r)	-	-	-
Detragiache and Demirgüc-Kunt (1998)	Logit	(r gr)	(gr)	-	(n gr)	(r) (/GDP) (r gr)	(r)	-	(r)	-
Eichengreen and Rose (1998)	Probit	(gr) (r gr glo)	-	(/GDP)	(rer)	(gr)	-	-	(glo)	(gap)
Hardy and Pazarbasioglu (1998)	Logit	(r gr)	(gr)	-	(rer gr)	(/GDP gr)	(/GDP gr)	-	(r)	-
(1998)	Event Analysis	(r gr)	(gr)	(/GDP)	(n gr) (rer)	(/GDP)	(r)	(gr)	-	(gap /GDP)
Kaminsky (1998)	Signals	(r)	-	-	(rer gap)	(/GDP)	(r) (r gap)	(n)	(r) (r glo)	-
Brüggemann and Linne (1999)	Signals	(gr)	-	-	(gr rer)	(gr /GDP)	(gr)	-	(r)	(r gr)
Gourinchas, Valdes, and Landerretche (1999)	Descriptives	(r) (gap)	(gr)	(/GDP)	(rer gap)	(/GDP gap)	-	-	(n) (r glo)	(/GDP gap)
Gonzalez-Hermosillo (1999)	Logit FE	(r gr)	(n)	-	(n gr)	(r) (/GDP)	-	-	(r) (r diff)	-
Hutchison and McDill (1999)	Signals	(r gr)	(gr)	-	(n gr)	(r gr)	(r)	(gr)	(n gr) (r gr)	(gap)

Continued on next page

Publication	Method	Domestic		External		Financial				Fiscal
		GDP	CPI	CA	ER	Bank Assets	Money	Stock Prices	Interest Rates	Public Debt
Kaminsky and Reinhart (1999)	Signals	(gr)	-	(r)	(rer gap)	(/GDP)	(r) (r gap)	(gr)	(r) (r diff)	-
Lindgren (1999)	Frequency	-	(gr)	(gap /GDP)	-	(/GDP) (r gr)	-	(/GDP)	-	(/GDP gap) (/GDP)
Rossi (1999)	Logit, FE	(r) (r gr)	(gr)	-	-	(r gr) (/GDP)	-	-	(r)	-
Demirguc and Detragiache (2000)	Logit	(r gr)	(gr)	-	(gr)	(r gr)	(r)	-	-	-
Glick and Hutchison (2000)	Signals Probit	(r gr)	(gr)	-	(gr)	-	-	-	-	-
Hawkins and Klau (2000)	Indices	-	-	-	-	(/GDP gr)	(/GDP gap)	-	(r)	-
Honohan (2000)	Mean comp.	-	-	-	-	(r) (gr)	-	-	-	(gap)
Goldstein, Kaminsky, and Reinhart (2000)	Signals	(r gr)	-	(gr)	(rer gap)	(/GDP gr)	(r gap) (r gr) (gr)	(gr)	(r) (diff)	(/GDP) (gr) (gap /GDP)
Bordo et al. (2001)	Logit	(r) (r gr)	(gr)	-	-	-	(r)	-	-	(r gap)
Borio and Lowe (2002)	Signals	-	-	-	(r gap)	(/GDP gap)	-	(r gap)	-	-
Eichengreen and Arteta (2002)	Probit	(gr) (r gr glo)	-	(/GDP)	(rer gap)	(gr)	-	-	(glo)	(gr /GDP)
Hutchison (2002)	Probit	(r)	(gr)	-	(gr)	-	-	-	-	-
Mendis (2002)	Logit FE IA	(r gr)	(gr)	-	(rer)	(/GDP)	(r gr)	-	(n)	(/GDP gap)
Borio and Lowe (2004)	Signals	(r gap)	-	-	-	(/GDP gap)	(/GDP gap)	(r gap)	-	-
Demirguc-Kunt and Detragiache (2005)	Logit	(r gr)	(gr)	-	(n gr)	(/GDP) (r gr)	(r)	-	-	-
Davis and Karim (2008*a*)	Signals Logit	(gr) (r gr)	(gr)	(r)	(rer gap) (n gr)	(/GDP) (/GDP) (r gr)	(r) (r gap)	(gr)	(r) (r diff)	-
Davis and Karim (2008*b*)	Logit, $\mathcal{CT}$	(r gr)	(gr)	-	(n gr)	(/GDP) (r gr)	(r)	-	(r)	-

Continued on next page

Publication	**Method**	Domestic		External		Financial				Fiscal
		GDP	CPI	CA	ER	Bank Assets	Money	Stock Prices	Interest Rates	Public Debt
Borio and Drehmann (2009)	Signals	-	-	-	-	(/GDP gap)	-	(r gap)	-	-
Schularick and Taylor (2011)	OLS Logit, FE	-	(gr)	-	-	(r gr) (r gap) (/GDP)	(r gr)	(r) (r gr)	(r) (n)	-
Jordà, Schularick, and Taylor (2011)	Logit	(r)	(gr)	(/GDP) (gr /GDP)	-	(gr /GDP)	-	(r gr)	(r) (n)	-
Alessi and Detken (2011)	Signals	(r gap) (r gr)	(gr) (gap)	-	(rer gap) (rer gr)	(r glo gap) (/GDP gap) (r gr) (r gap)	(r gap) (r gap glo) (r gr) (/GDP gap)	(r gap) (/GDP gap)	(n gap) (n) (glo gap) (r gap) (r) (r glo)	-
Casu, Clare, and Saleh (2011)	Signals	-	-	-	-	(gr gap) (/GDP gap) (n gap)	-	(n gap)	(n gap)	-
Duttagupta and Cashin (2011)	$\mathcal{CT}$	(r gr)	(gr)	(r)	(n gr)	(r gr)	(r)	-	(r)	-
Gourinchas and Obstfeld (2011)	Logit	(r gap)	-	(/GDP)	(rer gap)	(/GDP)	-	-	-	(/GDP)
Drehmann and Juselius (2012)	Signals	-	-	-	-	(/GDP gap)	-	-	(/GDP)	-
Frankel and Saravelos (2012)	OLS Probit	(r) (r gr) (r gr gap)	(gr gap)	(r) (/GDP)	(rer gap)	(gr /GDP)	(r gap)	(r gap)	(r)	-
Eicher, Christofides, and Papageorgiou (2012)	BMA	(r) (r gr) (r gr gap)	(gr gap)	(r) (/GDP)	(rer gap)	(gr /GDP)	(r gap)	(r gap)	(r)	-
Hahm, Shin, and Shin (2012)	Probit RE	(r gr glo)	(gr)	-	-	(/GDP gap)	(r)	-	(glo)	-
Drehmann (2013)	Signals	-	-	-	-	(/GDP gap)	-	-	-	-
Jordà (2013)	Logit IA, FE	-	-	(/GDP gap)	-	(/GDP gap)	-	-	(n)	(/GDP gap)
Drehmann and Juselius (2013)	Signals	(r gr)	-	-	-	(r gr) (/GDP gap)	(r)	(n gr) (gap)	(r)	-

Notes: (r) real; (n) nominal; (gr) growth; (glo) global GDP-weighted average; (gap) percentage deviation from (one-sided) HP-trend ($\lambda = 1600$); $\mathcal{CT}$ Classification Tree; IA Interaction Terms; FE Fixed Effects; RE Random Effects; BMA Bayesian Model Averaging; MIMIC Multiple Indicator Multiple Cause Model. There is an overlap between 3rd

generation currency crises and banking crises, also termed twin crises. Publications on these 3rd generation currency crises have been included in the table if they exhibit a focus on the banking crisis aspect. The table does not list all variables the authors use in each publication. Instead I mapped the variables into the variable systematization I use throughout this article. The focus is on macroeconomic predictors (see Gavin and Hausmann, 1996) while microeconomic, (e.g. González-Hermosillo, 1996; Caprio and Klingebiel, 1996; Gonzalez-Hermosillo, 1999), political or institutional (see Acemoglu et al., 2003) factors are not listed in the table. Also, interaction terms between any two variables included in the table are not made explicit. In some cases the mapping is rather coarse. For example, non-core liabilities (Hahm, Shin, and Shin, 2012) are listed as a real monetary variable in table A4.The main aim of the table is to give an impression of the variability and selectivity *in* - and not an exact overview *of* - variable selections in the literature on banking crises.