Detection of rare events: a machine learning toolkit with an application to banking crises

Jérôme Coffinet¹ and Jean-Noël Kien²

Banque de France – Directorate general statistics

VERY PRELIMINARY – PLEASE DO NOT QUOTE

Abstract

In this paper, we consider a machine learning toolkit applied to the detection of rare events, namely banking crises. For this purpose, we consider a broad set of macroeconomic series (credit-to-GDP gap, house prices, stock prices, inflation rates, long-term and short-term interest rates, etc.), in combination with their leads and lags, various filtering methodologies, and data science models that complement time series analysis. The main advantages of the approach are its robustness, its flexibility and its prediction performance. Based on the best model specification, our methodology allows to compute an indicator representing the probability of banking crisis along with an alert threshold up to 6 quarters ahead in real time for various developed economies.

Keywords: artificial neural networks; banking crises; early warning system; financial stability

JEL classification: C52; C53; E58; G21

¹ jerome.coffinet@banque-france.fr. The results presented in the paper do not necessarily coincide with the views of the Banque de France. All errors remain our sole responsibility.
² jean-noel.kien@banque-france.fr
1. Introduction

Understanding the roots of financial crises is a crucial stake for today’s economics but predicting them in order to reduce their costs remains a key challenge for economic policy (Zidong et al., 2018). Following the global financial crisis, a wide bunch of economists has proposed various indicators and numerous methods aiming at a better identification of risks to financial stability.

Among them, the *credit-to-GDP* gap is currently widely used in the euro area and in developed countries as the early warning indicator of risks to financial stability. Precisely, its importance rests on its easiness to compute. As a matter of fact, many influential papers originating the Bank for International Settlements demonstrate its usefulness in determining whether or not a countercyclical capital buffer should be activated (Drehmann et al., 2010; Drehmann et al., 2011; Drehmann, 2013; Drehmann and Tsatsaronis, 2014). In that context, the Basel Committee on Banking Supervision provided a “Guidance for National Authorities Operating the Countercyclical Capital Buffer” (Basel Committee on Banking Supervision, 2010), where the computation of the credit-to-GDP was considered as a standard tool, computed under certain circumstances, for implementing the Countercyclical capital buffer.

Like many Basel standards, its use is not binding but has to be transposed into national or regional law for that purpose. For instance, in the euro area, the European Systemic Risk Board (ESRB) recommendation of the 18 June 2014 adopted it on the basis of Articles 135 and 136 of Capital Requirements Directive IV. Hence, for instance in France, the *Haut Conseil de Stabilité Financière*, in charge of setting the countercyclical capital buffer, relies on the *credit-to-GDP* gap defined by the Basel Committee in order to determine when and what countercyclical capital buffer should be applied.

Nonetheless, what is more interesting, the ESRB regulation leaves open the possibility of using alternative methods (‘When the designated authorities consider that another method of measuring and calculating the credit-to-GDP ratio gap better reflects the specificities of the national economy, it is recommended that they measure and calculate a quarterly difference in the ratio additional credit-GDP, in addition to the difference calculated in accordance with point 1.’). As a result, a different method of calculation can be used but the underlying economic quantities should remain.
Our paper rests on this open question and proposes a machine learning toolbox so as to anticipate banking crises as accurately as possible, with the aim to deliver unambiguous forward-looking and operationally-relevant signals to measures risks to financial stability. For that purpose, we examine the predictive power of various economic and financial indicators of banking crises. Hence, we provide a model of the probability of banking crisis along with an alert threshold up to 6 quarters ahead in real time and for 15 developed countries over the period 1970Q1-2017Q2.

From the economic point of view, our contribution is twofold: first, we test whether additional variables to the credit-to-GDP gap are likely to improve the detection of risks to financial stability (O’Brien and al., 2018), and second we consider explicitly the possibility that the variables of interest are not synchronized, which improves our optimization process. Hence, we show that our strategy improves the standard methodology.

From a purely methodological point of view, our contribution is twofold. First, we introduce additional filter methodologies to the HP-filter retained in the standard credit-to-GDP methodology, in particular considering Kalman filters. Second, as many authors have decided to tackle this challenge adopting a wide variety of methodologies, from standard econometric models (Barrell et al., 2010) to machine learning methods (Ristolainen, 2018), we also build a comprehensive toolkit for crisis identification that includes not only basic logit models, but also artificial neural networks, which proved promising in several general public communications (Spanish Foundation for Science and Technology, 2015) and more specialized academic publications (Chirita, 2012; Ecer, 2013).

All in all, we demonstrate that additional variables are useful complements to the usual credit-to-GDP gap in order to assess the risks to banking and financial stability. Based on the best model specification, our methodology allows to compute a reworked indicator with the ability to produce the probability of banking crisis along with an alert threshold up to 6 quarters ahead in real time for various developed economies. Using additional variables to the credit-to-GDP improves significantly the prediction of crisis for about two thirds of the countries. As a result, we end up with more reliable probabilities of banking crisis than the ones stemming from the credit-to-GDP gap taken alone. Our results are crucial to the operational implementation and policy-making decisions related to the countercyclical capital buffer.
The remainder of the paper is structured as follows. Section 2 discusses the Basel Committee methodology. Section 3 presents the data at hand and our data science framework. Section 4 interprets the results. Section 5 concludes.

2. Designing a comprehensive methodology

2.1 The usual credit-to-GDP methodology

In Europe, the countercyclical buffer is implemented by national authorities that compute and publish every quarter a countercyclical buffer rate of reference that matches the Basel Committee on Banking Supervision guidance. This guidance relies on the usual credit-to-GDP gap defined at each time \( t \) (quarter) as follows:

\[
\text{basel}\_\text{gap}_t = \frac{\text{credit}_t}{\text{gdp}_t} - \tau_t
\]

where \( \tau_t \) is the trend of the credit-to-GDP ratio obtained with a recursive Hodrick-Prescott filter (Hodrick & Prescott, 1997) with \( \lambda = 400000 \) for the smoothing parameter. With this definition, the \( \text{basel}\_\text{gap}_t \) is also called the cycle component. Figure 1 presents the trend and the usual credit-to-GDP obtained from the standard methodology for France.

[Figure 1]

Using this standard credit-to-GDP at hand, the Basel Committee provides an \textit{ad hoc} rule (Basel Committee on Banking Supervision, 2010) to determine when and how a countercyclical buffer should be activated. A buffer is activated when the usual credit-to-GDP gap reaches 2% point. Its rate grows linearly until the gap reaches 10% point (Figure 2).

[Figure 2]

2.2 Some improvements to the usual method
The Basel methodology (hereafter ‘benchmark’ methodology) strongly relies on the credit-to-GDP gap and its computation. In fact, the countercyclical buffer is activated or deactivated based on a threshold and the rate of the buffer depends on the amplitude of the gap. Hence, the amplitude of the usual credit-to-GDP directly determines the countercyclical buffer.

The filtering step

The proposed calculation of the credit-to-GDP gap uses a very high smoothing parameter ($\lambda = 400000$). However, Hodrick and Prescott (1997) suggest using a smoothing parameter $\lambda = 1600$ for quarterly time-series. This value is still a standard for the analysis of economic cycles in time series. The benchmark credit-to-GDP methodology breaks this principle, on the basis of a result from Ravn and Uhlig (2002) stating that the smoothing parameter $\lambda$ should be adjusted depending on the length of the cycles. Then, the proposed methodology includes the a priori hypothesis that credit cycles are from 3 to 4 times longer on average than economic cycles. The corresponding values for the smoothing parameter $\lambda$ are from $3^4 \cdot 1600 \sim 125000$ to $4^4 \cdot 1600 \sim 400000$. The best $\lambda$ is chosen such that the noise-to-signal ratio is minimal (Drehmann et al., 2010). This suggests choosing a value of $\lambda = 400000$ to get the best results.

Nevertheless, some choices underlying the Basel gap hereafter ‘benchmark’ methodology can be challenged. The choice of the activation threshold (2%) seems to have been set a priori, thus the optimal $\lambda$ is only valid for this level. A signal of 1 (0) is judged to be correct if a crisis (no crisis) occurs any time within a three-year horizon. Hence, a flexible horizon as in Borio and Lowe (2002) is considered in this methodology (Drehmann et al., 2010). This three-year horizon offers flexibility but it is large and may lead to errors of prediction. The noise-to-signal ratio and consequently the choice of $\lambda$ can be biased by this flexibility. Finally, “while a simple moving average or a linear time trend could be used to establish the trend, the Hodrick-Prescott filter is used in this regime as it has the advantage that it tends to give higher weights to more recent observations. This is useful as such a feature is likely to be able to deal more effectively with structural breaks” (Barrell et al., 2010). The type of filter used has not been challenged yet and choosing a Hodrick-Prescott filter might not be the best option (Hamilton, 2016) because, by construction, the Hodrick-Prescott filter produces distortions on borders which limit its utilization in real-time.
The predictive property

The predictive property of the standard *credit-to-GDP* gap is questionable. In fact, the way it is built makes it a tool which is more explicative than predictive:

- At time \( t \), we have at hand the proposed *credit-to-GDP* gap at time \( t \) with the previous values. The only conclusion that can be made is whether or not there will be at least one crisis between \( t \) and \( t+3 \) years because \( \lambda \) has been chosen with a 3 years flexible horizon. This is too weak and approximate. No information concerning the number of crises, their probability of occurrence or a precise horizon of occurrence is provided;

- The standard *credit-to-GDP* gap is not reactive enough (see Figure 1 (a)). In fact, it takes too much time for the trend to include structural changes like the one between 1994 and 1999. The increasing trend during this period is questionable. More generally, the filtered gap just presents a delay compared to the raw series;

- The magnitude of the proposed *credit-to-GDP* gap directly determines whether or not a countercyclical capital buffer should be activated and the value of the buffer in percentage. However, depending on the threshold arbitrarily fixed, the methodology could have missed the internet bubble of 2001/2002 (see Figure 1 (b) with a threshold of 5% point) or in any case, could have activated an insufficient buffer;

- The methodology is not universal: all the parameters fixed and learnt (\( \lambda \) and the threshold of detection) are attached to the learning sample and there is no guarantee that these parameters will be identical for other countries or other sectors (private sector, public sector, etc.).

We have highlighted the rooms for improvement of the benchmark *credit-to-GDP* gap methodology. In order to tackle these issues we will present our data science framework in the next section.

The desynchronization of credit and GDP series
The standard methodology considers the ratio of credit-to-GDP for a given economy during a given time span. Yet, there might be, depending of the structural features of the economy, a natural dynamics of series that induces a non-zero phase between credit and GDP series (Coffinet and Montornès, 2014). As a result, when the credit series is forward-looking with respect to the GDP series, the credit-to-GDP ratio estimated at time $t_1$ (cf. Figure 3) is over the one estimated at time $t_2$, independently from any policy measure.

[Figure 3]

If credit constantly lags behind GDP, the credit-to-GDP ratio increases and decreases accordingly without signaling any divergence of credit compared to GDP. Our prior is that the credit-to-GDP ratio at the peak of each series taken separately, that corresponds to the natural dynamics of the variables, is more relevant from a policy point of view than the ratio of contemporaneous numerator and denominator. From the example above, the intuition is quite simple: deciding the implementation of a countercyclical capital buffer based on the ratio observed at time $t_1$ would be misleading because the natural dynamics of the series should yield an inferior ratio at time $t_2$. The risk is hence to impend the financing of the economy while being inefficient as regards financial stability purposes, and finally increase the risks of pro-cyclicality of the policy decision.

3. Building our Data science framework

3.1 The data

A systemic banking crisis is when a major disruption in the financial systems occurs. It consists in country’s corporate and financial sectors experiencing a large number of defaults and financial institutions and corporations facing great difficulties repaying contracts on time. As a result, non-performing loans increase sharply and a large part of the aggregate banking system capital is exhausted. This situation may be accompanied by depressed asset prices (such as equity and real estate prices) on the heels of run-ups before the crisis, sharp increases in real interest rates, and a slowdown or reversal in capital flows. In some cases, the crisis is triggered by depositor runs on banks.
Building a database of crisis occurrence is a complex task because opinions of experts diverge. Many authors over years have been studying such databases (Laeven and Valencia, 2008; Laeven & Valencia, 2010; Laeven & Valencia, 2012; Reinhart & Rogoff, 2013; Reinhart & Rogoff, 2011; Kaminsky & Reinhart, 1999; Caprio & Klingebiel, 2003). The data we chose is a quarterly database of systemic banking crisis occurrences for 41 countries and at most from 1970Q1 to 2010Q4. This database (Babecký, et al., 2014) has been built with experts from central banks, international institutions and universities on top of previous works cited above. It is also used in recent working papers such as Coudert & Idier (2016). Over the most recent period (from 2010 to 2017), we add information on European countries from Lo Duca et al. (2017) and on other countries from Baron et al. (2018).

The database of explanatory variables is composed of the following series:

- The ratio of credit-to-GDP and the related credit and GDP series come from the database on macroprudential indicators from the Bank for International Settlements, related to IMF-IFS Claims on private sector for the credit series. Credit encompasses the outstanding amount of loans and bonds to the non-financial private sector, which is a broad definition of the private sector financing. Consistently with the Basel definition, the GDP series consists of the nominal gross domestic product. The series is quarterly from 1969Q4 to 2017Q2;

- House prices statistics are available at a quarterly frequency over the whole sample. They encompass for each country the most important house price index as identified by the Bank for International Settlements, which is also the source of the series and allows for some cross-country homogeneity of the variables in that respect. Data gathered on the BIS website are complemented with national sources (national statistical institutes or national central banks) when needed;

- We also consider the following financial series, that are all taken from the International Financial Statistics database (IFS) of the IMF: the inflation rate is defined as the year-on-year growth rate of the consumer price index; the short-term interest rate is defined as the 3-month interbank money market rate when available or the overnight money market rate when the former is not available; the long-term interest rate is the 10-year government bond yield; share prices correspond to the
domestic stock market index. Consistently with the prior that the latest available information is relevant for policy decisions, all daily financial variables are transformed so as to take the end-of-quarter observations in the final dataset.

Data are collected over 32 countries: Brazil, Canada, Chile, China, Czech Republic, Greece, Hungary, India, Indonesia, Korea, Malaysia, Mexico, New Zealand, Poland, Portugal, South Africa, Thailand, as well as Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Netherlands, Norway, Spain, Sweden, Switzerland, United Kingdom and United States.

3.2 The algorithm

We build a Data science framework to predict the probability of banking crisis along with an alert threshold up to 6 quarters ahead in real time and for various developed economies. It is split in 5 steps:

1. **Importance of lags**: in this part, we present the need to include lags in the credit and the GDP when building the \( \frac{\text{credit}}{\text{gdp}} \) ratio;

2. **Filtering**: in this part, we explore alternatives to the Hodrick-Prescott filter that could overcome its limitations;

3. **Additional variables**: we investigate whether adding variables improves the forecast of crises;

4. **Predictive model**: in this part, we present the models used to predict crisis occurrences;

5. **Validation**: this part is devoted to the validation of the models and how best models are selected.

3.3 Importance of lags
Based on the rationale presented in Section 2.2, our proposal, in the paper framework, is to let the model optimize the best combination of leads and lags between credit and GDP, within the limit of 6 quarters, in order to predict banking crises.

3.4 Filtering

As we can notice in our proposed methodology, the filtering step is essential. In fact, it enables to remove the trend in time series. The remaining part is what is used to detect financial bubbles. Filtering the original time series has also another advantage. It can be viewed as a way to normalize the underlying phenomenon by focusing on its dynamics. Hence, a universal methodology across countries can be implemented.

We also have qualitatively tested several filters, namely Beveridge-Nelson, Savitzky-Golay, Baxter-King, Christiano-Fitzgerald, Butterworth, trigonometric regression, cubic splines, LOESS regression and different specifications of dynamic linear models. As a result, we decided to choose the class of dynamic linear models. Indeed, this class of models is very flexible, allowing modeling every component of a time series (trend, cycle, seasonal components). Moreover, certain filters can be set up as a specific dynamic linear model (the Beveridge-Nelson filter for instance), as explained in more details in Petris et al. (2009).

In a nutshell, dynamic linear models are a class of models for time series, including multivariate time series. They are particularly popular among Bayesians. Associated with a dynamic linear model, the Kalman filter allows us to update our beliefs about the current value of the unobserved vector (the trend) each time we incorporate a new observation.

Finally, after testing qualitatively different types of dynamic linear models, we have decided to extract the trend $\tau_t$ from the \textit{credit-to-GDP} ratio by modeling the phenomenon with a dynamic linear model named \textit{local level model or random walk + noise model}. This model is defined as the following state-space model:

$$\frac{\text{credit}_t}{\text{gdp}_t} = \tau_t + v_t, \quad v_t \sim N(0,V_t)$$

$$\tau_t = \tau_{t-1} + w_t, \quad v_t \sim N(0,W_t)$$
where $v_t$ is the noise of the observations and $w_t$ the noise of the model. In fact, at each time $t$, it is not the true trend that is observed but a noisy version of it. The unobserved true trend is modeled as a random walk with a noise. This noise $w_t$ is called the noise of the system or the noise of the model. In our specification, the noise of the model should be low because only the trend is modeled.

In our case, we will rather estimate the signal to noise ratio $\frac{w}{V}$ rather than $W$ and $V$ separately. This ratio will be optimized in the estimation of the models. To avoid being stuck in a local minima like this is the case with a quasi-Newton method, the optimization of the ratio is done by estimating models with the ratio picked from a grid of values (from 400 to 6000 by 200). Then, the value giving the best quality of prediction is selected.

As a result, we will use two filters: the recursive Hodrick-Prescott filter as a benchmark for our proposed credit-to-GDP gap methodology and the local level model. We will also refer to the local level model as the Kalman filter model by reference to the algorithm it uses. When using additional variables, a year-on-year variation is computed to the house prices and the share prices time.

3.5 Wrapping up in predictive models

We aim to model the banking crises in order to predict them. Usually, when the phenomenon is binary, the natural choice is to use a logistic model. We also have decided to use a neural network considering the possibilities to model complex and non-linear interactions between variables and robustness when structural changes occur.

The neural network used is a so-called “multilayer perceptron” from the class of feedforward artificial neural networks. We test 2 configurations: 1 hidden layer with 2 neurons and 2 layers with sequentially 4 and 2 neurons. This choice enables stable estimations with convergence of optimization steps at stake. The configuration with the best predictions is chosen.

For the variables included in the models, we have chosen 4 possibilities:
- Include only one variable from the set of ratios $\left\{ \frac{\text{credit}_{t-k}}{\text{gdp}_{t-l}}, \ k, l \in [1,6] \right\}$. As we noticed in Section 3.3, adding lags in both credit and GDP should be considered;

- Include only one variable from the set of ratios $\left\{ \frac{\text{credit}_{t-k}}{\text{gdp}_{t-l}}, \ k, l \in [1,6] \right\}$ with all the additional variables;

- Include all combinations of the $\frac{\text{credit}}{\text{gdp}}$ simultaneously. This means we include all variables from the set $\left\{ \frac{\text{credit}_{t-k}}{\text{gdp}_{t-l}}, \ k, l \in [1,6] \right\}$. The purpose here is not to include a single value but sequences of values that represent the piecewise dynamics;

- Include all combinations of the $\frac{\text{credit}}{\text{gdp}}$ simultaneously with all the additional variables.

The models have to give the probability of banking crisis up to 6 quarters ahead in real time. This forces the models to have the following form. Assuming we want to predict the probability of banking crisis in $h$ quarters, then the model is:

$$prob_t = f(X_{t-h})$$

where $prob_t$ is the probability of banking crisis at time $t$ and $X_{t-h}$ is all information set available up to time $t-h$. Once estimated, the predicted probabilities as follows read:

$$\hat{prob}_t = \hat{f}(X_{t-h})$$

where $\hat{prob}_t$ is the probability of banking crisis at time $t$ and $\hat{f}$ the model estimated. Then, the predicted probability of banking crisis in $h$ quarters from time $t$ can be obtained by shifting the model as follows:

$$\hat{prob}_{t+h} = \hat{f}(X_t)$$

As a result, when we want to predict the probability of banking crisis in $h$ quarters, the estimation of such a model consists in learning the crisis at time $t$ knowing the predictors up to time $t-h$. In fact, the predictors between time $t-h+l$ and $t$ are unknown once the model is shifted.
For instance, suppose we have estimated the following model:

\[ \text{prob}_t = \hat{f}\left(\frac{\text{credit}_{t-4}}{\text{gdp}_{t-3}}\right) \]
	hen, at time \( t \), we can predict the probability of banking crisis in at most \( h = \min(|4|,|3|) = 3 \) quarters.

We have recapitulated in Table 1 the following possible models in competition and the quality of prediction on a test sample will determine which one is best (see Section 3.6):

[Table 1]

3.6 Validation

At this step, we have a group of models in competition to predict the probability of banking crisis for each horizon \( h \). In order to compare them and determine what model is the best, we defined a test protocol that consists in stopping the learning process of all classes of models on a period ending in 2004Q4, and using the 2005Q1-2017Q2 period as a validation sample allowing for the validation of different models.

Our quantitative assessment of the model performance is based on the ‘Area Under the Curve’ criterion of the ‘ROC (Receiver Operating Characteristic) Curve’. More specifically, Area Under the Curve is a criterion for evaluating classification models by reporting "true positive" signals (the proportion of episodes actually detected) to "false positives" (proportion of episodes without crises identified as episodes of crisis ). When the area under the curve (cf. Figure 4) approaches 1 (green curve), the model is discriminating and makes it possible to identify the periods of crises and non-crises. Conversely, when the area approaches 0.5, the model does not distinguish between true and false signals and corresponds to a 'random' detection of episodes of crises. Conditionally to the choices of modeling, the optimization of the probability of detection and the probability of false alarm equally weighted in the out-of-sample prediction exercise makes it possible to determine the thresholds from which the probability of crisis can be interpreted.
In so doing, we gauge the capability of the different procedures to predict the last financial crisis, which acts as an ‘experimental proof’. Nonetheless, we reduce significantly our sample of interest as our specification imposes that for each country we can observe at least one crisis in the learning sample (before 2005) and at least one crisis in the test sample (after 2005). As long as we refer to the crisis identification explained in Section 3.1, we are compelled to retain only 15 countries presenting those characteristics: Belgium, Denmark, Finland, France, Germany, Ireland, Italy, Japan, Netherlands, Norway, Spain, Sweden, Switzerland, United Kingdom and United States.

Finally, although our framework is flexible because it allows selecting the best model for one country and one horizon of prediction, we decided to force the model to be the same for all horizons. This guarantees some stability between predictions during different horizon. The flexibility of the model by country is still conserved.

4. Results

4.1 Overall results

Figure 5 presents the distribution of obtained AUC over different classes of models, with reference to the benchmark model, which is the strict credit-to-GDP BIS-recommended model, with simultaneous credit and GDP observations.

At the first glance, we obtain the following results:

- On average, the best class of models is the one including a unique credit-to-GDP ratio with a Hodrick-Prescott filter, but with desynchronized credit and GDP as the AUC distribution is significantly over 0. Moreover, the distribution of AUC is quite small, which tends to indicate that the proposed method tends to maximize model homogeneity across countries. Hence, our assessment tends to confirm in general the
relevance of the proposed methodology, but underlines the importance of lags and leads in the dynamics of credit and GDP;

- Using multiple ratios does not provide any significant gain on the AUC distribution, while using neural networks seems quite interesting as it scarcely over performs the unique ratio models but is likely to improve the model performance at the top of the distribution: as a result, neural networks appear as a rare but not marginal gain on the prediction performance.

Figures 6 and 7 present the same distributions but re-ordered in order to analyze the effects of filtering choices and additional variables.

Figure 6 shows that the Kalman filter only improves the average prediction when it is used with a unique ratio and additional variables.

[Figure 6]

Figure 7 indicates that introducing additional variables is often beneficial for the different configurations, but the one using a unique ratio with HP filter. In particular, it improves the top of the distribution for some of model classes, especially the ones based on neural networks.

[Figure 7]

Taking the best models without continuity constraint, 40% retain additional variables from all horizons. The figure 8 below suggests that the higher the horizon, the less the additional variables are useful.

[Figure 8]

As it is difficult to isolate simply the marginal gain of using one specification rather than another one, and the contribution of each innovation, we carry out a variance analysis (ANOVA) on our distribution of AUC figures.
We find that all effects except additional variables are 95% significant. The additional variables are at 90%. Switzerland and the United States prove atypical countries with a significant negative effect on the quality of forecasting. All things being equal, estimating a model specifically for a country decreases the quality of forecasting. This means that we have a set of learning countries that provides useful information. The filtering is significant, and all else being equal, the Kalman filter decreases the quality of prediction, on average. The effects of Multivariate and Neural Networks models are not significant, which means that the model class does not significantly weigh on the quality of the prediction. Finally, the effect of the additional variables is significant with a positive effect on the forecast quality.

[Table 2]

4.2 Which model to use for each country?

Figure 9 presents the AUCs by country obtained on the one hand with the benchmark model without distinction of country and on the other hand with benchmark models estimated by country. It is interesting to note that a country model offers better forecast quality for Japan, Norway, United Kingdom, Belgium, France, Netherlands and the United States. On the other hand, a country can also benefit from the global information provided by other countries, such as the Spain, Switzerland, Italy, Denmark, Ireland and above all Germany.

[Figure 9]

Figure 10 presents the AUC by country obtained on the one hand with benchmark models estimated by country and on the other hand univariate logistic models by country. Each univariate logistic model is composed of the ratio lagged credit and lagged GDP by retaining only the most optimal lag combination by country. For all countries but Germany, univariate logistic models provide better forecasting quality.

[Figure 10]

By comparing (Figure 11) the performance of univariate logistic models by country with the benchmark model without distinction of countries, we succeed to get a better forecasting performance for Sweden for instance, which was not possible with the benchmark model by
country. Other countries (Spain, Denmark, Ireland and Germany) continue to benefit from the global information provided by other countries. The forecast quality for these same countries can be improved but it will then be a more sought-after model.

[Figure 11]

Here in Table 3 are the models we obtain country by country from the explained methodology used when imposing the same model by country whatever the horizon:

[Table 3]

To some extent, our results echo those of Borgy et al. (2014), who show that the predictive power of housing prices, stock prices, inflation, short-term and long-term interest rates complement that of credit-to-GDP, as also recognized in particular for house prices by Aldasoro et al. (2018). In that respect, Fielding and Rewilak (2015) explain that booms increase the probability of a crisis only in relatively fragile financial systems. Its proves that in our estimations, Kalman filters and Neural networks model are never included in the best models by country. However, the use of desynchronized ratios, of country-specific models and of additional variables is often of added value. In addition, it comes that the goodness-of-fit of the best model by country might strongly differ as shows the high dispersion of AUC statistics, and that the alarm threshold could also proves very diverse, which pleas agains a one-size-fits-all approach for estimating banking crises probabilities. As a result, we can recommend the use of our methodology for operational purposes, but with the additional information that using additional variables can guarantee its robustness across prediction horizons.

4.3 A normative view of the benchmark methodology

Our methodology builds endogenously for each possible configuration (model class, univariate vs. multivariate, neural network vs. logit, with or without additional variables) a threshold above which a crisis is detected. We name it the “alert threshold”. Thus, we are able to construct even for the benchmark methodology, which a strict application of the method defined by the Basel gap standards, a set of “alert threshold”.

17
Those are represented in the following Figure 12 (in red), together with the predicted probability of crisis (in blue). It illustrates that the benchmark model cannot be used as a one-size-fits-all method as the alert threshold might significantly differ across countries. It is also remarkable that in some cases the benchmark model is unable to deliver a reliable threshold (too close from 0 or 1). This highlights that using the benchmark methodology could be misleading because, whatever the threshold chosen, the power of detection cannot be accurate.

[Figure 12]

4.4 Some case studies

Results obtained on a set of countries allow for a comparison of different in-sample and out-of-sample features. Our results are based on the last available observation for each series in 2017Q4. The prediction for crisis detection is hence made up to 6 quarters ahead, that is to say until 2019Q2.

For most countries, we get a probability of banking crisis quite different, and often inferior, to that the one stemming from the standard credit-to-GDP model. This underlines the importance of using alternative and better designed models, as well as auxiliary variables, since the banking crisis detection based on the credit gap estimate would overvalue the risk of a crisis. In the following Figure 13, we present a comparison of the estimated probability of crisis stemming from the benchmark model and from our described methodology in France and Germany.

[Figure 13]

In Figure 14, we present the expected probability of crisis as estimated from the procedure (in blue) to be interpreted with respect to the estimated ceiling (in red): when the blue line goes above the red line, there is a risk of banking crisis in the corresponding country in the resultant time horizon.
From a more economic and financial stability point of view, we find that some countries experience a state of the economy that should be considered as perilous: Denmark, Ireland and Italy on the short-term, and Switzerland on the long-term.

[Figure 14]

5. Conclusion

Predicting banking crises is challenging. We have presented the standard credit-to-GDP methodology and its limitations. In fact, its ability to predict incoming crises precisely is questionable. In order to assess the predictive power of this method, we have proposed alternatives in terms of filtering and modelling.

We have proposed a local level model as a modern way to extract the trend from a time series. It has many advantages, including being easily usable in real-time. We have demonstrated that including several lag versions of the credit and the GDP series in the models helps to detect crises. In fact, dynamic evolutions are better suited than a single punctual value when it comes to describing a phenomenon that evolves over time. Concerning models, we have proven that it is better to model the crisis with the data than just using arbitrary thresholds. Besides, using additional variables such as the inflation rate, the short-term interest rate, share prices, house prices in addition to lagged versions of the credit and the GDP, improves the detection.

According to our study, the framework well suited to detect banking crisis is strongly country-specific, depending on the lags and leads used in the credit and GDP dynamics, the inclusion of several rations rather than only one, the use of additional variables or the recourse to country-specific vs. cross-country models. Nonetheless, the added value of the HP filter and that of logit models (rather than neural networks) appears somewhat marginal and restricted to some very specific cases. In that sense, our results tend to validate those choices of the benchmark method.

In future works, some improvements could be considered such as challenging the smoothing parameter \( \lambda \) of the HP filter, optimizing the specification of neural networks and tuning thresholds in order to prioritize the ability to detect crisis more than limiting the false alarm
rate. Last but not least, a new angle that could be considered is building a unique model for all countries with fixed effect to capture specificities.
References


Appendix

Figure 1

(a) The credit-to-GDP ratio (in black) and \( \tau_t \) the trend (in blue)

(b) The standard credit-to-GDP or cycle (in blue)

Figure 2

Figure 3
Figure 4: ROC Curve and Area Under the Curve

Figure 5: distribution of the obtained AUC over different classes of models
Figure 6: distribution of the obtained AUC over different filters

Figure 7: distribution of the obtained AUC with additional variables
Figure 8: share of best models by horizon including additional variables

Figure 9
Figure 12: crisis predictions – benchmark models

Belgium

Denmark

Finland

France

Germany

Ireland
Figure 13

<table>
<thead>
<tr>
<th>Country</th>
<th>Hendrick-Prescott filter</th>
<th>Logistic model</th>
<th>Best model</th>
<th>Cf. Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td></td>
<td>$\frac{credit_{t-k}}{gdp_{t-k}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td>$\frac{credit_{t-k}}{gdp_{t-k}}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model Number</td>
<td>Methodology</td>
<td>Filters</td>
<td>Models</td>
<td>Ratio Set</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>---------</td>
<td>--------</td>
<td>-----------</td>
</tr>
<tr>
<td>(M0) Basel methodology benchmark</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Logistic model&lt;br&gt;- Single ratio among the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M1) Univariate model with HP filter</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Logistic model&lt;br&gt;- Any single ratio among the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M2) Univariate model with LLM filter</td>
<td>- Local level model&lt;br&gt;- Logistic model&lt;br&gt;- Any single ratio among the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$&lt;br&gt;- Additional variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M3) Univariate model with HP filter with additional variables</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Logistic model&lt;br&gt;- Any single ratio among the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$&lt;br&gt;- Additional variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M4) Univariate model with LLM filter with additional variables</td>
<td>- Local level model&lt;br&gt;- Logistic model&lt;br&gt;- Any single ratio among the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$&lt;br&gt;- Additional variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M5) Multivariate model with HP filter</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Logistic model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M6) Multivariate model with LLM filter</td>
<td>- Local level model&lt;br&gt;- Logistic model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M7) Multivariate model with HP filter with additional variables</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Logistic model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$&lt;br&gt;- Additional variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M8) Multivariate model with LLM filter with additional variables</td>
<td>- Local level model&lt;br&gt;- Logistic model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$&lt;br&gt;- Additional variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M9) Neural network model with HP filter</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Neural Network model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M10) Neural network model with LLM filter</td>
<td>- Local level model&lt;br&gt;- Neural Network model&lt;br&gt;- All ratio from the set $\left{ \frac{\text{credit}<em>{t-k}}{\text{gdp}</em>{t-l}}, k, l \in [1,6] \right}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(M11) Neural network model with HP filter with additional variables</td>
<td>- Hodrick-Prescott filter&lt;br&gt;- Neural Network model</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- All ratio from the set \( \{ \frac{\text{credit}_{t-k}}{\text{gdpr}_{t-l}}, \ k, l \in [1,6] \} \)
- Additional variables

(M12) Neural network model with LLM filter with additional variables
- Local level model
- Neural Network model
- All ratio from the set \( \{ \frac{\text{credit}_{t-k}}{\text{gdpr}_{t-l}}, \ k, l \in [1,6] \} \)
- Additional variables

Table 2: regression table for the AUC estimated over different models

<table>
<thead>
<tr>
<th>Explained variable : AUC (Nb obs: 403)</th>
<th>Intercept</th>
<th>0.589513</th>
<th>0.00 ***</th>
</tr>
</thead>
<tbody>
<tr>
<td>Belgium</td>
<td>ref</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denmark</td>
<td>-0.017576</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Finland</td>
<td>-0.066035</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>-0.048294</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Germany</td>
<td>-0.044355</td>
<td>0.43</td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td>-0.108762</td>
<td>0.06</td>
<td></td>
</tr>
<tr>
<td>Italy</td>
<td>0.116842</td>
<td>0.04</td>
<td>*</td>
</tr>
<tr>
<td>Japan</td>
<td>0.04531</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>0.017133</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>Norway</td>
<td>0.030024</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>0.027517</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>Sweden</td>
<td>-0.007751</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>Switzerland</td>
<td>-0.164878</td>
<td>0.00</td>
<td>**</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>0.059598</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>United States</td>
<td>-0.1248</td>
<td>0.03</td>
<td>*</td>
</tr>
<tr>
<td>Country-specific model</td>
<td>-0.05026</td>
<td>0.02</td>
<td>*</td>
</tr>
<tr>
<td>Kalman</td>
<td>-0.147657</td>
<td>0.00</td>
<td>***</td>
</tr>
<tr>
<td>Model benchmark</td>
<td>-0.090246</td>
<td>0.24</td>
<td></td>
</tr>
<tr>
<td>Model LOGITMULTIVARIATE</td>
<td>-0.071053</td>
<td>0.23</td>
<td></td>
</tr>
<tr>
<td>Model LOGITUNIVARIATE</td>
<td>-0.004642</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>Model Neural Network</td>
<td>-0.048518</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Additional variables</td>
<td>0.063362</td>
<td>0.00</td>
<td>**</td>
</tr>
<tr>
<td>Country</td>
<td>Country-specific model</td>
<td>Filter</td>
<td>Model</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------</td>
<td>--------</td>
<td>------------------</td>
</tr>
<tr>
<td>Belgium</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Denmark</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_MULTIVARIATE</td>
</tr>
<tr>
<td>Finland</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>France</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Germany</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Ireland</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_MULTIVARIATE</td>
</tr>
<tr>
<td>Italy</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Japan</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Netherlands</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Norway</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Spain</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_MULTIVARIATE</td>
</tr>
<tr>
<td>Sweden</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>Switzerland</td>
<td>NO</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
<tr>
<td>United States</td>
<td>YES</td>
<td>HP</td>
<td>LOGIT_UNIVARIATE</td>
</tr>
</tbody>
</table>
## Appendix 1: Systemic banking crisis data sources combined in the database

<table>
<thead>
<tr>
<th></th>
<th>Sources</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caprio and Klingebiel (2003)</td>
<td>The annual dataset (1970–2002) includes information on 117 episodes of systemic banking crises in 93 countries and on 51 episodes of borderline and non-systemic banking crises in 45 countries. A systemic crisis is defined as “much or all of bank capital was exhausted.” Expert judgment was also employed “for countries lacking data on the size of the capital losses, but also for countries where official estimates underestimate the problem.”</td>
</tr>
<tr>
<td>2</td>
<td>Kaminsky and Reinhart (1999)</td>
<td>The monthly dataset (1970–1995) includes 26 episodes of banking crisis in 20 countries. Banking crises are defined by two types of events: “(1) bank runs that lead to the closure, merging, or takeover by the public sector of one or more financial institutions; and (2) if there are no runs, the closure, merging, takeover, or large-scale government assistance of an important financial institution (or group of institutions) that marks the start of a string of similar outcomes for other financial institutions.” The dataset of banking crises was compiled using existing studies of banking crises and the financial press.</td>
</tr>
<tr>
<td>3</td>
<td>Laeven and Valencia (2008, 2010, 2012)</td>
<td>The annual dataset (1970–2011) covers systemically important banking crises (147 episodes) in over 100 countries all over the world and provides information on crisis management strategies. A banking crisis is considered to be systemic if the following two conditions are met: “(1) Significant signs of financial distress in the banking system (as indicated by significant bank runs, losses in the banking system, and/or bank liquidations); and (2) Significant banking policy intervention measures in response to significant losses in the banking system.” The first year that both criteria are met is considered to be the starting year of the banking crisis, and policy interventions in the banking sector are considered significant if at least three out of the following six measures were used: “(1) extensive liquidity support; (2) bank restructuring costs; (3) significant bank nationalizations; (4) significant guarantees put in place; (5) significant asset purchases; and (6) deposit freezes and bank holidays.” The dataset is compiled using the authors’ calculations combined with some elements of judgment for borderline cases.</td>
</tr>
<tr>
<td>4</td>
<td>Reinhart and Rogoff (2013, 2011)</td>
<td>The annual dataset (1800–2010, from the year of independence) covers banking crises in 70 countries. The definition of banking crisis is the same as in Kaminsky and Reinhart (1999) (see above). The dataset of banking crises was compiled using existing studies of banking crises and the financial press.</td>
</tr>
</tbody>
</table>