Answering the Queen: Machine Learning and Financial Crises [*]

Jérémy Fouliard
Paris School of Economics

Michael Howell
CrossBorder Capital

Hélène Rey
LBS, CEPR and NBER

Vania Stavrakeva
LBS and CEPR

October 1, 2023

Abstract

Financial crises cause economic, social and political havoc. Macroprudential policies are gaining traction but are still severely under-researched compared to monetary and fiscal policy. We use the general framework of sequential predictions, also called *online machine learning*, to forecast crises out-of-sample. Our methodology is based on model aggregation and is "meta-statistical", since we can incorporate any predictive model of crises in our analysis and test its ability to add information, without making any assumption on the data generating process. We predict systemic financial crises twelve quarters ahead out-of-sample with high signal-to-noise ratio. Our approach guarantees that picking certain time dependent sets of weights will be asymptotically similar for out-of-sample forecasts to the best *ex post* combination of models; it also guarantees that we outperform *any* individual forecasting model asymptotically. We analyse which models provide the most information for our predictions at each point in time and for each country, allowing us to gain some insights into economic mechanisms underlying the building of risk in economies.

[*]We are highly indebted to Pierre Alquier for his generous help. We thank Gilles Stoltz for multiple discussions and very useful comments. We are also very grateful to our discussants Fabrice Collard, Frank Diebold, Christian Julliard, Sebnem Kalemli Ozcan, Luc Laeven and Philip Lane as well as to Richard Portes and Nicolas Vayatis and seminar participants at the NBER IFM meeting, the ASSA 2021 Meetings, the Society for Nonlinear Dynamics and Econometrics 2020 Meetings, the Applied Machine Learning, Economics and Data Science seminar, UC Berkeley, MIT, Boston University, Cambridge Judge, Oxford, Bocconi, ENSAE, KTO-GREDEQ-OFCE, Nowcasting, the LBS AQR academic symposium, the Annual BIS Conference, Chicago conference in the honour of Alejandro Justiniano, the Bank of Spain Macroprudential Conference and the Women in Economics Chicago Conference for comments. Our thanks also go to Virginie Coudert, Julien Idier, Mikael Juselius, Anna Kovner, Marco Lo Duca, to Benoit Mojon and to Katja Taipalus for help with the database and Pierre Gaillard and Yannig Goude for the opera package. Rey thanks the ERC (Advanced Grant 695722). We thank Francesco Amodeo for outstanding research assistance. Jérémy Fouliard benefited greatly from the hospitality of the BIS where part of this paper was written. This work does not represent in any way the views of the French Macroprudential Authority. Email: hrey@london.edu; jfouliard@london.edu.

1 Introduction

In November 2008, the Queen of the United Kingdom visited the London School of Economics. After the failure of Lehman Brothers in September, the financial crisis was on everyone's mind. As she was shown graphs emphasising the scale of imbalances in the financial system, she asked a simple question: "Why didn't anybody notice?"After a rather terse reply on the spot[1] it took several months before the British Academy wrote a three-page missive to Her Majesty blaming the lack of foresight of the crisis on the "psychology of denial" that was widespread in financial and political circles who tended to believe that "financial wizards had found new and clever ways of managing risks". "So in summary, Your Majesty, the failure to foresee the timing, extent and severity of the crisis and to head it off, while it had many causes, was principally a failure of the collective imagination of many bright people, both in this country and internationally, to understand the risks to the system as a whole." This paper is an attempt to bring back some imagination in the economics of crises.

Financial crises cause economic, social and political havoc. The average cumulative output loss in a banking crisis (in deviation from its trend) is around 20% over the length of the crisis, which is on average two years, according to the database of Laeven and Valencia (2020). Systemic financial crises lead to large fiscal costs, major increases in public debt and they disrupt the fabric of our societies. In order to decrease their frequency and their severity a new set of tools has been introduced in many countries. Macroprudential policies aim at increasing the resiliency of the financial system as a whole by introducing countercyclical capital buffers for banks, liquidity coverage ratio requirements and allowing for tightening of lending standards at discretionary times chosen by the macroprudential authorities. While there is an extensive body of academic research on monetary and fiscal policies, there is still relatively little work which can guide macroprudential policies. In particular, implementing those policies requires a timely understanding of the build up of risk in the economy. As shown in the classic Reinhart and Rogoff (2009) book "This Time is Different, Eight hundred years of Financial Follies", financial crises have occurred

[1]In private correspondance Luis Garicano, who was Professor of economics at LSE indicated that he actually gave the following answer to the Queen at the time: "At every stage, someone was relying on somebody else and everyone thought they were doing the right thing." See also Financial Times, November 14 2008.

repeatedly in emerging markets and advanced economies alike, and they exhibit some remarkable similarities. In a recent survey on financial crises, Sufi and Taylor (2021) emphasize that since crises do not occur randomly, it is important to understand better the pre-crisis periods. Crises are often, but not always, "credit booms gone bust" but they also display some differences in their mechanics. There are many different theoretical models in macroeconomics and in finance which have been developed to understand them. Some emphasise runs as in Diamond and Dybvig (1983). Most macrofinance models focus on the bust phase of the crisis and on amplification mechanisms. A few analyze the boom phase of the financial cycle: they emphasise limited liability and asset overvaluations due to risk-shifting (Coimbra and Rey (2017)), search-for-yield in low interest rates environments (Martinez-Miera and Repullo (2017)), or deviations from rational expectations and financial constraints (Gennaioli et al. (2012)). From an empirical point of view, a number of variables have been used to predict financial crises. Following the classic work of Kaminski and Reinhart (1999), the literature has very usefully described the behaviour of a number of key variables around crisis episodes (see e.g. Gourinchas and Obstfeld (2012)). Lowe and Borio (2002) and Schularick and Taylor (2012) underline the role of credit growth; Jordà et al. (2015) emphasize the joint importance of credit growth and asset prices; and Mian and Sufi (2009) show the critical importance of household debt. Most of the literature uses standard econometric methods such as panel data econometrics or event studies in order to identify early warning indicators (EWI) of financial crises[2]. Some recent attempts to introduce new forecasting methods from the machine learning literature can be found in Ward (2017) who uses classification trees or Bluwstein et al. (2020) who compares the forecasting performance of decision trees, random forests, extremely randomised trees, support vector machines, and artificial neural networks.[3]

Our starting point is that the ability of existing models to predict systemic crises out-of-sample early and accurately (with small type I and type II errors i.e. the ability to predict all crises which

[2]EWIs have been developed to predict financial crises in emerging economies (Frankel and Rose (1996); Bussiere and Fratzscher (2006)). They were also applied to large panels of advanced and emerging economies (Demirgüç-Kunt and Detragiache (1998); Eichengreen and Arteta (2002); Bordo et al. (2001)). After the 2008 crisis, a number of studies have shown the relevance of EWIs using univariate or multivariate regressions (see e.g. Borio and Drehmann (2009), Alessi and Detken (2011), Shin (2013), Frankel and Saravelos (2012), or model aggregation methodologies (Coudert and Idier (2016)). None has considered robust and agnostic aggregation methodologies like our paper.

[3]For a comparison of logits and a subset of machine learning models, see Beutel et al. (2019).

actually happened without crying wolf too often) is still limited. Turning points and non linear phenomena such as crises have been notoriously difficult to predict out-of-sample. Price-based early warning indicators tend to be more coincident indicators than good predictors. Predicting pre-crisis periods (twelve quarters before the crisis) in order to give macroprudential authorities the time to act proves to be extremely difficult. Yet financial stability policies need this type of input. The complexity and the interaction of many variables, some of them -like asset prices- very fast moving, may also render the understanding of financial crises exceptionally difficult. In such a context, the "failure of the collective imagination of many bright people" is likely to be a permanent feature of the world.

Ideally, we would like to forecast systemic financial crises without knowing the "true" model of the economy, using as much information as possible (many possible models of the economy) in a way which is flexible enough to do dynamic evolving forecasting (weights put on different forecasting models should vary over time). We also want to avoid the problem of overfitting which is often present in macroeconomic forecasting (Stock and Watson (1996)) and leads to poor out-of-sample forecasting power (see e.g. Meese and Rogoff (1983) and Rossi (2011)). For these reasons, we adapt the *framework of sequential prediction or online machine learning* (see Cesa-Bianchi and Lugosi (2006) and Cesa-Bianchi and Orabona (2021)) which is a *model aggregation* methodology precisely designed to overcome these difficulties.

Since the seminal work of Bates and Granger (1969), forecast combinations are viewed as a simple and effective way to perform better than individual models (Timmermann (2006), Elliott and Timmermann (2008), Diebold and Shin (2019)).[4] However, online machine learning is different from the real-time forecast combination literature as it does not estimate the coeffifficients of some underlying model, but rather, chooses the best combination of weights to form the most accurate out-of-sample predictions sequentially *without making assumptions on the underlying data generating process*. It is robust with respect to this out-of-sample forecast process and not

[4] In the presence of structural change, Diebold and Pauly (1987), Pesaran and Timmermann (2005) argue that forecast errors can be reduced through systematic combination of forecasts. Capistrán and Timmermann (2009) shows how to combine individual survey forecasts (with entry and exit of individual forecasters) for real time forecasting. Diebold and Shin (2019) proves that a LASSO-based procedure that sets some weights to zero and shrink the surviving models towards equality outperform simple average and median forecasts (and perform almost as well as the ex post best forecaster) in an application to the European Central Bank Survey of Professional forecasters.

with respect to an underlying stochastic process that it would have to fifit (since this is not what it does). This is why this approach is described as ”meta-statistical”.

Online machine learning is specifically geared at real-time prediction in situations where the true models driving outcomes are not known and can be different over time. Since we do not make any assumption on the way the sequence to be predicted is generated, there is no baseline to assess the forecaster's performance. Instead, it is measured by how well the forecaster uses the available information to make his own prediction. This available information is composed of reference forecasters, also called *experts*[5]. We estimate these experts using standard macroeconomic variables (debt, GDP, unemployment, investment, credit, interest rates, monetary aggregates, asset prices, commodity prices, housing prices, external imbalances). These variables are the ones which would have come naturally to the mind of any macroeconomist familiar for example with the important work of Kindelberger on *Manias, Panics and Crashes* (Kindleberger (1978)). But really, these same variables would be considered by anyone reading the *debt-deflation theory of great depressions* (Fisher (1933)).

Our approach aims at making the best prediction by aggregating experts' forecasts. The forecaster's error is then the sum of two errors : an *estimation error* defined as the error of the best combination of experts, known *ex post*, representing the best prediction the forecaster can make using the available information[6] and an *approximation error* measuring the difficulty to approach *ex ante* the best combination of experts [7]. Though based on model averaging with time varying weights, *on-line learning* is more general than Bayesian Model Averaging[8]; importantly and as already mentioned, it does not make any assumption on the data generating processes; further-

[5]We are aware that this terminology could be misleading. In the real-time forecast combinations literature, experts often refer to individual survey forecasters. The term ”expert” we use throughout the paper stands instead for any forecasting model, variable, or individuals survey forecasts. It comes from the online learning literature and reminds us that the problem is not a standard estimation problem but that it can be viewed as a ”meta-statistical” approach. It is also worth pointing out that the only potential source of overfitting in our analysis could come from the estimation of individual forecasting models but not from the aggregation rule (designed to avoid overfitting). Our contribution is not in the estimation of individual experts but in their optimal aggregation (overfitted experts will quickly get low weights in the aggregation rule). Nevertheless, in section 6 we present a placebo test as a robustness check for the presence of overfitting in the estimation of individual experts.

[6]This error can thus be attributed to the experts' performances.

[7]This error can thus be attributed to the aggregation rule.

[8]In some cases, even very simple ones (see Grunwald and van Ommen (2014)), Bayesian Model averaging does not converge due to heteroskedasticity.

more it allows for time-varying learning rates. Online learning guarantees that picking certain time dependent sets of weights will be asymptotically similar for out-of-sample forecast to the best fixed combination of models, known only *ex-post*[9]. The beauty of this method is that it makes sure that we do *at least as well* as the best existing forecasting model in central banks or elsewhere asymptotically. Indeed if any model performs well in out-of-sample forecasts, we can just include it in our set of experts and it will be picked by our algorithm. We are therefore guaranteed to do at least as well as the literature asymptotically (and it turns out we actually do better -and often much better- in most of the cases we looked at).[10]

Online learning is well-suited for our problem. Unlike in classical statistical theory, where the sequence of outcomes is assumed to be a realization of a stationary stochastic process, in our framework, pre-crises are the product of some unknown and unspecified mechanism, which could be deterministic, stochastic, or even adversarially adaptative to our own behavior (Orabona (2021)). In fact, the framework of sequential predictions has been introduced in the pioneering works on repeated games by Robbins (1951), Hannan (1957), and Blackwell (1956), where the data source consists of the opponents' plays in a two-person game. This is why this framework has an intimate connection with game theory and it runs even deeper than its origins. Cesa-Bianchi and Lugosi (2006) shows that simple bounds for the performance of online algorithms can be seen as applications of the classical minimax theorems in game theory and generalized minimax theorems such as Blackwell approachability theorem can be used to define good forecasters. [11]

To our knowledge online machine learning has never been applied to macroeconomics (one exception is Amat et al. (2018) for exchange rates) though it has been used in a number of applications outside economics, for example to forecast electricity consumption (Devaine et al. (2013)),

[9]Online learning is thus a very different methodology from the recursive forecasting approach of Casabianca et al. (2019), Döpke et al. (2017), Ng (2014).

[10]The literature has often provided AUROC and RMSE as diagnostics for forecasting performance. They are useful but not a panacea: an expert whose forecast probability of crisis would increase from 0 to 0.000001 at the right time would have an AUROC of 1 but would not provide good early warning signals. We provide a broader set of diagnostics, graph transparently our probability forecasts and those of our experts and show that our aggregation rule has a lower average loss and a lower cumulated loss compared to any individual experts and to the uniform aggregation of our experts.

[11]Even more surprisingly, a fascinating line of research shows that if all players in a repeated normal form game adopt a simple regret-minimizing prediction strategy similar to online learning algorithms, the induced dynamics leads to a certain equilibrium (see for instance Fudenberg and Levine (1995), Hart and Mas-Colell (2000), Hart and Mas-Colell (2001), Stoltz and Lugosi (2007)).

to track the performance of climate models (Monteleoni et al. (2011)), to model the network traffic demand (Dashevskiy and Luo (2011)), to forecast air quality (Mallet et al. (2009)) and to predict the outcomes of sports games Dani et al. (2012). Some of the online learning techniques like exponential-weighted aggregation have also been studied in the statistical literature (Dalalyan and Tsybakov (2008), Alquier and Lounici (2011), Dalalyan and Salmon (2012) and Rigollet and Tsybakov (2012)).

An advantage of the methodology is that it also allows us to track which models perform well over time in a given country. This is an important characteristic which sets it apart from black box approaches and makes it even more suited to inform macroprudential policies. This is often enlightening to understand sources of instability -though of course we cannot formally identify any *causal relationship* between variables having good forecasting power and the origins of the crisis. Most of the predictions we make in the paper are *quasi real time predictions* in the sense that we do out-of-sample forecasts using historical data which may have been revised by statistical agencies. We also present a set of *real time* predictions on French and UK data using exclusively vintage time-series, which reduces considerably the set of variables we can incorporate in our models but validates the power of our approach. Despite its generality and its flexibility, *online-learning* has of course some limitations. It will be unable to predict any crisis of a type that has never happened in history. For example, it will not be able to predict a hypothetical financial crisis caused by a cyber-attack as we never observed one so far, or a financial crisis potentially caused by a pandemic shock unless its correlates with characteristics of past crises.

The structure of the paper is as follows. We present our database on systemic financial crisis dates as well as the different variables which we use to build our "experts" (predictive models) in section 2. In section 3, we describe the general methodology of sequential predictions and show how we can adapt it to our specific problem. An important issue in our case is the delayed revelation of information since we are seeking to predict pre-crisis periods, an information that is revealed only when a systemic crisis happens twelve quarters after the beginning of the pre-crisis period. In section 4 we present a horse race between a number of "off-the-shelf" experts (predictive models) present in the literature to which we add a few more experts (elastic-net logits) as well as bayesian averaging models and machine learning and statistical models (random forests, support vector machines, general additive models) to illustrate the power of our methodology. We assess predictive ability using different model aggregation rules and we present a number of diagnostics. In all cases we uncover a *time-varying subset of models* which carry most of the information to predict financial crises. Among those models we also discuss which ones "flash red" at the right time. Our online aggregators improve on the literature, beat individual models and provide very informative signals for policy makers. Section 5 concludes.

2 Data on systemic crises and macroeconomic variables

We need two datasets: the dating of systemic crisis episodes and a dataset of economic indicators for a panel of countries in order to construct forecasting models ("experts"). Experts will be estimated either on country specific data or on the entire panel. Due to data availability, the period under consideration is 1985q1 to 2019q3. We consider seven countries : France, Germany, Italy, Spain, Sweden, the United Kingdom and the United States. They include the largest eurozone economies, a small open economy and the two largest financial centres (US and UK).[12]

2.1 Definition and Data on Systemic Crisis Episodes

We borrow the definition and the dates of systemic crises from the Official European database constructed by the European Central Bank and the European Systemic Risk Board (Lo Duca et al. (2017)). We also rely on their narratives of the crises. This database has been put together to establish a common ground for macroprudential oversight and policymaking in the European Union. The dating of systemic crises is in part based on quantitative indicators but it is ultimately based on the expert judgement of the relevant national authorities. The methodology used is a two-step approach. Following Duprey et al. (2017), it aims at first identifying historical episodes of elevated financial stress which were also associated with real economic slowdowns using a quantitative analysis. The financial stress indicator captures three financial market segments: i)

[12]Nothing in the methodology limits the number of countries.

equity market: stock price index, ii) bond market: 10-year government yields and iii) foreign exchange market: real effective exchange rate. Industrial production growth is used as measure of real economic activity (see more details in Appendix A). At the end of this first step, a list of potential systemic crisis events, characterised by six consecutive months of real economic slowdown occurring within one year of the financial stress period is drawn. The second step uses a qualitative approach. Each national authority distinguishes between **systemic crises** and **residual episodes** of financial stress following common criteria. An event is classified as a systemic crisis event if it fulfils one or more of the following three criteria: i) A contraction in the supply of financial intermediation or funding to the economy took place during the financial stress event, ii) The financial system was distressed (market infrastructures were dysfunctional and/or there were bankruptcies among large financial institutions) and iii) Policies were adopted to preserve financial stability (external support, extraordinary provision of central bank liquidity, direct interventions of the state). Residual events are episodes of financial stress which are not as wide and serious as systemic crises. National authorities are asked whether they want to complement the list of events or change the timing of events already flagged. Their judgements prevail. This official database of crisis episodes is available for European countries. We replicated the exact same methodology for the United States[13].

We focus on predicting systemic crises from twelve quarters ahead, that is we predict pre-crises periods which are the twelve quarters preceding a systemic crisis[14]. This time interval of three years allows macroprudential policies to be put in place. For example, there is typically a four quarter delay once the decision of an increase in the countercyclical capital ratio is taken and the implementation of the decision by the banking sector; the diagnostic of the decision and the decision process itself take several more quarters. We also provide some robustness analysis for eight quarter ahead predictions[15]. Formally, we denote the systemic crisis characteristic function $C_{n,t}$:

[13]We are very grateful to the New York Fed and to Anna Kovner in particular for the US data.

[14]In other words we predict whether we are in a crisis now and/or whether we will be in a crisis in some quarter within the next 12 quarters.

[15]Shortening the forecast horizon to four quarter ahead does not give enough lead time to macroprudential authorities to implement their policies. From the point of view of the algorithm it has also the disadvantage of decreasing considerably the number of pre-crisis periods.

$$C_{n,t} = \begin{cases} 1 & \text{If there is a systemic crisis in country } n \text{ at time } t \\ 0 & \text{Otherwise} \end{cases}$$

We define the pre-crisis indicator $I_{n,t}$:

$$I_{n,t} = \begin{cases} 1 & \text{if } \exists h \in H = [0, 12] \text{ such that } C_{n,t+h} = 1 \\ 0 & \text{otherwise} \end{cases}$$

The variable that we will seek to predict out-of-sample is therefore $I_{n,t}$.

2.2 Macroeconomic and financial variables

We consider a large set of standard macroeconomic and financial variables X_k. We take into account the main risks on financial markets, real estate markets, credit markets and macroeconomic conditions. The variables of our analysis are the ones which would have come naturally to the mind of the reader of Kindleberger (1978)[16] or of Minsky (1986)[17]. But really, these same variables would be considered by anyone reading in 1933 in Econometrica the *debt-deflation theory of great depressions* by Irving Fisher[18]. We do not deny that in the set of the exact measures

[16]"By no means does every upswing in business excess lead inevitably to mania and panic. But the pattern occurs sufficiently frequently and with sufficient uniformity to merit renewed study. What happens, basically, is that some event changes the economic outlook. New opportunities for profits are seized, and overdone, in ways so closely resembling irrationality as to constitute a mania. Once the excessive character of the upswing is realized, the financial system experiences a sort of "distress," in the course of which the rush to reverse the expansion process may become so precipitous as to resemble panic. In the manic phase, people of wealth or credit switch out of money or borrow to buy real or illiquid financial assets. In panic, the reverse movement takes place, from real or financial assets to money, or repayment of debt, with a crash in the prices of commodities, houses, buildings, land, stocks, bonds -in short, in whatever has been the subject of the mania" .

[17]"The economy consists of a mixture of hedge, speculative and Ponzi financing units. A hedge financing unit can fail to meet its obligations only if its gross profits after taxes fall below expectations. In the aggregate this can happen only if there is a sharp fall in aggregate demand. A speculative financing unit can fail to meet its obligations if its income is below expectations, if interest rates rise too much or if there is a breakdown in the normal functioning of some set of financial markets. A Ponzi financing unit can run into troubles for all of the reasons that a speculative unit can plus the capitalizing of interest can erode the margin of safety in equity so that lenders are unwilling to continue capitalizing interest. An economy in which the dominant financing form is hedge financing will be financially robust. The greater the proportion of firms that are speculative or Ponzi financing the more fragile the financial structure. The basic theorem of the financial instability hypothesis is that over an extended period of prosperous times the weight of speculative and Ponzi finance in the total financial picture increases, so that the economy migrates from being financially robust to being financially fragile".

[18]"While quite ready to change my opinion, I have, at present, a strong conviction that these two economic maladies, the debt disease and the price-level disease (or dollar disease), are, in the great booms and depressions, more

we use some of them would not have been available historically (such as the VIX) but most of them (and actually the ones that tend to matter) would have been; and the economic concepts that all these variables measure were the ones described by this classic literature. Our database contains commonly used Early Warning Indicators with transformations (1-y, 2-y, 3-y change and gap-to-trend) for a panel of countries. We have a total of 244 quarterly variables, including the transformations, for our forecasts in quasi real time. Whenever we de-trend a variable we make sure we use only data of the estimation sample (and no future data to avoid look-ahead bias). We make use of OECD's Main Economic indicators and National Accounts databases, the Bank for International Settlements data and of the database of Cross Border Capital data (CBC) which contains monthly data series on liquidity aggregates (public and private), capital flows and risk indices.[19] . Importantly the CBC variables are available in revised format as well as in real-time. We have a smaller total of 122 variables, including transformations, for our real time analyses. The full list of variables and their sources is provided in Appendix A.

3 The Framework of Sequential Predictions

To predict the pre-crisis periods out-of-sample, we use the general framework of sequential predictions, also called *online machine learning* or *on-line protocol*. Consider a bounded sequence of observations (the occurence or non-occurrence of pre-crisis periods) $y_1, y_2, ..., y_T$ in an outcome space $\mathcal{Y}$. The goal of the forecaster is to make the predictions $\hat{y_1}, \hat{y_2}, ..., \hat{y_T}$ in a decision space $\mathcal{D}$.

important causes than all others put together. Some of the other and usually minor factors often derive some importance when combined with one or both of the two dominant factors. Thus over-investment and over-speculation are often important; but they would have far less serious results were they not conducted with borrowed money. That is, over-indebtedness may lend importance to over-investment or to over-speculation. The same is true as to over-confidence. I fancy that over-confidence seldom does any great harm except when, as, and if, it beguiles its victims into debt. Another example is the mal-adjustment between agricultural and industrial prices, which can be shown to be a result of a change in the general price level. Disturbances in these two factors, debt and the purchasing power of the monetary unit, will set up serious disturbances in all, or nearly all, other economic variables. On the other hand, if debt and deflation are absent, other disturbances are powerless to bring on crises comparable in severity to those of 1837,1873, or 1929-33" Fisher (1933).

[19] We also use a few variables from diverse sources: house price forecasts from the Survey of Professional Forecasters. Experimenting with many more variables could be interesting and our methodology is well-suited for this. We leave that for future research.

This framework has two main specificities. First, the observations $y_1, y_2, \ldots$, are revealed in a sequential order. At each step $t = 1, 2, ..$, the forecaster makes a prediction $\hat{y}_t$ on the basis of the previous $t-1$ observations before the *t*th observation is revealed. This is why this approach is said to be "online" since the forecaster sequentially receives information. The optimal forecasting model is adaptable over time which is very convenient when the predictive content is unstable over time. This lack of stability is indeed a stylized fact in the forecasting literature (Stock and Watson (2012) and Rossi (2011)). Second, in contrast to the stochastic modelling approach, we *do not assume* that $y_1, y_2, \ldots$ are the product of a stationary stochastic process. The sequence $y_1, y_2, \ldots$ could be the result of any unknown mechanism which is in line with the fact that there is no consensus on a precise model of financial crises and that they may result from very complex non linear processes.

The forecaster predicts the sequence $y_1, y_2, \ldots$ using a set of "experts". Experts are predictive models. They can be statistical models, an opinion on y_t using private sources of information or a black box of unknown computational power (neural network prediction for example). Each expert $j = 1, \ldots, N \in \mathcal{E}$ makes the prediction $f_{j,t}$ based only on information available until date t-1. Of course the quality of our optimal forecast will be dependent on the quality of our set of experts. The methodology of *online learning* is extremely flexible and general as any forecasting model can be used to contribute to the optimal forecast. On the one hand, of course there is no magic, if all forecasting models are bad, the optimal forecast will also be bad. If we put "garbage in", we will get "garbage out". On the other hand, we are guaranteed to do at least as well as the best forecasting model asymptotically: if *any model* provides excellent out-of-sample forecasts, it will be picked with a weight of one by our algorithm[20]. Hence we are sure to outperform the literature and central banks' forecasting models asymptotically.

To combine experts' advice, the forecaster chooses a sequential aggregation rule $\mathcal{S}$ which consists in picking a time-varying weight vector $(p_{1,t}, \ldots, p_{N,t}) \in \mathcal{P}$. The forecaster's outcome is the

[20]We emphasize this could be a bayesian averaging model, any other machine learning model, subjective judgement, etc...

linear combination of experts' advice :

$$\hat{y}_t = \sum_{j=0}^{N} p_{j,t} f_{j,t}$$

After having computed $\hat{y}_t$ (based on information available until t-1), the forecaster and each expert incur a loss defined by a non-negative loss function : $\ell : \mathcal{D} \times \mathcal{Y}$. This framework is summarized in Algorithm 1 in Online Appendix D.

How do we measure the sequential aggregation rule's performance ? If the sequence $y_1, y_2, \ldots$ were the realisation of a stationary stochastic process, it would be possible to estimate the performance of a prediction strategy by measuring the difference between predicted value and true outcome. But we do not have any idea about the generating process of the observations. However, one possibility is to compare the forecaster's strategy with the best expert advice. Let's define the difference between the forecaster's loss and the loss of a given expert, cumulated over time:

$$R_{j,T} = \sum_{t=1}^{T} (\ell(\hat{y}_t, y_t) - \ell(f_{j,t}, y_t)) = \hat{L_T} - L_{j,T}$$

where $\hat{L_T} = \sum_{t=1}^{T} \ell(\hat{y}_t, y_t)$ denotes the forecaster's cumulative loss and $L_{j,T} = \sum_{t=1}^{T} \ell(f_{j,t}, y_t)$ is the cumulative loss of the expert j.

The *regret* of a sequential aggregation rule $\mathcal{S}$ is given by :

$$R(\mathcal{S}) = \hat{L}_T(\mathcal{S}) - \inf_{q \in \mathcal{P}} L_T(q)$$

where $\inf_{q \in \mathcal{P}} L_T(q) = \inf_{q \in \mathcal{P}} \sum_{t=1}^{T} \ell(\sum_{j=0}^{N} q_{j,t} f_{j,t}, y_t)$ is the cumulative loss of the best combination of experts (known *ex post*).

This difference is called "regret" since it measures how much the forecaster regrets not having followed the advice of this particular combination of experts. The regret is a way of measuring the performance of a forecaster's strategy by comparing the forecaster's predictions (based on information at date t-1) with the best prediction which could have been done had she followed a certain combination of experts based on realised value at date t.

Knowing that $\hat{y_t} = \sum_{j=0}^{N} p_{j,t} f_{j,t}$, the regret can be written as :

$$R(\mathcal{S}) = \sum_{t=1}^{T} \ell(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t) - \inf_{q\in\mathcal{P}} \sum_{t=1}^{T} \ell(\sum_{j=1}^{N} q_{j,t} f_{j,t}, y_t)$$

Minimizing the regret is for the forecaster a robustness requirement. When the regret is close to 0, it ensures that forecaster's strategy (determined at date t-1) is close to the best combination of experts, which is known at the end of the round (at date t). To get a robust aggregation rule, the forecaster wants, in addition to having the smallest bound possible for the regret, to obtain a "vanishing per-round regret" so that when T goes to infinity the superior limit of the regret taken over all possible observation and prediction sequences goes to zero:

$$\lim_{T\to\infty} \sup\left\{\frac{R(\mathcal{S})}{T}\right\} \leq 0$$

In this case, the forecaster's cumulative loss will converge to the loss of the best linear combination of experts known *ex-post*. This approach can be described as "meta-statistical" since the aim is to find the best sequential linear combination of experts. Indeed, the following decomposition:

$$\hat{L}_T(\mathcal{S}) = \inf_{q\in\mathcal{P}} L_T(q) + R(\mathcal{S})$$

indicates that the forecaster's cumulative loss is the sum of an "estimation error", given by the cumulative loss of the best linear combination of experts (known *ex post*), and by the regret which measures the difficulty to approach *ex ante* the best combination of experts ("approximation error")[21].

[21] The bound of the regret guarantees that forecasters performance will compete with the performance of the best convex combination of experts when T goes to ∞. Note that this combination of experts is fixed over time whereas forecasters strategy includes time-varying weights. The forecaster's strategy is often worse than the performance of the best convex combination of experts since the best convex combination is known *ex post*, but it is not a theoretical necessity. With time-varying weights, an excellent online strategy could be able to beat the best (fixed) convex combination of experts.

Whereas this approach is very popular in statistics, most econometric research uses a ”batch” framework, where one starts from estimating a model on a complete sample. For model averaging problems, one of the most popular ”batch” methodologies in econometrics is the Bayesian Model Averaging (BMA) framework which uses Bayesian decision theory. There is a link between Bayesian decision theory and the theory of sequential predictions[22]. For a specific loss function based on a specific aggregation strategy, Cesa-Bianchi and Lugosi (2006) show that the on-line learning weights approximate the posterior distribution of a simple stochastic generative model. In this situation, the online approach maps into a case where the Bayes decisions are robust in a strong sense because their performance can be bounded not only in expectation with respect to the random draw of the sequence but also for each individual sequence. However, the online learning approach differs from the BMA approach in a fundamental way. In the BMA framework, the learning rate is always equal to 1, which makes this framework non-robust to some misspecification issues. For instance, Grunwald and van Ommen (2014) show that Bayesian inference can be inconsistent in simple linear regression problems when the data are heteroskedastic. In this set-up, regularity conditions for BMA consistency established by De Blasi and Walker (2013) are violated. As a consequence, as sample size increases, the posterior puts its mass on worse and worse models of ever higher dimensions. A natural solution is to add a learning rate in a sequential setting (Vovk (1990); McCallester (2001); Barron and Cover (1991); Walker and Hjort (2001); Zhang (2006)). We note that since online learning can be seen as a ”meta-statistical approach” (or a ”meta-algorithmic approach”), it can incorporate Bayesian analysis and make it compete with the best combination of models.

3.1 Online learning with delayed feedback

Our exercise does not fully correspond to the standard framework of sequential predictions. In the standard framework previously described, the forecaster knows the true observation y_t at the end of period t. After that, she incurs a loss and can update her weights. In our case, this assumption is not valid anymore. Indeed, the pre-crisis period is an *ex-post* definition. *After* a

[22] We are grateful to Christian Julliard for his insights on this topic.

crisis occurs, the 12 quarters before the beginning of the crisis is defined as a pre-crisis period. As a consequence, at the end of period t, the forecaster still does not know whether t, $t-1$,..., $t-12$ were a pre-crisis or not : the feedback of the forecaster is delayed. We therefore develop the online learning with delayed feedback framework, where the feedback that concerns the decision at time t is received at the end of the period $t+\tau_t$. We build on the work of Weinberger and Ordentlich (2002) and of Joulani et al. (2013). In this framework, τ_t may have different forms. It could vary over time, be an i.i.d. sequence independent of the past predictions of the forecaster or depend on $\hat{y}_t$. In our case, τ is a constant which is equal to 12. We define $R'(\mathcal{S})$ as the regret of the sequential aggregation rule $\mathcal{S}$ in a delayed setting. Following Weinberger and Ordentlich (2002) it is straightforward that:

$$R'_{T,\tau}(\mathcal{S}) \leq R_T(\mathcal{S}) \times O(\tau)$$

Introducing a delayed feedback increases the bound of the regret - the approximation error - but does not violate our robustness requirement. We implement Algorithm 2 (see Online Appendix D).

3.2 Choosing a loss function

The loss function can take different forms. The only constraint is that it should be convex and bounded for minimizing the regret. In our case, we are seeking to predict a binary outcome so there is no issue. We use a squared loss function $\ell(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$ (but could also use an absolute loss function $\ell(\hat{y}_t, y_t) = |\hat{y}_t - y_t)|$). Which of them is more appropriate for a given problem is an empirical question though the squared loss function tends to have better out-of-sample performance.

3.3 Selecting aggregation rules

We only select robust aggregation rules, which compete with the best combination of experts *ex post*. We consider four aggregation rules with different properties to investigate the robustness of our results: the Exponentially Weighted Average aggregation rule (EWA), the Online-Gradient Descent aggregation rule (OGD), the Ridge aggregation rule (R) and the Fixed Share aggregation rule (FS). We discuss in the main text the characteristics of the EWA in order to provide some intuition but relegate the detailed discussion of the other rules to Appendix D.

3.3.1 Exponentially weighted average aggregation rule

At first, we consider convex aggregation rules. Convex aggregation rules combine experts' predictions with a time-varying vector $p_t = (p_{1,t}, ..., p_{N,t})$ in a simplex $\mathcal{P}$ of $\mathbb{R}^N$:

$$\forall j \in \{1, ..., N\}, p_{j,t} \geq 0 \text{ and } \sum_{k=1}^{N} p_{k,t} = 1$$

We use the exponentially weighted average (EWA) aggregation rule as it presents key advantages. First, the weights are computable in a simple incremental way. Second, the forecaster's predicted probability only depends on the past performance of the experts and not on her past prediction.

We use the gradient-based version of the EWA aggregation rule $\mathcal{E}_\eta^{grad}$ where weights are defined by :

$$p_{j,t} = \frac{exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{j,s}})}{\sum_{k=1}^{N} exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{k,s}})}$$

where $\tilde{L_{j,s}} = \nabla \ell(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s) \cdot f_{j,s}, \nabla$ is the gradient operator and η_t is the learning rate, the speed at which weights are updated. Weights are easy to interpret. If expert j's advice $f_{j,s}$ points in the direction of the largest increase of the loss function, i.e. if the inner products $\nabla \ell(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s) \cdot f_{j,s}$ has been large in the past, the weight assigned to expert j will be small. We implement Algorithm 3 described in Online Appendix D.

The strategy $\mathcal{E}_\eta^{grad}$ is robust in the sense that it competes with the best convex combination of experts. Indeed, the regret of the gradient-based exponentially weighted average forecaster can be bounded using only assumptions on the loss function and on the decision space (Cesa-Bianchi and Lugosi (2006)).

Theorem 1 (Vovk, 1990 and Stoltz, 2012) . Assume the loss function ℓ is convex, that its gradient $\nabla\ell$ exists and its inner product $\tilde{\ell}$ is included in $[m, M]$, the regret of the gradient-based exponentially weighted average forecaster satisfies for $\eta > 0$:

$$\sup(R_T(\mathcal{E}_\eta^{grad})) \;\leq \frac{\ln N}{\eta} + \eta\frac{(M-m)^2}{8}T$$

In our framework, $\mathcal{D} = [0,1]$,$\mathcal{Y} = 0,1$ and $\ell(x,y) = (x-y)^2)$. The inner product $\nabla\ell(\sum_{k=1}^N p_{k,s} f_{k,s}, y_s)\cdot f_{j,s}$ is therefore bounded and the required assumptions to bound the regret are satisfied.

Note that the bound of the regret depends on three parameters, two exogeneous (N and T) and one endogenous (η). An interesting property of the theorem is that the bound does not depend linearly on the number of experts, but on $\ln(N)$. A large number of experts will not drastically increase the difference between the forecaster's cumulative loss and the cumulative loss of the best combination of experts.

To satisfy our robustness requirement, the regret bound can be optimized over η :

$$\min_{\eta>0}\left\{\frac{\ln N}{\eta} + \eta\frac{(M-m)^2}{8}T\right\} = (M-m)\sqrt{\frac{T}{2}\ln N}$$

for the theoretical optimal choice of $\eta^* = \frac{1}{M-m}\sqrt{8\ln N T}$. Then the strategy $\mathcal{E}_\eta^{grad}$ satisfies our robustness requirement:

$$\sup\{R_T(\mathcal{E}_\eta^{grad}) = o(T)$$

The disadvantage of this optimal tuning of the learning rate relies on the fact that it requires a knowledge of the horizon T in advance. This is why the litterature traditionnaly let the parameter η depend on the round number t [23]. However, the theoretical values of η_t suggested by the litetrature are generally too smal and lead to poor forecasting performance, as illustrated by Mallet et al. (2009) and Devaine et al. (2013) . This is why decide to adopt the empirical strategy proposed **?**to calibrate η_t by minimizing the cumulative loss function (see Apendix D for more

[23]for a discussion of the calibration of η, especially with the so-called doubling trick or the theoretical calibration of adaptative learning, see **?** and **?**. They show that these calibrations can achieve the same regret bounds as if T was known in advance

details) [24]:

$$\eta_t \in \arg\min_{\eta>0} \hat{L}_{t-1}(\mathcal{E}_\eta)$$

3.3.2 Other aggregation rules

We present in Appendix D three other aggregation rules: the Fixed Share aggregation rule (FS), which builds directly on the EWA; the Online-Gradient Descent aggregation rule (OGD) and the Ridge aggregation rule (R) and explain how to implement these aggregation rules in an environment with delayed feedback. These rules offer some diversity in the way the aggregation is performed and the speed at which the learning parameter is evolving. For the Ridge, the aggregation weights are not bounded between zero and one. For the EWA, the FS and the Ridge, the learning parameter is optimised empirically. For the OGD, the learning rate is theoretically calibrated (see Appendix D) . Due to the delayed feedback and the relatively small size of the sample, the relative performance of the different rules is an empirical question.

3.3.3 Online learning with small samples

As opposed to a traditionnal online exercice, our framework has two specificities. First, the forecaster has to take a delayed feedback into account which inevitably reduces its performance. Second, due to the sample size, the online sample is necessarily small and only contains one pre-crisis period. Indeed, forecaster's performance relies on its capacity to discard bad experts on a sufficient amount time to predict the pre-crisis period.This is why the computation of the initial weight vector is critical : with a delayed back, starting with a wrong information would be detrimental for forecaster's performance. To limit this risk, we decide to implement a non-initial weight vector (see Appendix D for theoretical guarantees with a non-uniform initial weight vector)

[24] Note that we compute every aggregation rule using the opera package created by P.Gaillard and Y.Goude, which is also used to forecast electricity consumption. See : https://cran.r-project.org/web/packages/opera/index.html

3.4 Designing experts

To design the experts, the forecaster faces the following arbitrage. On the one hand, it is critical to include a sufficient number of experts to get the maximum amount of information, in order to reduce the approximation error. On the other hand, the regret increases with the log of the number of experts. This tradeoff is even more accurant with a small sample. Too many experts could increase the number of information the forecaster has to deal decreasing his ability to predict on time the next pre-crisis period.

With this tradeoff in mind, we pick different sets of experts in Section 4: some are "off-the shelf" experts used in the literature and in central banks to predict financial crises, others are bayesian averaging models and machine learning models such as random forests. The beauty of our approach is that we can include *any* type of experts and therefore be very œcumenical in terms of methodology. Experts' parameters are first estimated on the batch sample. Then they are estimated recursively on the online sample, *using only the information available at t-12.* As a consequence, during the first dozen observations of the online sample, experts' parameters are not updated.

4 An œcumenical approach to crisis prediction

We include in our set of experts several models used by academics and by central banks in their effort to construct a set of early warning indicators for macro prudential policies: Dynamic Probit Models, Panel logit models, bayesian model averaging. Some of these models were summarised by the Macroprudential Research Network of the ECB (Alessi et al. (2015)). To those, we add models from the econometrics and statistical and machine learning literature: General Additive Model (GAM), random forests, Support Vector Machine (SVM). We then add several Logits with elastic net penalties[25] as these models have been found to be particularly well suited for out-of-sample forecasts. We design those by grouping variables by themes: a subset of the logits describe the real economy, another subset the housing market, another the credit market etc... This is in or-

[25]This is a regularized regression method that combines linearly the penalties of the LASSO and the Ridge with certain weights.

der to ease the economic interpretation of our results. Note that our models incorporate various horizons of changes for the variables so that inflexion points can be captured. In a small number of cases, when we use models of the literature we could not include one variable of the model as it was not publicly available. All the models have been re-estimated with our variables on our sample. Some models are estimated on a panel, others are estimated country by country. Therefore our experts incorporate information from the *entire set of countries* and account for potential interactions and global effects. We note that we could consider many more variables and models. We emphasise that *our contribution is not about finding the best possible experts*. Our contribution is to optimally aggregate those experts. If an excellent expert can be found our method would give it a high weight and by construction we would do at least as well as this excellent expert. We could also extend the country sample. The methodology is flexible enough to incorporate all these improvements. We end up with 26 experts that we briefly describe below. Some of these models are generic in the sense that the specification is exactly the same for all countries. Others use country specific variables, which we select using the Area under the Receiving Operator Curve (AUROC) criteria. Our eclectic choice of models will allow us to see whether a-theoretical models such as random forests dominate or not models based on economic mechanisms (such as credit growth) to produce out-of-sample forecasts. We refer the reader to Appendix B and C for a detailed description of these models and for all the precise specifications.

Our first set of experts are taken from the economic literature on macroprudential policies on panel data (see Appendix B and C): **Expert P1**. Dynamic Probit Model: variables selected with a country-specific AUROC on the batch sample panel (Antunes et al. (2014)); **Expert P2**. Panel logit fixed effect: variables selected with a country-specific PCA Analysis on the batch sample panel (Bush et al. (2015)); **Expert P3**: Panel logit fixed effect. We follow the literature for the exact specification (Behn et al. (2013)); **Expert BMA**: Bayesian Model Averaging. Variables selected with a country-specific AUROC on the batch sample panel. Our second set of experts come from the Machine Learning and statistical literature (see Appendix B and C): **Expert GAM**: General Additive Model; **Expert RF**: Random Forest; **Expert SVM**: Support Vector Machine. Our third set of experts are constructed using Logits with elastic-net penalty[26]. All the Logits include each variable in level as well as the 1-year change and the 2-year change. Quantities are expressed as a fraction of GDP. These Logits are organised around sets of variables belonging to a specific sector of the economy. For example we construct a Logit credit (**Expert Lcr**) using Total credit to non-financial sector; Banking Credit to non-financial sector; Total Credit to Households; Total Credit to non-financial corporations. The Logit Foreign (**Expert Lfor**) will have Cross Border Flows; Real Effective Exchange Rate; Dollar Effective Exchange Rate; Current Account; Terms of Trade. We have a valuation Logit, two real economy Logits, a housing Logit, a monetary Logit, etc... We also allow for combinations. For a detailed description of these 19 additional models please see Appendix B and C. We now have experts of all stripes and shapes including some models with common components, global variables, Bayesian averaging and random forests. Some are estimated on panel data, others are country specific. Our models contain most of the variables that have been shown to be important in the literature and that a well-read international economist would have considered since the beginning of the 20th century: asset valuations, credit; household debt; house prices, financial condition indices, current accounts, real exchange rates, etc... Our œcumenical approach can accommodate many more. Our only restriction is data availability. For example it would be desirable to test the information content of variables based on individual bank's balance sheets but the timing of the first crisis and the twelve quarter lags means that in practice those variables cannot be incorporated in the analysis.

5 Results

We focus on countries such as France, the United Kingdom, Germany and Italy which experienced a systemic crisis at the beginning of our sample in the 1980s or 1990s. This allows our algorithm to learn about systemic crises and enables it to predict out-of-sample thereafter. Spain and the US do not experience any systemic crisis at the beginning of the sample. We present

[26]First introduced by Zou and Hastie (2005), the good performance of elastic-net penalty compared to other regularization methods has been confirmed in various applications Mol et al. (2009); Mol et al. (2009); Destrero et al. (2009). They prevent overfitting which comes from model complexity by shrinking coefficient estimates.

a series of results using quasi-real time data (i.e. historical data which may have been revised). For France and the UK we also present results using real time data. We note that the timing of the systemic crises in all those countries is different not only in the 1980s or 1990s but also around 2008. They have commonalities but also country specific characteristics (this is why we symbolically wrote the section headings below in the national languages). Most of the literature focuses on in-sample results and attempts to predict crises (not pre-crises). We present results for *out-of-sample pre-crisis* prediction. We show a time series of our predicted probability of crisis as this has the advantage of being very transparent and of allowing us to assess straight away the usefulness of our predictive model as an early warning indicator. If the signal tends to be monotonically increasing before a crisis it is likely to be a useful early warning indicator, provided it does not have too many false positive. For each country we present in the main text our estimated probability of pre-crisis using the EWA aggregating rule. We show some additional results in Appendix. We also present results on the time-varying weights assigned by our aggregation rule on each model and the contribution of each expert to the prediction in order to gain some insights in the transmission mechanisms. Finally we report diagnostics regarding the fit of our model (mean squared errors and AUROCs) for the different aggregation rules, as well as loss functions for our EWA rule compared to individual experts. This allows us to compare the performance of our model to the performance of any of the experts (the literature and more) and to the uniform aggregation of the experts. Recall that the forecasters' performance can be decomposed in two elements: the approximation error, which reflects the difficulty to optimally combine experts' advice and the estimation error. The approximation error is measured by the difference between the aggregation rule and the best combination of predictors (known *ex post*). The estimation error reflects the performance of the different experts. It can be illustrated by computing the uniform aggregation rule (where each expert has the same weight over time) or the performance of the best expert.

6 Les crises systémiques en France

There are two *systemic crises* in France during our sample period from1985Q1 to 2019Q3. The first one is from 1991 Q2 till 1995 Q1 and the second one from 2008 Q1 to 2009 Q4. This makes France a good case to apply our methodology as both batch and online samples are relatively long. There are also two *residual events* which correspond to the burst of the IT bubble in 2002 Q3 till 2003 Q2 and the euro area sovereign debt crisis from 2011 Q1 till 2013 Q4. The 1991Q2-1995 Q1 French systemic crisis, on which our algorithm learns, was linked to real estate. As described in Lo Duca et al. (2017) on which we draw, France experienced a period of high GDP growth and deregulation after 1987, which led to a sizeable increase in residential and commercial real estate prices. Increasing oil prices and a deteriorating international economy brought a severe slowdown after 1990 Q2 and a plunge in real estate prices. The French banks saw an increase in non-performing loans and a fall in value of real estate property assets in portfolios. They reduced their supply of loans to property developers and sellers. The large decline in commercial real estate prices, used as collateral had a negative impact on the financial position of borrowers and led to some defaults. The economy was then damaged by the European Exchange Rate Mechanism crisis of 1992 and the fragility of the banking sector with the near bankruptcy of the Crédit Lyonnais (due to the real estate market downturn and excessive risk taking). The trough of the recession was reached in 1993 Q1.

6.1 Out-of-sample prediction of crises: France. Quasi real time data.

Figure1 illustrates the timing of pre-crises and crises in France on the period during which we forecast out-of-sample which starts in 2001Q3. We aim at forecasting the systemic pre-crisis period (2005Q1 to 2008Q1). We estimate the expert models on the batch sample 1987Q3-2001Q2 (1987Q3 is the earliest possible date we can start because of data availability). We present results for out-of-sample pre-crisis prediction for 2001Q3 to 2019Q3. This includes the period of the second systemic crisis (2008 Q1 to 2009 Q4)[27]. That systemic crisis followed the collapse of Lehman

[27]For the US the systemic crisis is dated 2007Q3-2009Q4.

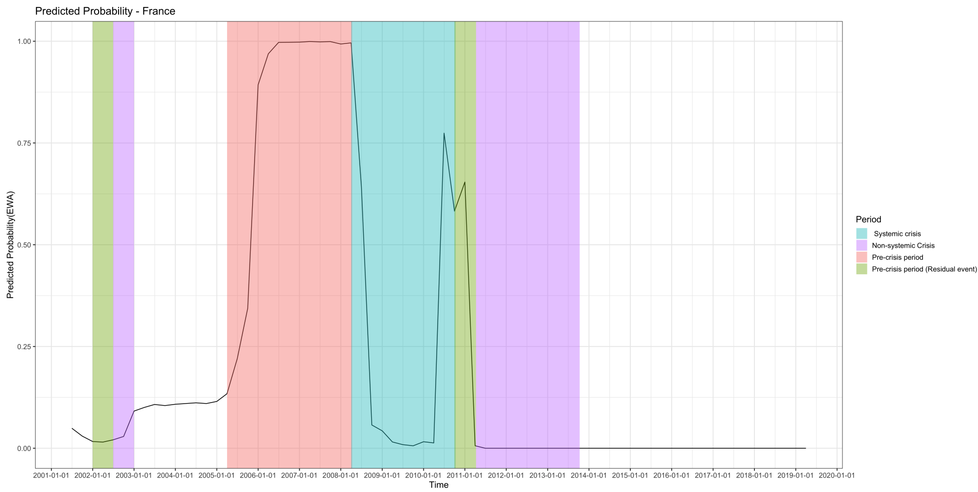

Figure 1: France: Predicted probability of a crisis - EWA aggregation rule

Brothers after an era of growing GDP, falling unemployment, excessive credit growth and booming real estate prices. As described in Lo Duca et al. (2017), the spillovers from the US financial crisis triggered a recession with a fall in investment and consumption, as private agents tried to deleverage in front of a deteriorating and highly uncertain economic environment together with a collapse of international trade. France entered a recession in Q3 2008, for four quarters. Unemployment rate rose from 7.5% to 9.5%. There was a 10% decline in residential real estate prices after a boom in the 1995-2007 period. Policy interventions included a restructuring and capital injection into Dexia, a Franco Belgium bank, a French bank guarantee scheme (November 2008-2009), a recapitalisation scheme (December 2008 and March 2009) and a merger and capital injection into Banque Populaire-Caisse d'Epargne (May 2009). In Q3 2009, GDP growth turned positive again and unemployment started to fall. This out-of-sample forecasting period also includes the euro area sovereign debt crisis (2011 Q1 till 2013 Q4), which is not classified as a systemic crisis in France but as a residual event. That period however saw spillovers from the crises in some euro area countries both in terms of real activity and via exposure of French banks to the periphery.

Pre-crisis probability.

Figure 1 presents the results for the EWA aggregation rule. The entire period is out-of-sample and we aim at forecasting the systemic pre-crisis period (2005Q1 to 2008Q1). It shows that the probability of being in a systemic pre- crisis before 2004 Q4 was low with a sharp increase starting in 2005 Q1. Since the probability increases over time and increases steeply, the model provides a very good early warning system. The 12 quarter ahead crisis probability reaches 1 and remains there till 2008 Q1. The model performs very well as the crisis starts in 2008 Q1 and accordingly the probability starts dropping -we are predicting the pre-crisis *not* the crisis. After 2008 Q4, the probability of a systemic pre- crisis remains very close to zero until 2010 Q1where the probability of crisis goes back up again. This corresponds more or less to the timing of the pre-crisis for the euro area crisis, which is classified as a "residual event" in our data base (from the point of view of the algorithm this is therefore a mistake). The probability goes back down to low levels at the end of the pre-euro crisis period and remains close to zero till the end of the sample. It seems therefore that the algorithm learns on the 1991 Q2 -1995 Q1 systemic crisis all that is necessary to be able to predict the 2008 crisis as early as 2005 Q1 (and it gives a smaller warning before the residual event of the euro area crisis). We show in Appendix the results for the FS, OGD and Ridge aggregation rules. The FS rule also gives a clear and rising signal in 2005 Q1 well before the 2008 systemic crisis. For the OGD aggregation the results are somewhat similar to the FS aggregation rule. The Ridge does not perform very well. This is possibly a consequence of our small sample: EWA type rules are more robust in that case (the gradient -based EWA is very reactive due to the weights computation). Three aggregation rules manage to predict the pre-crisis period for the 2008 systemic crisis (the Ridge predicts mostly the euro area crisis, which is not systemic). For all the aggregation rules there is a second probability spike, usually smaller, linked to the pre-euro area crisis period. One of the main difference across the different aggregation rules in terms of methodology is the way the learning rate is picked. For the EWA, the FS and the Ridge it is optimised upon empirically whereas for the OGD the theoretically calibrated value of the learning rate is used. This said, the results across the aggregation rules are often consistent (except for the Ridge). The EWA is the simplest rule and it appears to be the most robust when samples are small.

Diagnostics

Table 1 presents the Root Mean Squared Errors (RMSE) and Area under the Receiving Operator

Curve (AUROC) of our different aggregation rules and compares them to the best fixed convex combination of experts known *ex post* and to the uniform aggregation rule (equal weights on all experts). The ROC curve represents the ability of a binary classifier by plotting the true positive rate against the false positive rate for all thresholds. If the model made a perfect prediction the area under the curve (AUROC) would be equal to 1; if it were as bad as a coin flip, the AUROC would be 0.5. We note that the EWA, the FS and to a lesser extent the OGD RMSE are close to their theoretical asymptotic value of the best convex combination of experts (0.26, 0.31 and 0.33 respectively versus 0.28 for the best convex combination known *ex post*) despite the relatively small sample size. The EWA does even better as its weights are time varying whereas the best convex combination has fixed weights. The approximation error is therefore very small. Since the uniform aggregation performs relatively well, this also suggests that experts are good and that the estimation error is not too high. The EWA and FS aggregation rules have an AUROC remarkably close to 1, at 0.98 and 0.92. All aggregation rules do better than uniform weights except the Ridge which performs badly. Note that the prediction of the euro area crisis is counted as an error by the algorithm as this episode is not classified as a systemic crisis but as a residual event. We do not want to emphasize particular diagnostics but do report them to allow comparisons with the literature[28]. AUROCs are by no means a panacea to assess the performance of a model: an expert that would increase his probability of pre-crisis from 0 to 0.00001 at the right time would have an AUROC of 1 but would not be particularly useful. Similarly, RMSE give information on the average performance of an expert but is not necessarily highly correlated with the quality of the signal given. What we do want to emphasize is that our out-of-sample graphs of the time-varying probability of systemic crises provide a transparent way of assessing the performance of our methodology and of our experts. A sizable, monotonic increase at the right time is a reliable and clear pre- crisis signal, which is valuable for macroprudential authorities. At which point those authorities may want to respond to that increased probability of crisis is a matter of judgement, which cannot be mechanical and will depend on external parameters, which we

[28]For comparison purposes, Schularick and Taylor (2012) report an out-of-sample AUROC of 0.646. Richter et al. (2021) report AUROCs between 0.70 on their whole sample and 0.86 on a reduced sample. Bluwstein et al. (2020) reports AUROCs of between 0.77 and 0.87 depending on the sample. These papers make use of the macro history dataset of Jorda Schularick and Taylor. None of them report the time series of the estimated probability as we do.

Online Aggregation Rule	RMSE	AUROC
EWA	0.26	0.98
FS	0.31	0.92
OGD	0.33	0.85
Ridge	0.52	0.70
Best fixed convex combination	0.28	0.97
Uniform	0.36	0.79

Table 1: RMSE and AUROC of different aggregation rules. France

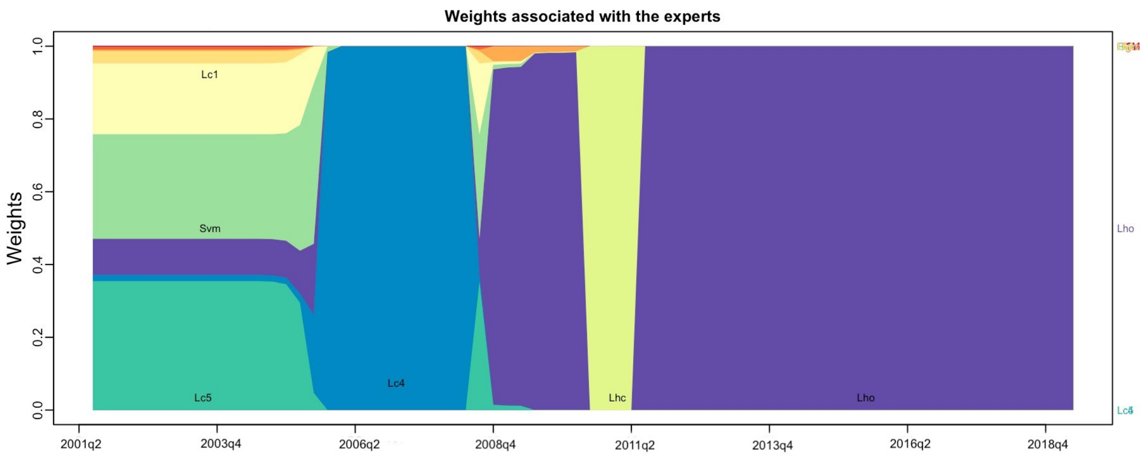

Figure 2: France: Weights. Quasi-real time. EWA aggregation rule.

do not attempt to model. We emphasize further that, by design, our methodology ensures that asymptotically we always do at least as well as the best expert (including any kind of combination or transform of experts). In practice, when we compare average losses or cumulative losses, we show that in most cases our aggregation rule outperforms *any* of our individual experts as well as the uniform aggregation of our experts (see **Figure 4** and **Figure 5**).

Dominant experts and their roles.

Our online learning methodology is not a black box. It allows us to track which models get an endogenously higher weight in the forecast at a given point in time and which ones give the crisis signal. Interestingly some models dominate the forecast. **Figure 2** shows the time varying weights associated to each of our experts for the EWA aggregation rule and **Figure 3** presents the contribution of the experts to the forecast (the dashed line is the pre-crisis period we are seeking

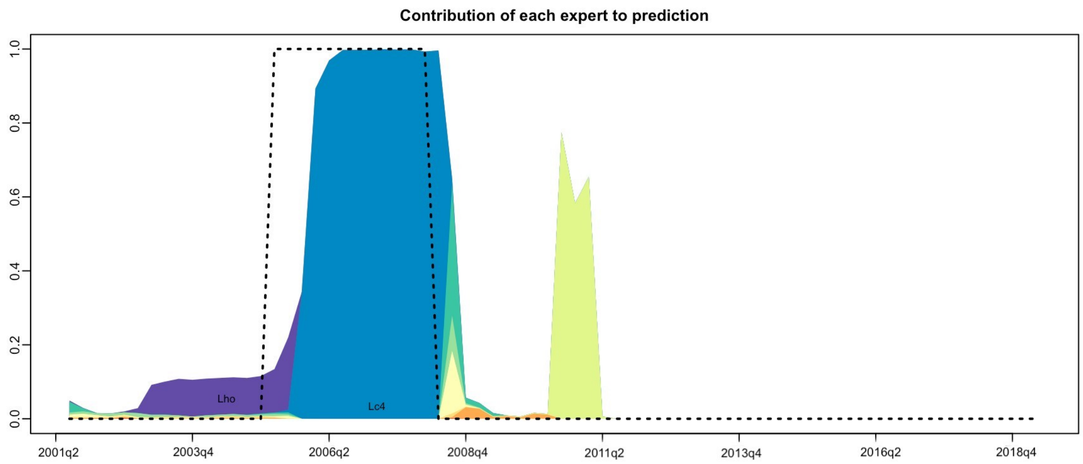

Figure 3: France: Experts. Quasi-real time. Contribution to forecast. EWA aggregation.

to predict). The optimal forecast for the EWA rule puts some positive weights on several models. Among those, in **Figure 3**, we see that the ones giving the crisis signal are **Lho**[29] and **Lc4** which is the one really spiking; **Lc4** is a logit elastic net mainly on housing, credit and investment[30]. **Lc5**[31] and **Lhc**[32] are also informative. We can look specifically at the estimated coefficients of each variable for each expert. We find that for the best expert **Lc4**, real estate variables such as the rent price index (1y and 2y), the price to income and fluctuations in quantities of credit (bank credit to the private non-financial sector and total credit to the private non-financial sector) are particularly informative to predict systemic crises.

Figure 11 in the Appendix shows the time varying weights associated to each of our experts for the FS aggregation rule and **Figure 12** presents the contribution of each expert to the forecast. Interestingly **Lc4** also plays the central role and gives the pre-crisis signal. According to the OGD aggregation rule, it is also **Lc4** and **Lc5** which give the strongest signal for the systemic crisis. So the results are very consistent across three aggregation rules (EWA, FS and OGD) for the prediction of the pre-systemic crisis period (the Ridge is the outlier in terms of performance).

[29] **Lho**'s variables are: Price-to-rent, price-to-income, real estate price, rent price index.

[30] **Lc4**'s variables are: Real estate price, GDP, Total Credit to Households, Rent Price Index, Loans, Banking Credit to private non-financial sector, Price-to-income, Investment, Share price index, Equity Holdings.

[31] **Lc5**'s variables are Price-to-rent, Short-term interest rate, Terms of Trade, Housing 2, Total Credit to Households, Banking Credit to private non-financial sector, Total Credit to private non-financial corporations, Rent Price index, Investment, Share Price index, equity Holdings.

[32] **Lhc**'s variables are Price-to-rent, price-to-income, real estate price, rent price index, Total credit to non-financial sector, Banking credit to nonfinancial sector, total credit to households, total credit to non-financial corporation.

For the FS rule, the euro area pre-crisis peak in crisis probability is due to **Lhc**. Similar experts are picked by the OGD and the Ridge aggregation rules for the pre-euro area peak.

Our results are not totally surprising. Real estate markets are seen as having played an important role in the Great Financial Crisis in the US. They seem to have done so as well in France and this is consistent with the historical narrative of the French crisis.[33] In the case of France, we see however that banking credit and total credit to non-financial corporations are also very informative, as well as price-to-rent and price-to-income. For the euro area pre-crisis (residual event), financial market stress indicators, global factor in asset prices, interest rates and international flows and exchange rate variables seem to play a bigger role. This is true across all the aggregation rules we considered. These variables were picked *ex ante* out-of-sample by the algorithm and they make economic sense given the *ex post* known narrative on the French crisis. Of course, no causality can be established.

In **Figure 4** we show the average loss suffered by the experts and in **Figure 5** the cumulative loss of all the experts over time. These two graphs show that the EWA aggregation rule *does better than all the experts*. It also confirms the good performance of the expert **Lc4** in the case of France. The excellent performance of our aggregation rule (the cumulative loss of the forecaster is always strictly smaller than the ones of the experts in **Figure 5**) could surprise given the relative short sample and low number of crises. We note that the algorithm learns both from the ones (pre-crisis) and the zeroes which allows for accurate predictions by removing quickly the experts over-predicting pre-crises thanks to the learning rate. In **Figure 17** in Appendix we show the predicted probability of some of the experts and of the EWA rule. Some experts predict crises with a high probability but get the timing wrong (logit housing) and are then discounted while others (SVM) do no predict steep increases in probability.

Additional Robustness Checks

On top of checking our results are consistent across different aggregation rules, we also re-estimated our EWA aggregation using an 8 quarter pre-crisis period as opposed to a 12-quarter

[33]We note that the timing of the systemic crisis is not exactly the same in France (2008Q1-2009Q4) and in the US (2007Q3-2009Q4); we also note that the euro area crisis affected France subsequently.

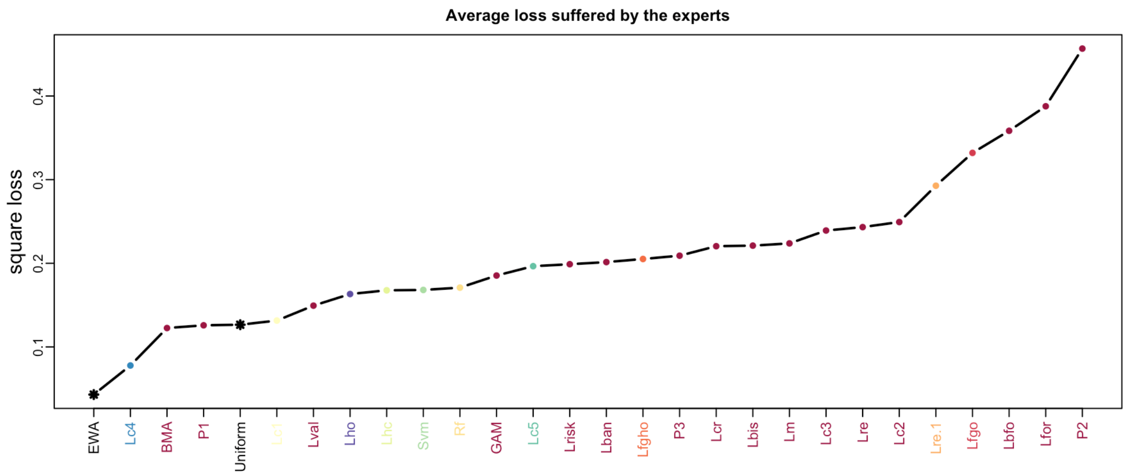

Figure 4: France: Experts. Quasi-real time. Average Loss. EWA aggregation.

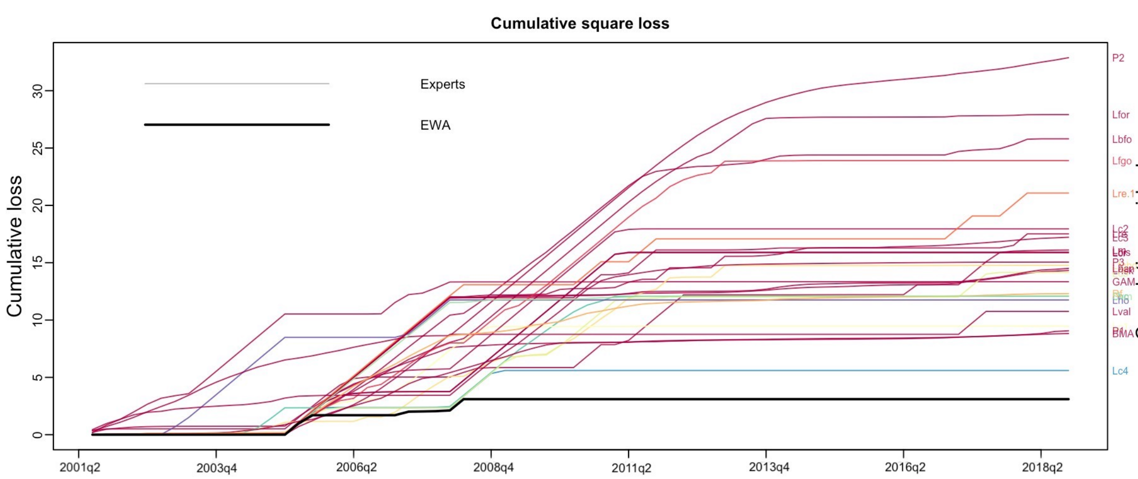

Figure 5: France: Experts. Quasi-real time. Cumulative Loss. EWA aggregation.

period. The two models picked are **Lc4** and **Lbfo** and the model giving the signal is **Lbfo**[34] (see Appendix). So our aggregation method is able to give a very clear signal of the systemic crisis in 2008 both 3 year (mostly with housing variables and credit volume) and 2 year ahead (mostly with international variables and risk taking variables). Since the importance of different actors (banks versus hedge funds versus households for example) vary along the financial cycle, it is not surprising that different models and variables show forecasting power at different horizons. For example, price variables have some forecasting power at short horizons while quantity variables are more important at longer horizons (Coimbra et al. (2021)).

We also revisit the problem of overfitting, which may stem from model complexity or the size of the sample. By design of the theory of sequential predictions the only source of overfitting could be the experts -but not the aggregation rule. However, many experts (all the elastic net logits) are already regularized. Nevertheless, we present a placebo test as an ultimate check for the presence of overfitting. We randomly assigned a crisis in the batch sample and another one in the online sample. The algorithm is unable to predict the crisis. These results are presented in the Appendix in **Figure 17**. We find that experts are unable to predict the crisis. There is no evidence of overfitting.

6.2 Out-of-sample prediction of crises: France. Real time data.

We test our methodology for real time out-of-sample prediction using vintage data for France and the EWA aggregation rule. Unfortunately, we have been able to obtain vintage data only for a subset of our variables. In particular we are missing long enough series for GDP, credit and housing market related variables. Fortunately however we can rely on Cross Border Capital vintage data series for the whole panel of countries (liquidity indices built on real time flow data as well as risk taking indices built on asset price data)[35]. We also use exchange rates and asset price data which are not revised, and specifically for France M3 and inflation data which are not revised. We go from 244 variables down to 122 variables. We reestimate all our experts on

[34] Share price index, Equity Holdings, Risk Appetite, Total Liquidity Index, Crossborder flows, Real effective exchange rate, dollar effective exchange rate, current account, Terms of Trade.

[35] For a description of the Cross Border Capital Data see Online Appendix.

the 1987Q3-2001Q3 sample using *only vintage data* and we use also only vintage data for the out-of-sample exercise[36]. Despite the strong data limitations, we still get good results for the predictability of the systemic crisis as shown in **Figure 6**. The probability of pre-crisis goes up in 2005 Q2 (1 quarter later than in quasi real time) and remains high until the systemic crisis unfolds. There is a spike as before for the euro area crisis but it occurs a bit later. The main difference has to do with the existence of spikes in 2018 which were not there when we used the quasi-real time data with the EWA rule (there were small spikes with the other rules). So the real-time estimates, which are based on fewer series seem noisier and more prone to false positive. It is hard to make a meaningful comparison of the weights of the models with the quasi real time results as the variables used in the models are now very different due to data restrictions. There are two models which are picked by the EWA aggregation rule: the statistical expert **GAM**[37] and a new **Lc5** expert[38]. It is the **GAM** expert which gives the signal of a pre- systemic crisis before 2008. In the absence of any credit data and housing data which were very important in our quasi-real time exercise, it is the interest rate, exchange rate and capital flow data which trigger the alarm. **Lc5**, which measures mostly financial stress and asset price variables is responsible for the subsequent spikes. Those are false positive. More than the real time versus quasi-real time dimension it seems to us plausible that it is the lack of data availability in terms of credit, real variables, and housing market vintage variables which are responsible for the deterioration in forecasting ability. The estimation error has gone up as an increase in the RMSE of the experts aggregated uniformly shows (the experts do not perform as well as in the quasi-real time case) but the RMSE and the AUROC are still very good (see Table 2) when compared to the Best convex combination of experts (based on *ex post* information) or on a uniform aggregation. On-line learning methodologies have been developed precisely to do real time forecasts.

[36]Our real time out-of-sample exercise is overly strict. Indeed, we even estimate our experts on the batch sample using vintage time series.

[37]**GAM**'s variables are Short-term interest rate 2y, Cross Border Flows 1y, Dollar effective exchange rate 2y.

[38]**Lc5**'s variables are Financial Condition Index, Domestic Sector Liquidity Stock, Private Sector Liquidity Stock, Equity Exposure Index, Total Liquidity Stock.

Online Aggregation Rule	RMSE	AUROC
EWA	0.36	0.84
Best convex combination	0.32	0.84
Uniform	0.40	0.54

Table 2: RMSE of different aggregation rules. France: **real time**

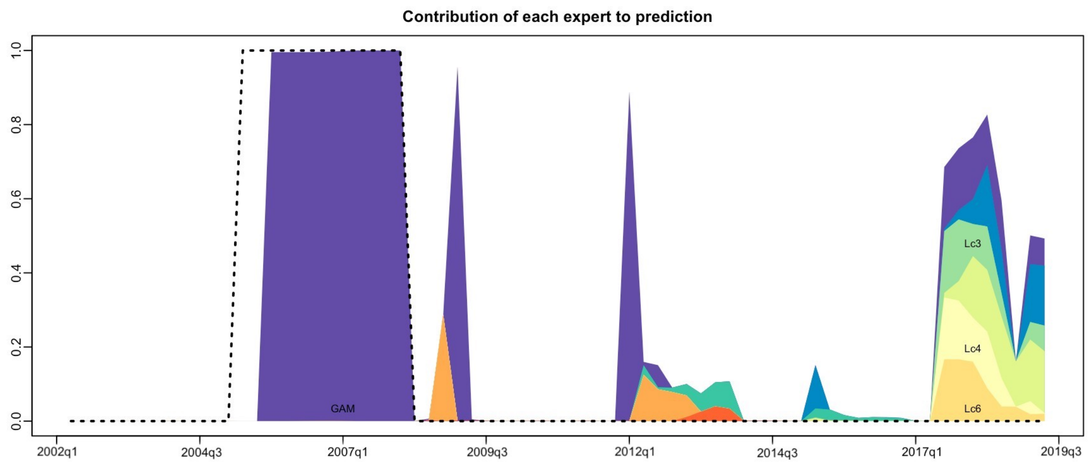

Figure 6: France: Experts. **Real time** - EWA

7 Systemic crises in the United Kingdom

We now turn to the UK. As for France we have a relatively long sample for batch and online estimations. For the UK, the crisis started in 2007 Q2 and ended in 2010 Q1, which is a different timing from France. The previous systemic crisis was from 1991 Q2 till 1994 Q2. As described in Lo Duca et al. (2017), that crisis was linked to excessive credit growth, high real estate prices and leverage. Rapid credit expansion took place in the 1980s (including in property-related assets). Even though some small institutions failed from June 1990 there was no reaction or concern from authorities until counterparties were unable to access their funds at the BCCI (Bank of Credit and Commerce International). The event generated panic and the people moved their money to larger institutions. The Exchange Rate Mechanism (ERM) forced the Bank of England to keep a high interest rate. This exacerbated the economic slowdown, accelerating the fall of property prices. The second systemic crisis 2007 Q2 till 2010 Q1 is predicted out-of-sample. The episode relates to the subprime crisis. The instability came from weaknesses within the financial system that developed during the global credit boom characterised by rapid balance sheet expansion.

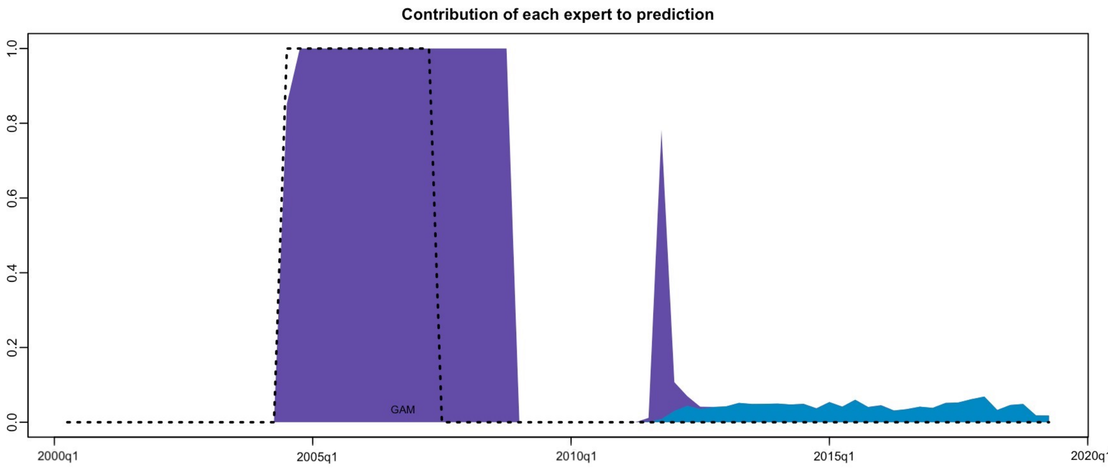

Figure 7: UK: Contribution of experts to forecasts. Quasi-real time - EWA

Too many assets whose liquidity and credit quality were uncertain were created and funding structures were risky and fragile. We present results for out-of-sample prediction for 2000Q2 to 2019Q3. Unlike France, there are no residual events in the data after the systemic crisis.

7.1 Out-of-sample prediction of crises: UK. Quasi real time data

Figure 7 presents the predicted probability of a pre-crisis in the UK for the EWA aggregation rule with the contribution of the experts to the forecast. The dashed line is the pre-crisis period we are seeking to predict. The probability of being in a pre-crisis in 2004 rises very quickly. The probability of a subsequent crisis is very low after the Great Financial crisis except for a peak in 2011. **Table 3** shows that the EWA rule performs well and just like in the case of France, it performs better than the other rules (unreported). Two experts are doing most of the work: the **GAM**[39] and the Logit risk **Lrisk**[40]. It is the **GAM** expert that gives the signal before the 2008 crisis. That expert combines information on the housing market and on long term interest rate. The second expert **Lrisk** reflects risk taking. The algorithm can also predict well the crisis two year ahead as shown in **Figure 18** in the Appendix. The model giving the signal in that case is **Lval**[41], which reflects valuations in different asset markets and risk taking. **Figure 19** and

[39]**GAM**'s variables are long-term interest rate and Price-to-rent.

[40]**Lrisk**'s variables are VXO, Risk Appetite, Equity Holding.

[41]**Lval**'s variables are Share price index, Real Estate price, Global Factor in Asset Prices, Short-term interest rate, long-term interest rate, dollar effective exchange rate.

Online Aggregation Rule	RMSE	AUROC
EWA	0.29	0.92
Best convex combination	0.29	0.94
Uniform	0.43	0.66

Table 3: RMSE of different aggregation rules. UK: quasi-real time from 2001Q1 to 2019Q4

Figure 20 in the Appendix show respectively the average and the cumulative loss of the experts. Interestingly, the GAM expert has about the same average loss as the EWA. They both greatly outperform the other experts. The performance of our aggregation rule is evident when one compares its forecast to a subset of the experts as in **Figure 21**.

7.2 Out-of-sample prediction of crises: UK. Real time data.

We reestimate all our experts using only vintage data for the out-of-sample exercise. Despite the strong data limitations, we still get good results for the predictability of the systemic crisis as shown in **Figure 8**. The probability of pre-crisis goes up as before but it remains high longer than in quasi-real time and after the beginning of the crisis. There are only very small spikes during the euro area crisis and a small spike in 2018 so the results are consistent with the quasi-real time ones. There are two models which are picked by the EWA aggregation rule and these are one statistical and one machine learning model: the **GAM**[42] and the **SVM** expert. It is the **GAM** expert which gives the signal of a pre-crisis before 2008. For the UK, it is therefore clearly the behaviour of the real time liquidity variables and the exchange rate which trigger the alarm.

8 Systemische Krisen in Deutschland

We now turn to Germany. Both the timing of the first and the second systemic crises (2001 Q1 till 2003 Q4 and 2007 Q2 till 2013 Q2 respectively) are different from the ones in France and in the UK. There is less time for the algorithm to learn from the first crisis. We therefore expect a worse performance than for France and the UK. As described in Lo Duca et al. (2017), the 2001Q1-

[42]The **GAM**'s variables are the Dollar effective exchange rate, Private Sector Liquidity Stock (2y), Domestic Liquidity Stock local (2Y).

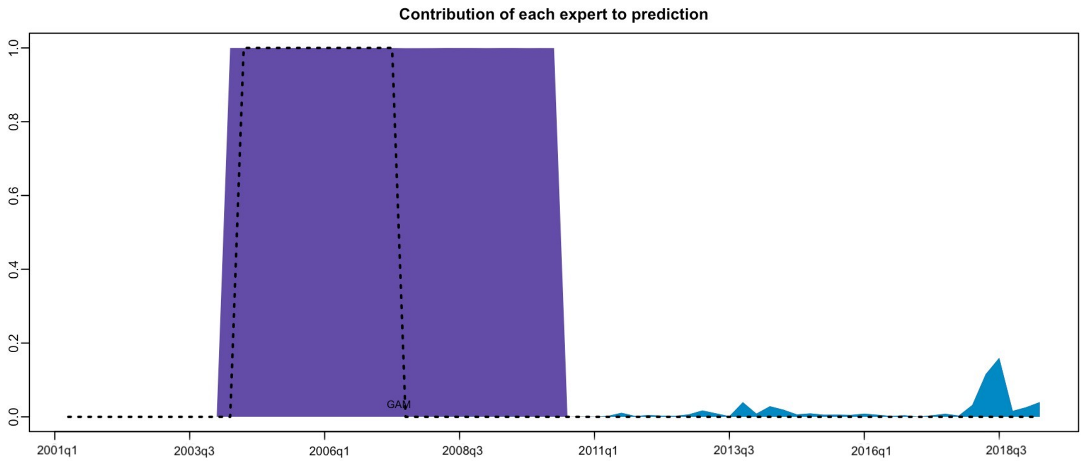

Figure 8: UK: Experts. **Real time** - EWA

2003 Q4 crisis was due to exposure concentration, excessive credit growth and leverage (financial and non financial) and excessive risk taking. The cyclical downturn, following a domestic credit boom and the implosion of the tech bubble, put significant stress on the German financial sector which had low profitability. Some of the largest institutions had to adjust their balance sheets and to tighten their lending standards with negative feedbacks effects. The second systemic crisis 2007 Q2 till 2013 Q2 is predicted out-of-sample. During the years preceding the Lehman Brothers bankruptcy, some German financial institutions became strongly interconnected in international markets and involved in the build-up of systemic risks. The drying up of market and funding liquidity was a key destabilising factor in the crisis. In addition to securitizations, some banks in Germany had important exposures to commercial real estate and the shipping industry. High leverage increased the risk of pro-cyclical fire sales and of a credit crunch. In the later stage of the crisis exposures to stressed euro area sovereigns and banking systems affected the financial sector in Germany. We present results for out-of-sample prediction for 2000Q3 to 2019Q3. Unlike France, there are no residual events during that out-of-sample forecast period but a longer systemic crisis and less time between the batch sample and the out-of-sample systemic crises.

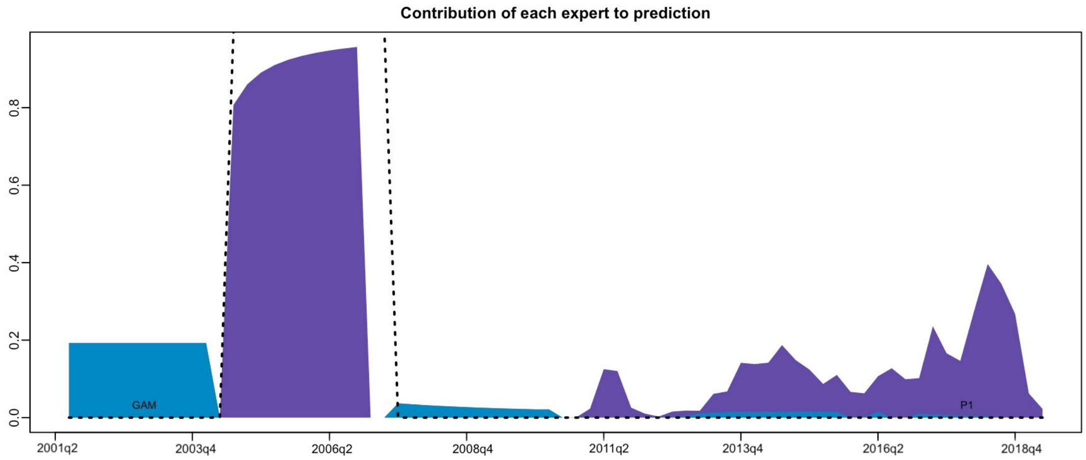

Figure 9: Germany: Contribution of experts to forecasts. Quasi-real time - EWA

8.1 Out-of-sample prediction of crises: Germany. Quasi-real time data.

Figure 9 presents the predicted probability of a pre-crisis in Germany for the EWA aggregation rule. The probability of being in a pre-crisis reaches a very high level in 2004. There are however some subsequent smaller peaks during the 2011-2018 period. We see that when the pre-crisis probability peaks, it is the **P1** expert which is carrying all the weight and giving the signal[43].

Table 4 presents the RMSE and the AUROCs. The EWA performs best once more in all our aggregation rules (unreported) but it does not do as well as the best linear convex combination. This suggests that we have a good pool of experts but that the learning could be improved further. This may be linked to the fact that, for Germany, the two systemic crises are not far apart in time. **Figure 22** and **Figure 23** in the Appendix show respectively the average and the cumulative loss of the experts. Interestingly, expert P1 has actually a slightly lower average and cumulated loss than EWA. They both greatly outperform the other experts. The performance of our aggregation rule is evident when one compares its forecast to a subset of the experts as in **Figure 24**.

[43]**P1**'s variables are Price-to-rent, Real estate price, Banking credit to private non-financial sector, Long term interest rate.

Online Aggregation Rule	RMSE	AUROC
EWA	0.21	0.84
Best convex combination	0.19	0.84
Uniform	0.41	0.78

Table 4: RMSE of different aggregation rules. Germany: quasi-real time

9 Le crisi sistemiche in Italia

Italy experienced a systemic crisis at the beginning of our sample from 1991Q3 to 1997Q4. According to Lo Duca et al. (2017) this crisis was mostly linked to currency markets and the ERM crisis and the distress in the economy and in the banking sector, especially in the South. The systemic crisis we are trying to predict out-of-sample starts later than in the previous economies: it runs from 2011Q2 to 2013Q4. Italy also experienced a "residual event" just before the systemic crisis from 2008Q1 to 2011Q3 due to financial market stress though there was little exposure of Italian banks to US mortgage markets.

9.1 Out-of-sample prediction of crises: Italy. Quasi-real time data

In the case of Italy, the EWA aggregation rule puts almost all its weight on one expert **Lc2** which is a Logit combination of Consumption, Investment, Housing 1, Housing 2, Total Credit to Households and the Global Factor in asset prices. That expert is able to give an accurate forecast of the pre-crisis period in Italy. It also has smaller spikes later in the sample in 2016 and 2018. The RMSE and AUROCs are reported in **Table 5**: the EWA does not perform as well as the best combination of experts suggesting that the learning could be improved. The high RMSE of the uniformly weighted forecast suggest that the pool of experts is not very good on average. **Figure 25** and **Figure 26** in the Appendix show respectively the average and the cumulative loss of the experts. They show that two expert models outperform the EWA in our sample. All the other experts are greatly dominated.

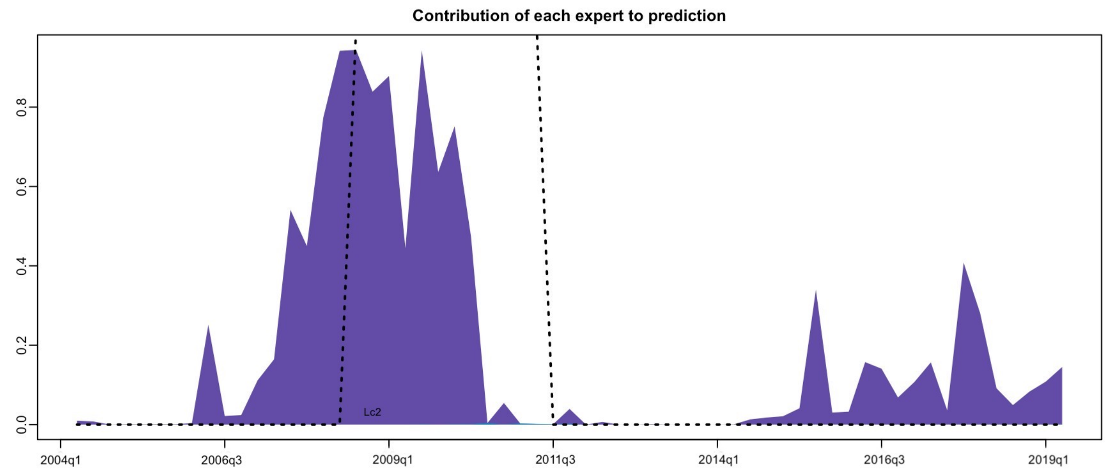

Figure 10: Italy: Contribution of experts to forecasts. Quasi-real time - EWA

Online Aggregation Rule	RMSE	AUROC
EWA	0.34	0.77
Best convex combination	0.28	0.94
Uniform	0.42	0.70

Table 5: RMSE of different aggregation rules. Italy: quasi-real time

10 Conclusions

Online learning algorithms build predictive models by processing data sequentially and do not rely on any statistical assumptions on the data source (see Cesa-Bianchi and Lugosi (2006) and Orabona (2021)).[44] The online-learning methodology has the ability to run a horse race among a very eclectic set of experts (estimated on a batch sample) and to aggregate them in order to produce an optimal out-of-sample forecast of financial crises. We rely on very standard macroeconomic variables, suggested by the literature on financial crises as far back as Fisher (1933). We use a mix of 26 experts, some of them central bank financial crises models, including bayesian averaging models, some of them machine learning and statistical models (random forests, SVT, GAM). Some of our forecasting models make use of the panel dimension of the data and include global factors, others are country specific. We find that for France, UK, Germany, and Italy we are able to predict a systemic financial crisis 3 years ahead out-of-sample with high signal-to-noise

[44]Online learning finds its origins in the pioneering works on repeated games by Robbins (1951), Hannan (1957) and Blackwell (1956). The connections with game theory are deep (see Fudenberg and Levine (1995), Hart and Mas-Colell (2000) and Hart and Mas-Colell (2001)).

ratios and low RMSE compared to the existing literature. Our methodology ensures that asymptotically we always do at least as well as the best expert. In practice, when we compare average losses or cumulative losses, we show that indeed in most cases our aggregation rule outperforms *any* of our individual experts as well as the uniform aggregation of our experts. By design, our methodology is robust to overfitting. We present graphs of our predicted probabilities, which are very transparent indicators of the performance of our methodology. We perform a variety of robustness analyses: we predict the crisis two years ahead instead of three; we use real time vintage data; we test four different aggregation rules; we perform placebo tests. In another paper, Fouliard et al. (2021), we use our methodology with a different specification on a completely different dataset, the historical Jorda-Schularick-Taylor dataset (1870-2017) at annual frequency (see Jordà et al. (2017)) and can confirm the power of our approach.

Our methodology and results are valuable for the conduct of macro-prudential policies, which need to be put in place in a discretionary fashion at the time of the risk build up in order to try to contain very socially costly systemic risks. Unlike monetary and fiscal policy authorities, macro-prudential authorities have so far benefitted from very little modelling of the economy. Yet they are a pillar of the stability of the financial system. Our predicted probabilities of crises, constitute a clear and transparent signal as they increase sizably and monotonously during pre-crisis periods. At which point the macroprudential authorities may want to respond to that increased probability of crisis is a matter of judgement, cannot be mechanical and will depend on external parameters, which we do not attempt to model in this paper. Of course, our methodology is unable to test for causality but our estimates give a clear warning about potential instability and, by overweighting some models, can suggest some areas of the economy that macro-prudential authorities should investigate further with more granular and qualitative data.

There are important lessons we can draw from our estimates. First, echoing Bates and Granger (1969), model aggregation works[45] : the systemic crises of our sample are all predictable ahead of

[45]We underline once more however that online machine learning is different from the real-time forecast combination literature as it does not estimate the coefficient of some underlying model, but instead chooses the best combination of weights to form the most accurate out-of-sample predictions sequentially without making assumptions on the unknown data generating process (see Cesa-Bianchi and Orabona (2021) and Cesa-Bianchi and Lugosi

time -we see our pre-crisis probability increasing significantly and monotonously- and our aggregation methodology outperforms individual forecasting models or their uniform aggregation. Second, and this is in accordance with the *ex post* qualitative narratives of the crises, there is some heterogeneity across countries in terms of which models and variables forecast better. Third, the EWA aggregation rule seems to be the most robust rule in our small size sample with delayed feedback. It performs better than the OGD, the FS rule or the Ridge across countries. Despite the relatively small sample size it converges often towards the best linear combination of experts (asymptotic results). The algorithm learns using ones and zeroes to quickly discard the badly performing models, based on their out-of-sample predictive ability. Fourth, there is considerable time variation in the information content of various models as more information gets revealed. For out-of-sample predictions in quasi-real time, aggregation rules tend to put a high weight on models with credit, housing and risk taking variables but those weights are heterogeneous depending on the countries. For France, credit and real estate variables contribute jointly to give a signal three year ahead. International variables, risk taking and prices are more important two years ahead. This echoes other findings in the literature showing that quantity variables have better predictive ability than asset prices at longer horizons (when asset prices signal trouble, it is often too late - see Coimbra et al. (2021)). For the UK, long term interest rates and price-to-rent, give most of the signal three year ahead while asset prices, risk taking and the exchange rate are more important two year ahead. For Germany it is long term rates, banking credit and real estate variables which seem more informative. For Italy, real activity, credit, housing market and international conditions seem most informative. Clearly it is very important to allow for time varying weights. Real estate variables, credit, risk appetite and monetary and real variables are important at different times. This heterogeneity is broadly in line with the qualitative *ex post* narratives of the crises and their causes. This is where the online nature of our algorithm is key as standard methodologies would not be able to extract enough information from the sample. As a "meta-statistical" approach, our method is very flexible: we could incorporate many more experts (deep learning, subjective judgement, any kind of models combinations) and potentially

(2006)).

increase further the performance of our EWA forecast. Across our countries we see differences in the performance of our pool of experts. They may be well suited for France, Germany, the UK and may be less so for Italy. When we switch to the use of vintage data (for France and the UK), we lose a lot of relevant information, particularly the credit quantity variables, which are informative to predict financial stability. This experiment is very strict as we use vintage data even for the estimation of the experts on the batch sample. Strikingly the model is however still able to predict the pre-crisis period for both France and the UK, though with a lower accuracy. It relies on financial stress, interest rates, liquidity and international variables.

To conclude, we could add to the letter of the British Academy addressed to the Queen that we should use machine learning tools in our policy institutions. They can give precious hints to guide humans in charge of financial stability regarding when and where they should up their game, summon their imaginative capacity and exercise their best judgement.

References

Addo, Peter Martey, Dominique Guégan, and Bertrand Hassani (2018) "Credit Risk Analysis using Machine and Deep learning models," *Documents de travail du Centre d'Economie de la Sorbonne*, No. 18003, February.

Alessi, Lucia and Carsten 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*, Vol. 27, No. 3, pp. 520–533.

Alessi, Lucia, António Antunes, Jan Babeckỳ, Simon Baltussen, Markus Behn, Diana Bonfim, Oliver Bush, Carsten Detken, Jon Frost, Rodrigo Guimaraes et al. (2015) "Comparing different early warning systems: Results from a horse race competition among members of the macro-prudential research network," *http://dx.doi.org/10.2139/ssrn.2566165*.

Alquier, Pierre and Karim Lounici (2011) "PAC-Bayesian bounds for sparse regression estimation with exponential weights," *Electronic Journal of Statistics*, Vol. 5, pp. 127–145.

Amat, Christophe, Tomasz Michalski, and Gilles Stoltz (2018) "Fundamentals and exchange rate forecastability with simple machine learning methods," *Journal of International Money and Finance*, Vol. 88, pp. 1–24.

Antunes, António, Diana Bonfim, Nuno Monteiro, and Paulo Rodrigues (2014) "Early warning indicators of banking crises: exploring new data and tools," *Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies*.

Barron, Andrew and Thomas Cover (1991) "Minimum complexity density estimation," *Information Theory, IEEE Transactions on*, Vol. 37, pp. 1034 – 1054, 08.

Bates, John M and Clive WJ Granger (1969) "The combination of forecasts," *Journal of the Operational Research Society*, Vol. 20, No. 4, pp. 451–468.

Behn, Markus, Carsten Detken, Tuomas A Peltonen, and Willem Schudel (2013) "Setting countercyclical capital buffers based on early warning models: would it work?" *ECB working paper*.

Beutel, Johannes, Sophia List, and Gregor von Schweinitz (2019) "Does machine learning help us predict banking crises?" *Journal of Financial Stability*, Vol. 45, p. 100693.

Blackwell, David (1956) "An analog of the minimax theorem for vector payoffs," *Pacific Journal of Mathematics*, Vol. 6, No. 1, pp. 1–8.

Bluwstein, Kristina, Marcus Buckmann, Andreas Joseph, Miao Kang, Sujit Kapadia, and Özgür Simsek (2020) "Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach," *Bank of England Working Paper*, No. 848.

Bordo, Michael, Barry Eichengreen, Daniela Klingebiel, and Maria Soledad Martinez-Peria (2001) "Is the crisis problem growing more severe?" *Economic policy*, Vol. 16, No. 32, pp. 52–82.

Borio, Claudio EV and Mathias Drehmann (2009) "Assessing the risk of banking crises–revisited," *BIS Quarterly Review, March*.

Bush, Olivier, Rodrigo Guimaraes, and Hanno Stremmel (2015) "Beyond the credit gap: bank balance sheets, credit and price of risk," *Manuscript, Bank of England*.

Bussiere, Matthieu and Marcel Fratzscher (2006) "Towards a new early warning system of financial crises," *Journal of International Money and Finance*, Vol. 25, No. 6, pp. 953–973.

Capistrán, Carlos and Allan Timmermann (2009) "Forecast combination with entry and exit of experts," *Journal of Business & Economic Statistics*, Vol. 27, No. 4, pp. 428–440.

Casabianca, Elizabeth Jane, Michele Catalano, Lorenzo Forni, Elena Giarda, Simone Passeri et al. (2019) "An Early Warning System for banking crises: From regression-based analysis to machine learning techniques," *Marco Fanno Working Papers 235*.

Cesa-Bianchi, Nicolo and Gabor Lugosi (2006) *Prediction, Learning and Games*: Cambridge University Press.

Cesa-Bianchi, Nicolò and Francesco Orabona (2021) "Online Learning Algorithms," *Annual Review of Statistics and Its Application*, Vol. 8, pp. 165–190.

Coimbra, Nuno and Hélène Rey (2017) "Financial Cycles with Heterogeneous Intermediaries," *NBER Working Papers*, No. 23245, March.

Coimbra, Nuno, Daisoon Kim, and Hélène Rey (2021) "Central Bank Policy and the concentration of risk: Empirical estimates," *Journal of Monetary Economics*.

Coudert, Virginie and Julien Idier (2016) "An Early Warning System for Macro-prudential Policy in France," *Banque de France Working Paper 609*.

Dalalyan, Arnak S and Joseph Salmon (2012) "Sharp oracle inequalities for aggregation of affine estimators," *The Annals of Statistics*, Vol. 40, No. 4, pp. 2327–2355.

Dalalyan, Arnak and Alexandre B Tsybakov (2008) "Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity," *Machine Learning*, Vol. 72, No. 1-2, pp. 39–61.

Dani, Varsha, Omid Madani, David M Pennock, Sumit Sanghai, and Brian Galebach (2012) "An empirical comparison of algorithms for aggregating expert predictions," *arXiv preprint arXiv:1206.6814*.

Dashevskiy, Mikhail and Zhiyuan Luo (2011) "Time series prediction with performance guarantee," *IET communications*, Vol. 5, No. 8, pp. 1044–1051.

De Blasi, Pierpaolo and Stephen Walker (2013) "Bayesian Estimation of the Discrepancy with Misspecified Parametric Models," *Bayesian Analysis*, Vol. 8, pp. 781–800, 12.

Demirgüç-Kunt, Asli and Enrica Detragiache (1998) "The determinants of banking crises in developing and developed countries," *Staff Papers*, Vol. 45, No. 1, pp. 81–109.

Destrero, Augusto, Christine Mol, Francesca Odone, and Alessandro Verri (2009) "A Regularized Framework for Feature Selection in Face Detection and Authentication," *International Journal of Computer Vision*, Vol. 83, pp. 164–177, 06.

Devaine, Marie, Pierre Gaillard, Yannig Goude, and Gilles Stoltz (2013) "Forecasting electricity consumption by aggregating specialized experts," *Machine Learning*, Vol. 90, pp. 231–260, 01.

Diamond, Douglas W and Philip H Dybvig (1983) "Bank Runs, Deposit Insurance, and Liquidity," *Journal of Political Economy*, Vol. 91, No. 3, pp. 401–19, June.

Diebold, Francis X and Peter Pauly (1987) "Structural change and the combination of forecasts," *Journal of Forecasting*, Vol. 6, No. 1, pp. 21–40.

Diebold, Francis X. and Minchul Shin (2019) "Machine learning for regularized survey forecast combination: Partially-egalitarian LASSO and its derivatives," *International Journal of Forecasting*, Vol. 35, No. 4, pp. 1679–1691.

Döpke, Jörg, Ulrich Fritsche, and Christian Pierdzioch (2017) "Predicting recessions with boosted regression trees," *International Journal of Forecasting*, Vol. 33, No. 4, pp. 745–759.

Duprey, Thibaut, Benjamin Klaus, and Tuomas Peltonen (2017) "Dating Systemic Financial Stress Episodes in the EU Countries," *Journal of Financial Stability*, Vol. 32, 08.

Eichengreen, Barry and Carlos Arteta (2002) "Banking crises in emerging markets: presumptions and evidence," *Financial policies in emerging markets*, pp. 47–94.

Elliott, Graham and Allan Timmermann (2008) "Economic forecasting," *Journal of Economic Literature*, Vol. 46, No. 1, pp. 3–56.

Fisher, Irving (1933) "The Debt-Deflation Theory of Great Depressions," *Econometrica*, Vol. 1, No. 4, pp. 337–357.

Fouliard, Jérémy, Hélène Rey, and Vania Stavrakeva (2021) "Is this Time Different? Financial Follies across Centuries," *London Business School and CEPR.*

Frankel, Jeffrey A and Andrew K Rose (1996) "Currency crashes in emerging markets: An empirical treatment," *Journal of International Economics*, Vol. 41, No. 3-4, pp. 351–366.

Frankel, Jeffrey and George Saravelos (2012) "Can leading indicators assess country vulnerability? Evidence from the 2008–09 global financial crisis," *Journal of International Economics*, Vol. 87, No. 2, pp. 216–231.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2010) "Regularized Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, Vol. 33, 02.

Fudenberg, Drew and David K Levine (1995) "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, Vol. 19, No. 5-7, pp. 1065–1089.

Gennaioli, Nicola, Andrei Shleifer, and Robert Vishny (2012) "Neglected risks, financial innovation, and financial fragility," *Journal of Financial Economics*, Vol. 104, No. 3, pp. 452 – 468.

Gourinchas, Pierre-Olivier and Maurice Obstfeld (2012) "Stories of the Twentieth Century for the Twenty-First," *American Economic Journal: Macroeconomics*, Vol. 4, No. 1, pp. 226–65.

Grunwald, Peter and Thijs van Ommen (2014) "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It," *Bayesian Analysis*, Vol. 12, 12.

Hannan, J (1957) "Approximation to Bayes risk in repeated plays. Dresher M, Tucker A, Wolfe P, eds. Contributions to the Theory of Games, Vol. III."

Hart, Sergiu and Andreu Mas-Colell (2000) "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, Vol. 68, No. 5, pp. 1127–1150.

——— (2001) "A general class of adaptive strategies," *Journal of Economic Theory*, Vol. 98, No. 1, pp. 26–54.

Hastie, T.J. and R.J. Tibshirani (1986) *Generalized additive models*, Vol. 1, pp.297-318.

Jordà, Oscar, Moritz Schularick, and Alan M. Taylor (2015) "Leveraged bubbles," *Journal of Monetary Economics*, Vol. 76, pp. S1–S20.

Jordà, Òscar, Moritz Schularick, and Alan M Taylor (2017) "Macrofinancial history and the new business cycle facts," *NBER macroeconomics annual*, Vol. 31, No. 1, pp. 213–263.

Joulani, Pooria, Andras Gyorgy, and Csaba Szepesvári (2013) "Online learning under delayed feedback," in *International Conference on Machine Learning*, pp. 1453–1461.

Kaminski, Graciela and Carmen Reinhart (1999) "The Twin Crises: The Causes of Banking and Balance of Payments Problems," *American Economic Review*, Vol. 89, No. 3, pp. 473–500, June.

Kindleberger, Charles (1978) "Manias, Panic and Crashes," *New York Basic Books.*

Laeven, Luc and Fabian Valencia (2020) "Systemic Banking Crises Database: A Timely Update in COVID-19 Times," *CEPR Discussion Paper No. 14569.*

Lo Duca, Marco, Anne Koba, Marisa Basten, Elias Bengtsson, Benjamin Klaus, Piotr Kusmierczyk, Jan Hannes Lang, Carsten Detken, and Tuomas Peltonen (2017) "A New Database for Financial Crises in European Countries," *ECB Occasional Paper*, No. 194.

Lowe, Philip and Claudio Borio (2002) "Asset prices, financial and monetary stability: exploring the nexus," BIS Working Papers 114, Bank for International Settlements.

Mallet, Vivien, Gilles Stoltz, and Boris Mauricette (2009) "Ozone ensemble forecast with machine learning algorithms," *Journal of Geophysical Research: Atmospheres*, Vol. 114, No. D5.

Martinez-Miera, David and Rafael Repullo (2017) "Search for yield," *Econometrica*, Vol. 85, No. 2, pp. 351–378.

McCallester, David (2001) "PAC-Bayesian Stochastic Model Selection," *Machine Learning*, Vol. 51, 08.

Meese, Richard A and Kenneth Rogoff (1983) "Empirical exchange rate models of the seventies: Do they fit out of sample?" *Journal of International economics*, Vol. 14, No. 1-2, pp. 3–24.

Mian, Atif and Amir Sufi (2009) "The Consequences of Mortgage Credit Expansion: Evidence from the U.S. Mortgage Default Crisis," *The Quarterly Journal of Economics*, Vol. 124, No. 4, pp. 1449–1496, 11.

Minsky, Hyman (1986) "Global Consequences of Financial Deregulation," *Washington University Working Paper*, No. 96.

Mol, Christine, Ernesto De Vito, and Lorenzo Rosasco (2009) "Elastic-Net Regularization in Learning Theory," *Journal of Complexity*, Vol. 25, pp. 201–230, 04.

Mol, Christine, Sofia Mosci, Magali Traskine, and Alessandro Verri (2009) "A Regularized Method for Selecting Nested Groups of Relevant Genes from Microarray Data," *Journal of computational biology*, Vol. 16, pp. 677–90, 06.

Monteleoni, Claire, Gavin A Schmidt, Shailesh Saroha, and Eva Asplund (2011) "Tracking climate models," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 4, No. 4, pp. 372–392.

Ng, Serena (2014) "Boosting recessions," *Canadian Journal of Economics/Revue canadienne d'économique*, Vol. 47, No. 1, pp. 1–34.

Orabona, Francesco (2021) “A Modern Introduction to Online Learning,” *arXiv:1912.13213v4*.

Pesaran, Hashem and Allan Timmermann (2005) “Real-Time Econometrics,” *Econometric Theory*, Vol. 21, No. 1, pp. 212–231.

Reinhart, Carmen M. and Kenneth S. Rogoff (2009) *This Time Is Different: Eight Centuries of Financial Folly*, Princeton, New Jersey: Princeton University Press.

Richter, Björn, Moritz Schularick, and Paul Wachtel (2021) “When to lean against the wind,” *Journal of Money, Credit and Banking*, Vol. 53, No. 1, pp. 5–39.

Rigollet, Philippe and Alexandre B Tsybakov (2012) “Sparse estimation by exponential weighting,” *Statistical Science*, Vol. 27, No. 4, pp. 558–575.

Robbins, Herbert (1951) “Asymptotically subminimax solutions of compound statistical decision problems,” in *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pp. 131–149, University of California Press.

Rossi, Barbara (2011) “Advances in Forecasting Under Instability,” *Handbook of Economic Forecasting*, Vol. 2, 09.

Schularick, Moritz and Alan M. Taylor (2012) “Credit Booms Gone Bust: Monetary Policy, Leverage Cycles, and Financial Crises, 1870-2008,” *American Economic Review*, Vol. 102, No. 2, pp. 1029–61.

Shin, Hyun Song (2013) “Procyclicality and the search for early warning indicators,” *International Monetary Fund Working Paper 258*.

Stock, James H and Mark W Watson (1996) “Evidence on structural instability in macroeconomic time series relations,” *Journal of Business & Economic Statistics*, Vol. 14, No. 1, pp. 11–30.

Stock, James and Mark Watson (2012) “Disentangling the Channels of the 2007ÃÂfiÂ?!09 Recession [with Comments and Discussion],” *Brookings Papers on Economic Activity*, pp. 81–156, 01.

Stoltz, Gilles (2010) “Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l’air et à celle de la consommation électrique,” *Journal de la Société Francaise de Statistique*, Vol. 151, 01.

Stoltz, Gilles and Gábor Lugosi (2007) “Learning correlated equilibria in games with compact sets of strategies,” *Games and Economic Behavior*, Vol. 59, No. 1, pp. 187–208.

Sufi, Amir and Alan M Taylor (2021) “Financial crises: A survey,” *National Bureau of Economic Research Working Paper 29155*, August.

Tibshirani, Robert (1996) “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, pp. 267–288, 01.

Timmermann, Allan (2006) “Chapter 4: Forecast Combinations,” Vol. 1 of Handbook of Economic Forecasting: Elsevier, pp. 135–196.

Vovk, V.G. (1990) “Aggregating strategies,” *COLT*, pp. 371–386, 01.

Walker, Stephen and Nils Hjort (2001) “On Bayesian consistency,” *Journal of the Royal Statistical Society Series B*, Vol. 63, pp. 811–821, 02.

Ward, Felix (2017) "Spotting the danger zone: Forecasting financial crises with classification tree ensembles and many predictors," *Journal of Applied Econometrics*, Vol. 32, pp. 359–378.

Weinberger, Marcelo J and Erik Ordentlich (2002) "On delayed prediction of individual sequences," *IEEE Transactions on Information Theory*, Vol. 48, No. 7, pp. 1959–1976.

Zhang, Tong (2006) "Information-theoretic upper and lower bounds for statistical estimation," *Information Theory, IEEE Transactions on*, Vol. 52, pp. 1307 – 1321, 05.

Zhang, Yongli (2012) "Support Vector Machine Classification Algorithm and Its Application,", pp. 179–186, 09.

Zinkevich, Martin (2003) "Online convex programming and generalized infinitesimal gradient ascent," in *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936.

Zou, Hui and T. Hastie (2005) "Regularization and Variable Selection via the Elastic Nets," *Journal of the Royal Statistical Society: Series B*, Vol. 67, pp. 301–320, 01.

Figure 11: France: Weights. Quasi-real time. FS aggregation rule.

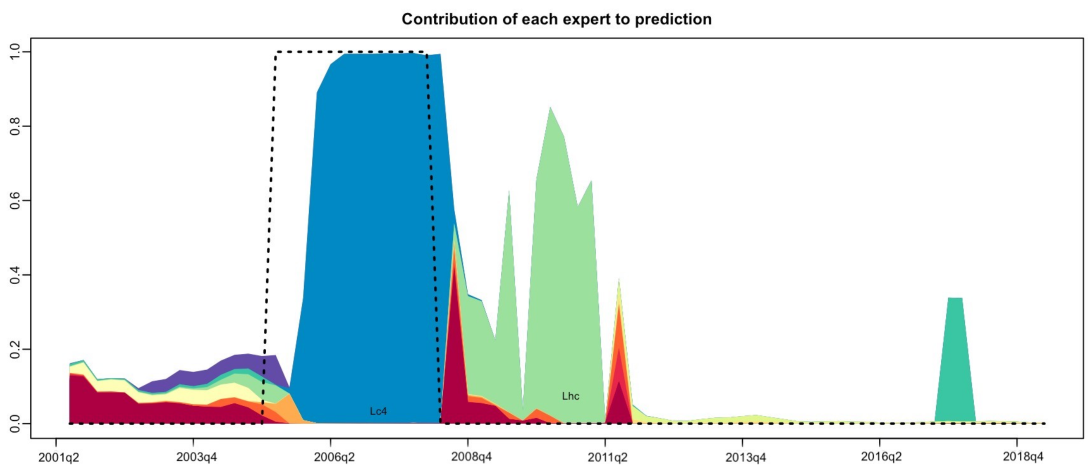

Figure 12: France: Experts. Quasi-real time. Contribution to forecast. FS aggregation rule.

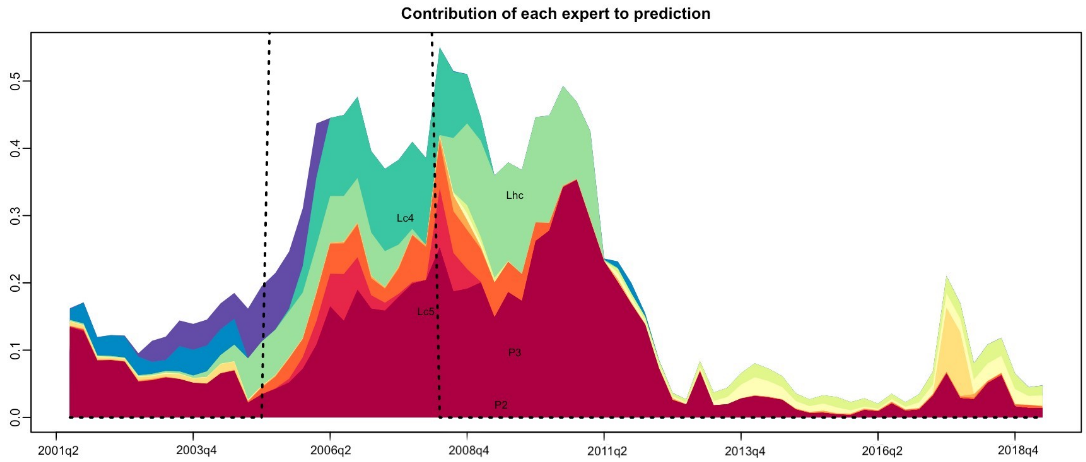

Figure 13: France: Experts contribution to forecast. OGD aggregation rule

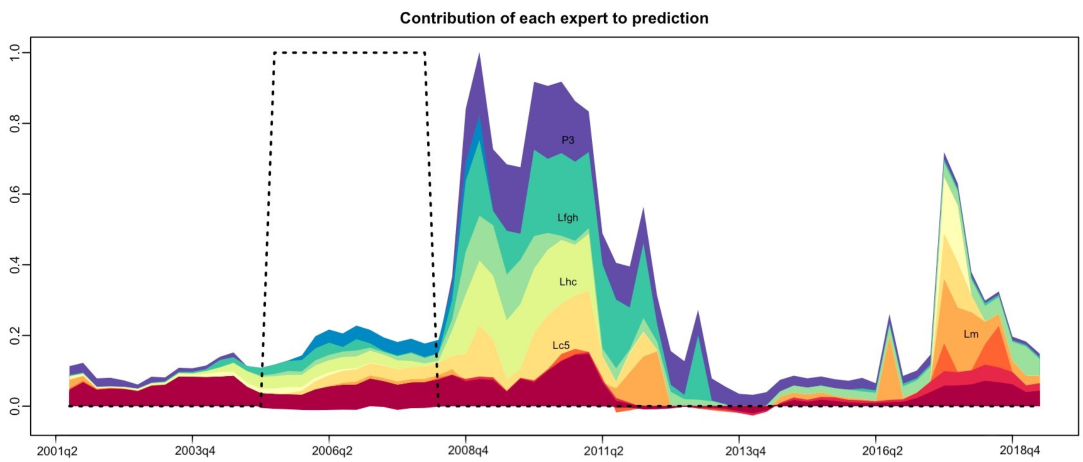

Figure 14: France: Experts contribution to forecast. Ridge aggregation rule

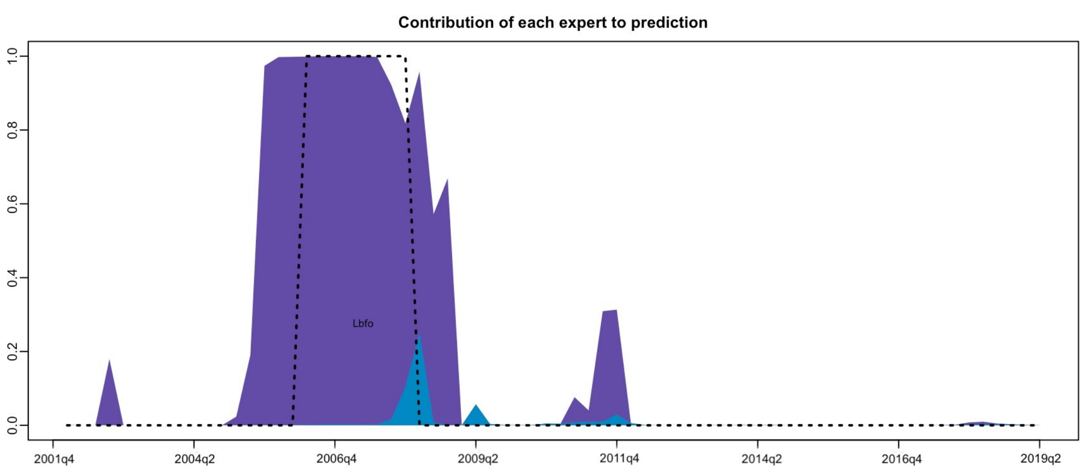

Figure 15: France: Experts contribution to forecast. EWA aggregation rule. **2 year pre-crisis period**.

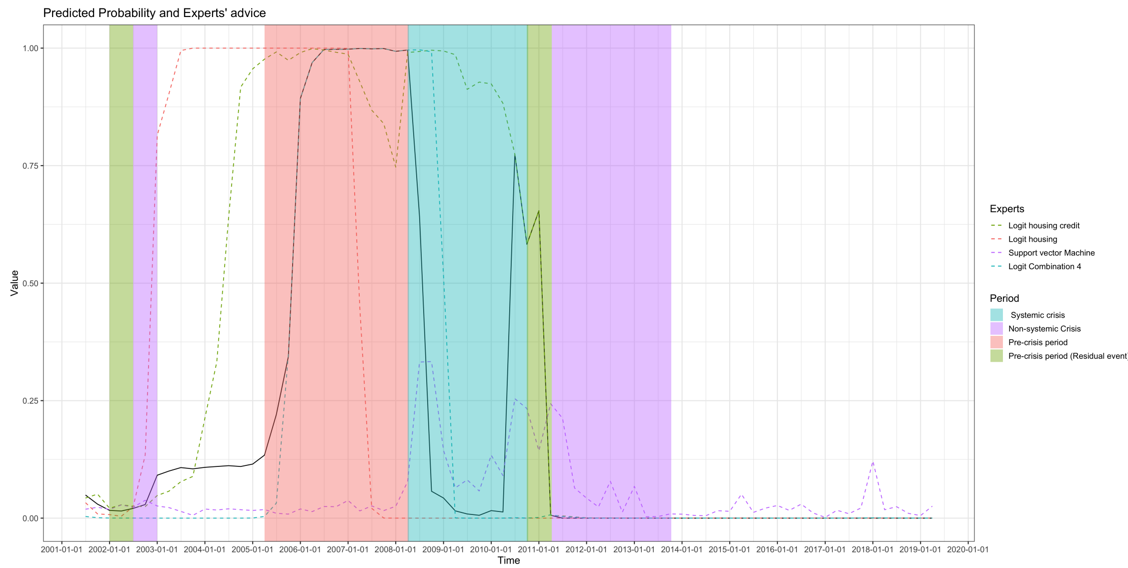

Figure 16: France: A Subset of Experts' Predicted Probability of Crises versus EWA aggregation.

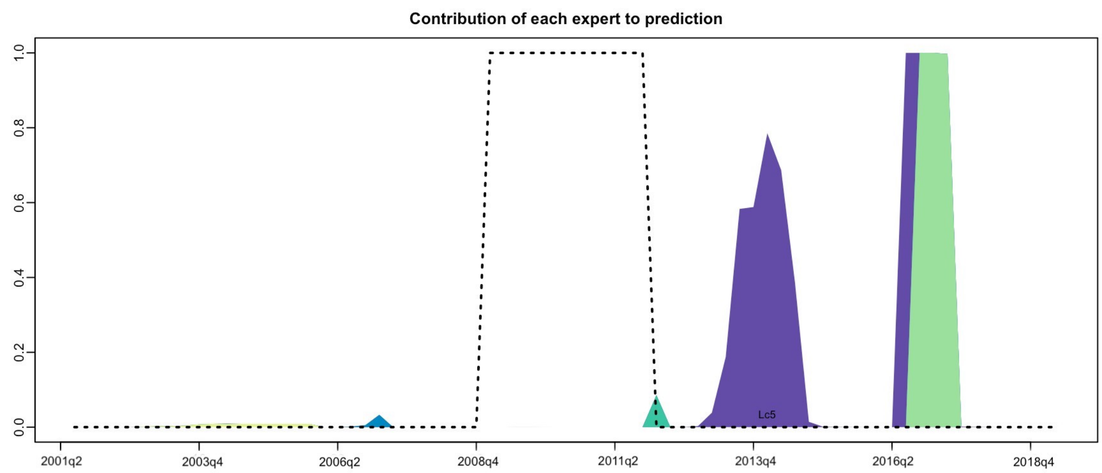

Figure 17: France: **Placebo test**. EWA aggregation.

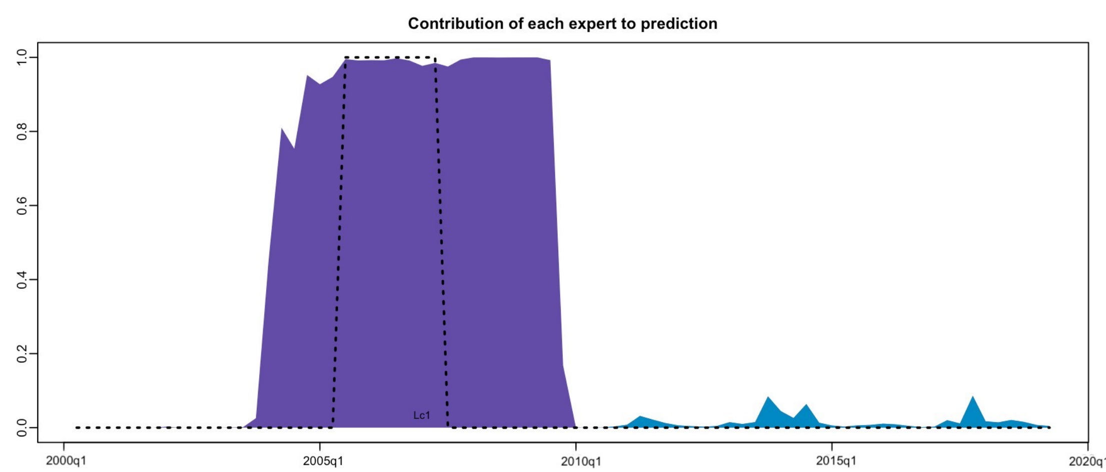

Figure 18: UK: Experts contribution to forecast. EWA aggregation rule. **2 year pre-crisis period**.

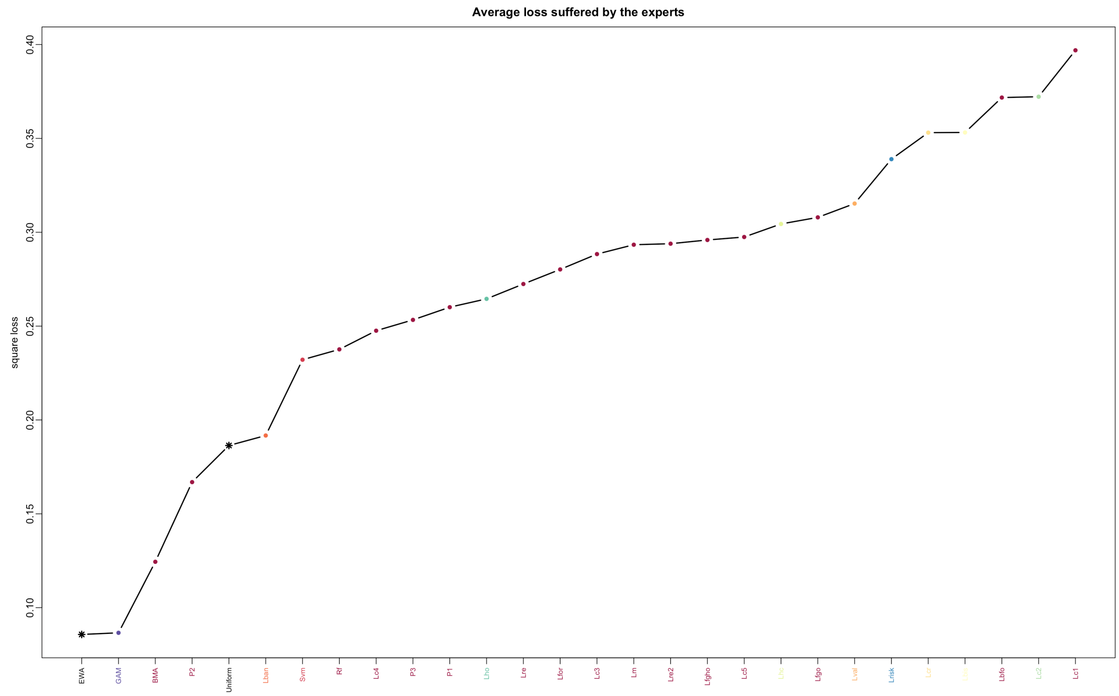

Figure 19: UK Experts. Quasi-real time. Average Loss. EWA aggregation.

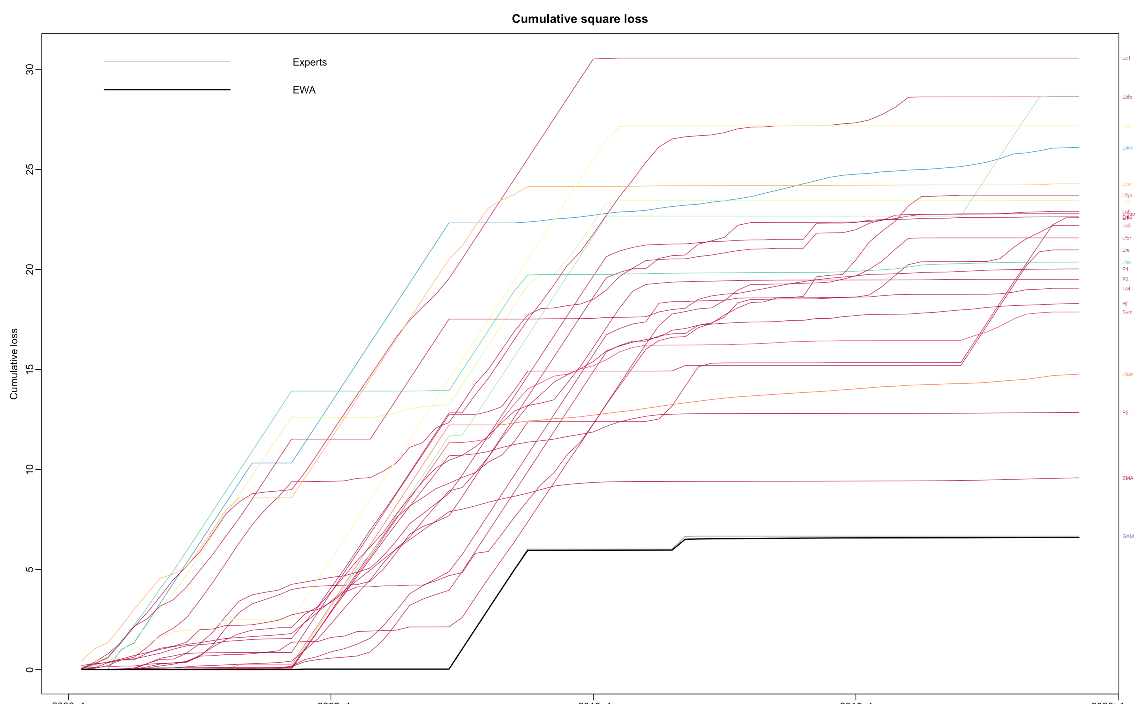

Figure 20: UK Experts. Quasi-real time. Cumulative Loss. EWA aggregation.

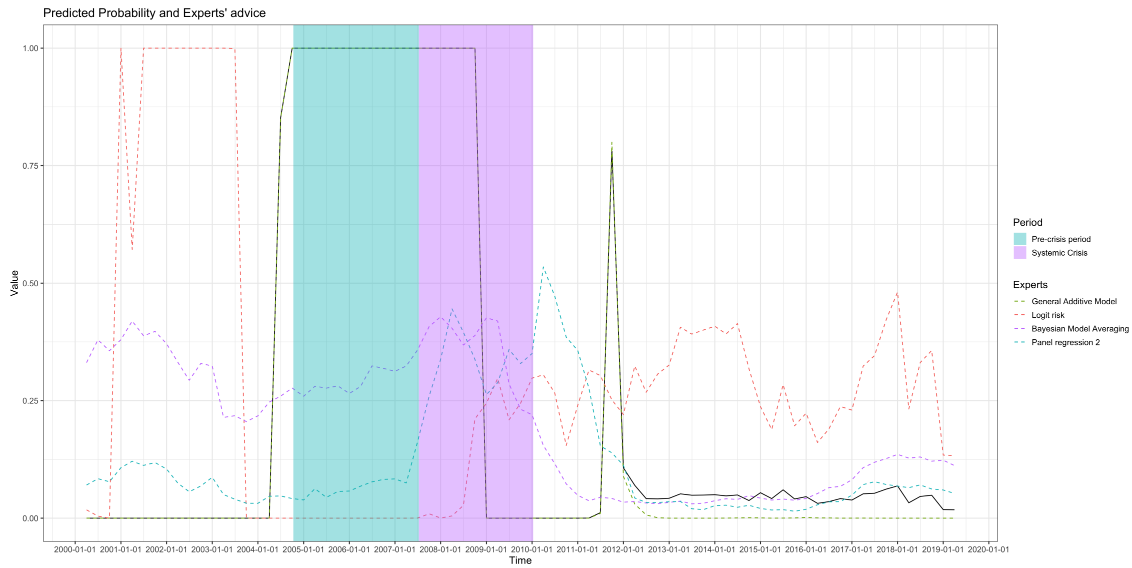

Figure 21: UK: A Subset of Experts' Predicted Probability of Crises versus EWA aggregation.

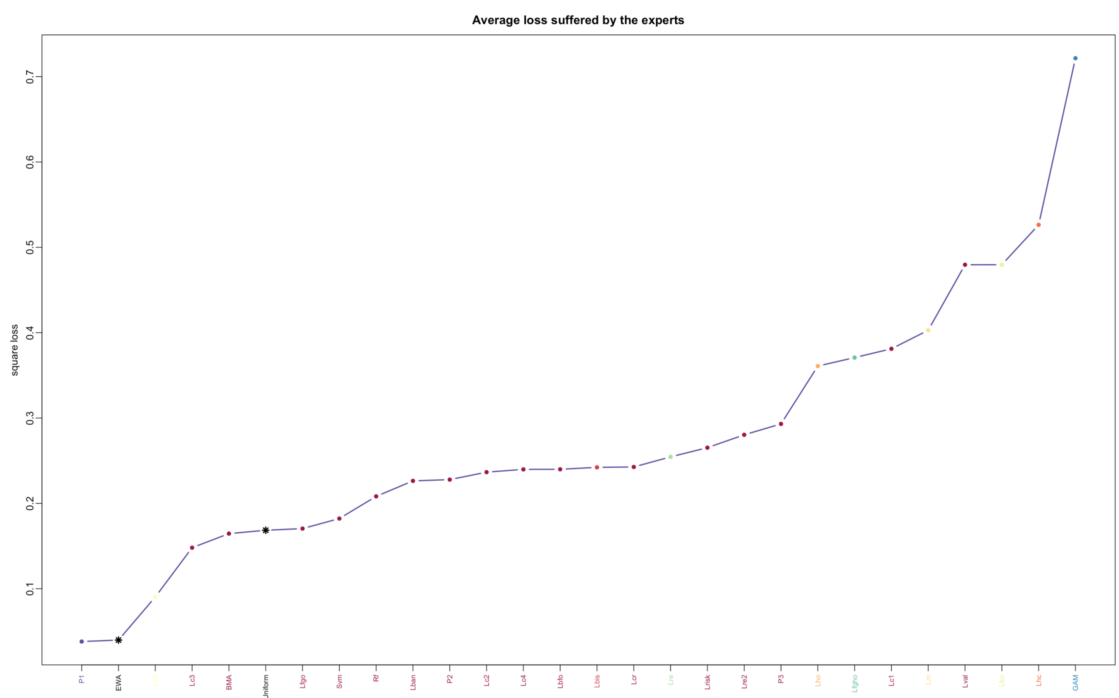

Figure 22: Germany Experts. Quasi-real time. Average Loss. EWA aggregation.

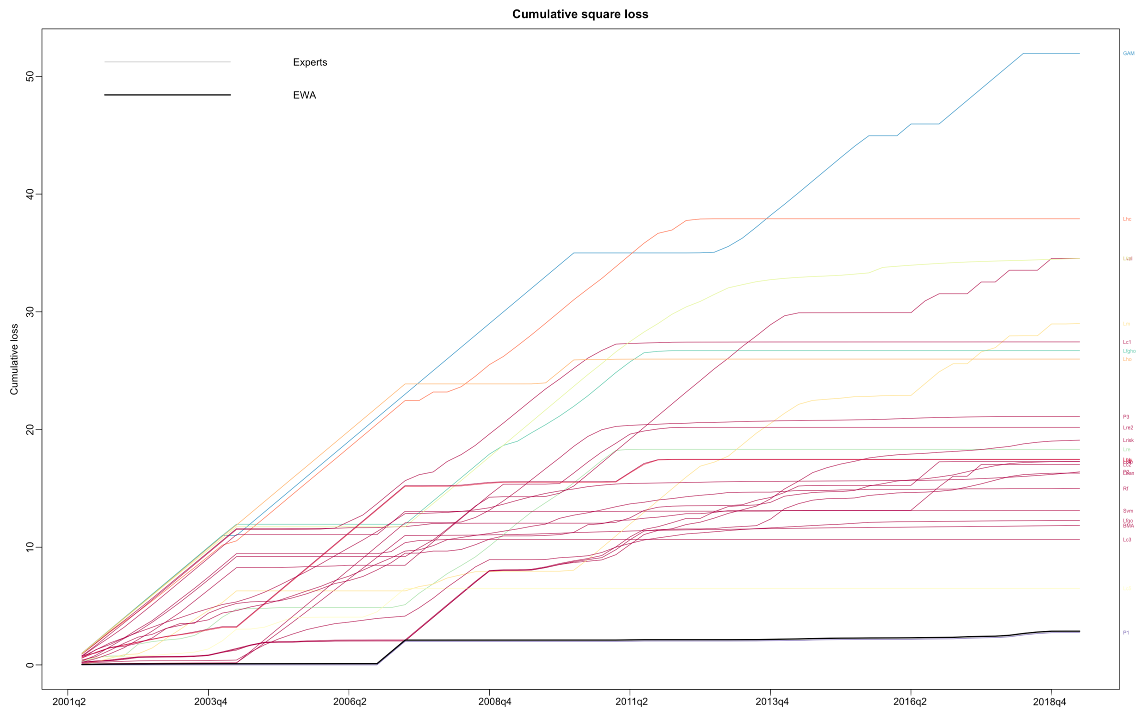

Figure 23: Germany Experts. Quasi-real time. Cumulative Loss. EWA aggregation.

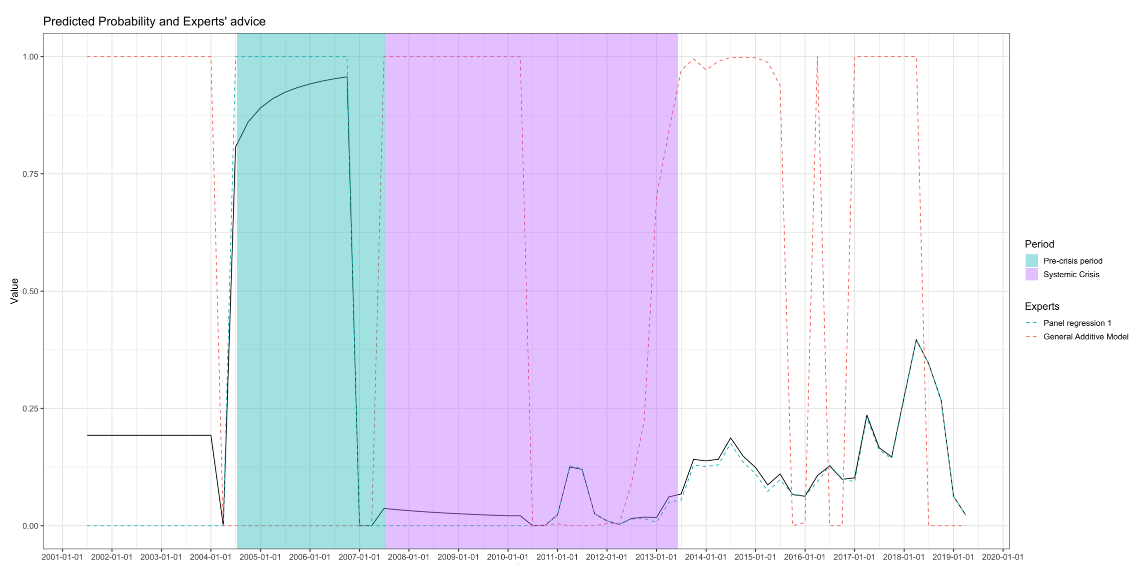

Figure 24: Germany: A Subset of Experts' Predicted Probability of Crises versus EWA aggregation.

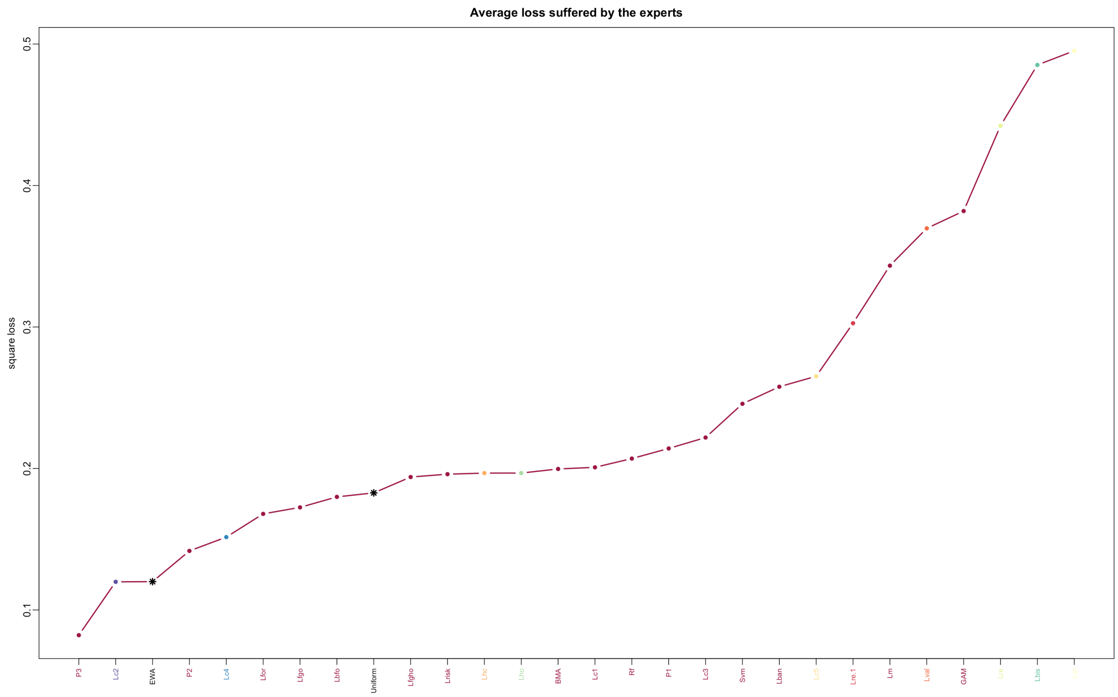

Figure 25: Italy Experts. Quasi-real time. Average Loss. EWA aggregation.

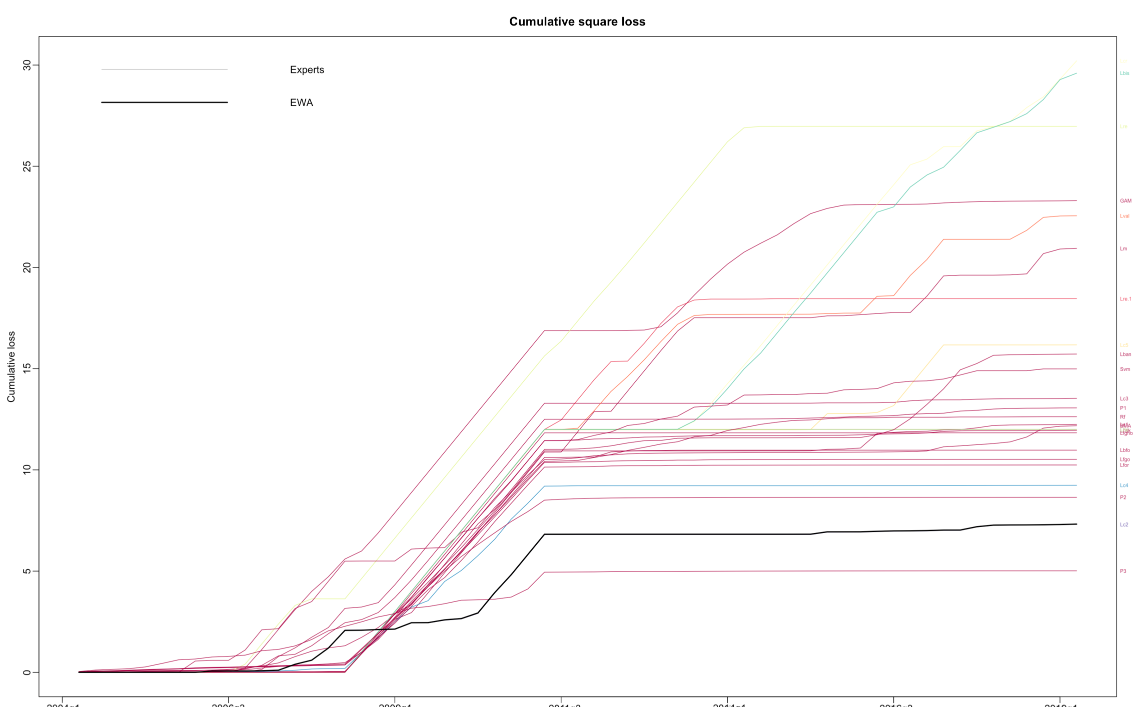

Figure 26: Italy Experts. Quasi-real time. Cumulative Loss. EWA aggregation.

Online Appendix

A Data

A.1 Database of systemic crises

We use the official database of systemic crises provided by Lo Duca et al. (2017) and replicate the same methodology for the US, the only non-european country in our database. This approach consists in two steps. First, it aims at identifying historical episodes of elevated financial stress which were also associated with economic slowdowns. This step provides a preliminary list of potential systemic crisis events for consideration. Then, each national authority distinguish between systemic crisis and residual episodes of financial stress and exercises judgement on the precise dates.

As in Lo Duca et al. (2017), we construct a country-specific financial stress index which captures three financial market segments :

- **Equity market**. We capture market stress with two variables : the quarterly average of absolute log-returns of the real stock price index (VSTX) and the cumulative maximum loss (CMAX) that corresponds to the maximum loss compared to the highest level of the stock market over two years. Before computing volatilities, we divide the data by a 10 years trailing standard deviation.
- **Bond market**. We capture stress in the bonds market with two variables : the quarterly realised volatility (VR10) is computed as the quarterly average of absolute changes in the real 10-year government bond yields and the increase of a 10-year fibond indexfi compared to the minimum (CMIN) over a two-year rolling window.
- **Foreign exchange market** . We capture foreign exchange market stress with two variables: the realised volatility (VEER) is computed as the absolute value of the growth rate of the real effective exchange rate and the cumulative change (CUMUL) over 2 quarters.

Then, we apply a Markov Switching model to endogenously determine low and high financial stress events. Finally, in order to produce a list of potential systemic crisis events, we only select financial stress episodes associated with real economic stress : i) with at least six consecutive months of negative industrial production growth ii) which overlap at least partly with a decline in real GDP during at least two possibly non-consecutive quarters.

During the second step, each national authority has to identify systemic crises among the list of potential systemic crisis events, following common guidelines - for the US, we contacted the Federal Reserve Bank of New York. An event of financial stress is classified as a systemic crisis if it fulfils one or more of the following three criteria :
i) A contraction in the supply of financial intermediation or funding to the economy took place during the potential crisis event. The financial system played a role in originating or amplifying shocks, thereby contributing substantially to negative economic outcomes. Examples : Despite remaining solvent, banks significantly contract the supply of credit to the real economy due to market distress and funding difficulties. Foreign capital is withdrawn and the supply of credit to the domestic economy shrinks (currency crisis).
ii) The financial system was distressed during the potential crisis event. Examples : Market infrastructures were dysfunctional. There were bankruptcies among large/significant financial institutions.
iii) Policies were adopted to preserve financial stability or bank stability during the potential crisis event. Examples : External support (IMF interventions). Extraordinary provision of central bank liquidity. Direct interventions of the state in support of the banking system (liability guarantees, recapitalisation or nationalisation of banks, assisted/forced mergers among institutions and creation of bad banks and/or asset management companies). Monetary policy actions with a financial stability angle.

A.2 Variables

A.2.1 Quasi-real time data

- **Macroeconomic indicators** : GDP, GDP per person employed, GDP per capita, GDP per hour worked, Unemployment rate, Consumer Price Index, General Government Debt, Golden rule (gap of real long term interest rate to real GDP), Political Uncertainty Index, Oil price index, Consumption, Investment, Multifactor Productivity.

- **Credit and Debt indicators** : Total credit (to households, to private non-financial sector, to non-financial firms), Debt Service Ratios (household, non-financial corporations, private non-financial sector), Household Debt, General Government Debt.

- **Banking sector indicators**: Banking credit to private sector, Bank assets, Bank equity.

- **Interest rates and monetary indicators** : 3-month rate, 10-year rate, slope of the yield curve (10y-3m), monetary aggregate M3.

- **Real estate indicators** : Loans for House purchase, Residential real estate prices, Price-to-income ratio, Price-to-rent ratio, rent price index, house price forecasts.

- **Market indicators**: Share prices, Financial Conditions Index, Risk Appetite Index, oil price, Equity holdings, Financial assets, VXO, Global Factor in Asset Prices.
- **External condition indicators**: Cross-border flows, Real effective exchange rate, Dollar effective exchange rate, Current account, Shipping indicator; export growth, import growth, terms of trade, growth of Foreign Exchange Reserves, External Debt.
- **Liquidity Indicators**: Total Liquidity, Domestic Liquidity, Policy Liquidity.

A.2.2 Real time data

Because of the lack of vintage data, we only use market indicators, external condition indicators (except current account), liquidity indicators and some monetary indicators (3-month rate, 10-year rate, slope of the yield curve (10y-3m), monetary aggregate M3) for real-time forecasts. Due to the importance of real estate, credit and debt variables to predict systemic crises, this lack of vintage data is problematic. As a consequence, we add several market, liquidity, monetary and external condition real-time indicators from the CrossBorder Capital vintage database.

- **Market indicators**: Equity Exposure Index, Bond Exposure Index, Financing Risk Index, Forex Risk Index, Composite Risk Index.
- **External condition indicators**: Foreign Exchange Reserves, Gross Capital Flows, Currency Exposure Index, Exposure Risk Index.
- **Interest rates and monetary indicators** : Central Bank Intervention.
- **Liquidity Indicators**: Quantity Liquidity Index, Momentum index.

We use the following data sources. Macroeconomic, external, real estate and monetary indicators come from the OECD whereas credit and debt indicators come from the BIS database. Liquidity data and some market indicators (Risk Appetite Index, Financial Condition Index) come from CrossBorder Capital. The notation ###### - ###### denotes different starting and ending dates depending on the country. Additional data used to predict systemic crisis in real-time come from the CrossBorder Capital vintage database.

Variable name	Frequency	Time Range (base:1985)	Source
Dollar effective exchange rate	Q	1985Q1-2018Q1	BIS
Real effective exchange rate	Q	1985Q1-2018Q1	BIS
GDP per capita, constant prices	Q	1985Q1-2018Q1	OECD
GDP per hour worked, constant prices	Q	1985Q1-2018Q1	OECD
GDP per person employed, constant prices	Q	1985Q1-2018Q1	OECD
Price-to-rent ratio	Q	1985Q1-2019Q1	OECD
Price-to-income ratio	Q	1985Q1-2019Q1	OECD
Banking credit to private sector	Q	1985Q1-2019Q1	BIS
Total credit to households	Q	1985Q1-2019Q1	BIS
Total Credit to private non-financial sector	Q	1985Q1-2019Q1	BIS
Total credit to non-financial firms	Q	1985Q1-2019Q1	BIS
Debt Service Ratio (Households)	Q	1985Q1-2016Q1	BIS
Debt Service Ratio (non-financial corporations)	Q	1985Q1-2017Q4	BIS
Debt Service Ratio (private non-financial sector)	Q	1985Q1-2017Q4	BIS
Consumer prices	Q	1985Q1-2019Q2	OECD
Monetary aggregate M3	Q	1985Q1-2018Q1	OECD
Real estate prices	Q	1985Q1-2019Q1	BIS
Share prices	Q	1985Q1-2019Q3	OECD
Unemployment rate	Q	1985Q1-2019Q3	GFD
Current account	Q	1985Q1-2019Q3	OECD
Rent Price Index	Q	1985Q1-2019Q3	OECD
Gross domestic product - expenditure approach	Q	1985Q1-2019Q1	OECD
Loans for House Purchasing	Q	###### - ######	OECD
Long-term interest rates (10Y)	Q	1985Q1-2019Q3	Datastream

Short-term interest rate (3M)	Q	1985Q1-2019Q3	Datastream
Slope of the yield curve (10Y - 3M)	Q	1985Q1-2019Q4	Datastream
Household Debt	Q	######-######	OECD
Equity holdings	Q	1985Q1-2019Q2	CrossBorder Capital
Financial assets	Q	1985Q1-2019Q2	CrossBorder Capital
Oil price	Q	1985Q1-2019Q2	CrossBorder Capital
Shipping indicator	Q	1985Q1-2019Q2	CrossBorder Capital
Golden rule	Q	1985Q1-2018Q4	built
VIX	Q	1990Q1-2019Q3	FRED
Export growth	Q	1985Q1-2019Q2	OECD
Import growth	Q	1985Q1-2019Q2	OECD
Terms of trade	Q	1985Q1-2019Q2	OECD
Growth of foreign exchange reserves	Q	1985Q1-2019Q2	OECD
External debt	Q	######-2019Q1	BIS
Multifactor productivity	A	1985-2017	OECD
General Goverrnment Debt	A	1985-2019	AMECO
Financial Conditions Index	Q	1985Q1-2019Q2	CrossBorder Capital
Risk Appetite	Q	1985Q1-2019Q2	CrossBorder Capital
Cross-border flows	Q	1985Q1-2019Q2	CrossBorder Capital

Economic Political Uncertainty Index	M	######-2019M9	PolUncertainty
Consumption Investment	Q Q	1985Q1-2019Q1 1985Q1-2019Q2	OECD OECD
GDP	Q	1985Q1-2019Q2	OECD
Global Factor	Q	1985Q1-2019Q2	Miranda-Agrippino, Rey
Housing 1 Forecast	Q	1985Q1-2019Q2	FED
Housing 2 Forecast	Q	1985Q1-2019Q3	FED
Domestic Liquidity Stock	Q	1985Q1-2019Q2	CrossBorder Capital
Policy Liquidity Index	Q	1985Q1-2019Q2	CrossBorder Capital
Domestic Liquidity Index	Q	1985Q1-2019Q2	CrossBorder Capital
Private Liquidity Index	Q	1985Q1-2019Q2	CrossBorder Capital
Quantity Liquidity Index	Q	1985Q1-2019Q2	CrossBorder Capital
Total Liquidity Index	Q	1985Q1-2019Q2	CrossBorder Capital
Policy Liquidity Stock	Q	1985Q1-2019Q2	CrossBorder Capital
Policy Liquidity Flows	Q	1985Q1-2019Q2	CrossBorder Capital
Total Liquidity Stock	Q	1985Q1-2019Q2	CrossBorder Capital

Total Liquidity Flows	Q	1985Q1-2019Q2	CrossBorder Capital
Central Bank Intervention	Q	1985Q1-2019Q2	CrossBorder Capital
Total Liquidity Stock	Q	1985Q1-2019Q2	CrossBorder Capital
Total Liquidity Flows	Q	1985Q1-2019Q2	CrossBorder Capital
Central Bank Intervention	Q	1985Q1-2019Q2	CrossBorder Capital
Financial Assets	Q	1985Q1-2019Q2	CrossBorder Capital
Fixed Income Holdings	Q	1985Q1-2019Q2	CrossBorder Capital
Equity Holdings	Q	1985Q1-2019Q2	CrossBorder Capital
Risk Appetite	Q	1985Q1-2019Q2	CrossBorder Capital
Private Sector Liquidity	Q	1985Q1-2019Q2	CrossBorder Capital
Gross Capital Flows	Q	1985Q1-2019Q2	CrossBorder Capital
Momentum	Q	1985Q1-2019Q2	CrossBorder Capital
Monetized Savings Index	Q	1985Q1-2019Q2	CrossBorder Capital

Bond Exposure Index	Q	1985Q1-2019Q2	CrossBorder Capital
Currency Exposure Index	Q	1985Q1-2019Q2	CrossBorder Capital
Exposure Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Financing Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Forex Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Composite Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Exposure Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Financing Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Forex Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital
Composite Risk Index	Q	1985Q1-2019Q2	CrossBorder Capital

Estimation periods

Samples are defined so that the batch sample contains one pre-crisis period and the online sample has enough observations according to data availability. QRT refers to quai real time. RT refers to real time. 3y is 3-year ahead. 2y is 2-year ahead.

Country	Batch sample	Online sample
France - QRT - 3y	1987q3 - 2001q2	2001q3 - 2019q3
France - QRT - 2y	1987q3 - 2001q2	2002q3 - 2019q3
France - RT - 3y	1987q3 - 2002q1	2002q2 - 2019q3
UK - QRT - 3y	1987q3 - 2000q1	2000q2 - 2019q3
UK - QRT - 2y	1987q3 - 2000q1	2000q2 - 2019q3
UK - RT - 3y	1987q3 - 2001q1	2001q2 - 2019q3
Germany - QRT - 3y	1987q3 - 2001q2	2001q3 - 2019q3
Italy - QRT - 3y	1987q3 - 2003q4	2004q1 - 2019q3

B Experts: quasi-real time data

We report whether experts are **Generic** experts (same specification for all the countries) or whether the specifications are country specific because variables have been selected via country specific AUROC. In that case, the country specification is reported below the main expert list.[46] We have a total of 26 experts.

B.1 Experts from the literature

Our first set of experts are taken from the economic literature on macroprudential policies:

1. **Expert P1**. Dynamic Probit Model: variables selected with a country-specific AUROC on the batch sample.

2. **Expert P2**. Panel logit fixed effect: variables selected with a country-specific PCA Analysis on the batch sample.

3. **Expert P3 Generic**: Panel logit fixed effect. We follow the literature and use the following specification: Banking credit to private sector gap-to-trend [47]; Banking credit to private sector 1y change; Real GDP 1y change; Consumer Prices; Share Prices 1y change; Rent Price Index 1y change; Banking credit to private sector gap-to-trend (global[48]); Banking credit to private sector 1y change (global); Real GDP 1y change (global); Consumer Prices (global); Share Prices 1y change (global); Interaction: Banking credit to private sector gap-to-trend (global)*Banking credit to private sector 1y change; Interaction : Banking credit to private sector gap-to-trend (global)* Banking credit to private sector gap-to-trend; Interaction: Banking credit to private sector 1y change * Banking credit to private sector 1y change (global).

4. **Expert BMA**: Bayesian Model Averaging. Variables selected with a country-specific AUROC on the batch sample.

B.2 Experts from Machine Learning

Our second set of experts come from the Machine Learning literature:

1. **Expert GAM**: General Additive Model

[46] 1-year change, 2-year and 3-year change are also included for each variable.

[47] Trend is computed with hp filter (1600) on the batch sample, and extrapolated with ARIMA forecasts for the online sample.

[48] Global variables are a simple average of the variable.

- Generalized additive models (**GAM**) provide a general framework for extending a standard linear model by allowing non-linear functions of each of the variables, while maintaining additivity. We consider here a General Additive Model such as :

$$y_t = \beta_0 + f_1(x_{1,t}) + f_2(x_{2,t}) + f_2(x_{12t})$$

The model is fitted with smoothing splines [see Hastie and Tibshirani (1986)].

2. **Expert RF**: Random Forest

- A random forest (**RF**) consists in three steps :

i) Build a number of decisions trees on bootstrapped training samples. ii) Each time a split in a tree is considered, a random sample of m predictors is chosen as split candidate. iii) Aggregate the prediction of each tree.

3. **Expert SVM**: Support Vector Machine

- A Support Vector Machine (**SVM**) expert classifies observations by constructing a hyperplane which has the largest distance to the nearest training-data point of any class. The aim is to find the separating hyperplane that is farthest from the data, that is to say which experiences the smallest perpendicular distance from each training observation, i.e. the smallest margin. In case of non-linear separable data, SVM extends the methodology used in a support vector classifier by enlarging the feature space using kernels. Indeed, a kernel function transforms the data into a higher dimensional feature space to make it possible to perform a linear separation.

Our basic Support Vector Machine (SVM) works in three steps :

i) Choose an optimal hyperplane which maximizes margin. ii) Applies penalty for misclassification. Indeed, a cost function specifies the cost of a violation to the margin. When the cost argument is small, then the margins will be wide and many support vectors will be on the margin or will violate the margin. When the cost argument is large, then the margins will be narrow and there will be few support vectors on the margin or violating the margin.This cost function is fitted using a grid on the batch sample. iii) If non-linearly separable data points, transform data to high dimensional space where it is easier to classify with linear decision surfaces. We use here a radial kernel. For more details, see Zhang (2012).

B.3 Experts Elastic-net Logits by themes

Our third set of experts are regularized logistic regressions. All the regularized regressions include each variable in level as well as the 1-year change, the 2-year change and the 3-y change.

Let's recall that Im is the pre-crisis indicator taking values in $G = 0, 1$. Let $p(x_i) = Pr(Im = 1|x_i) = \frac{1}{1+e^{-(\beta_0 + x_i\beta_i)}}$ be the probability for observation i at a particular value for the parameters (β_0, β). We solve :

$$\min_{(\beta_0,\beta)\in\mathcal{R}^{p+1}} \left\{ \frac{1}{N} \sum_{i=1}^{N} I(y_i = 1)log(p(x_i)) + I(y_i = 0)log(1 - p(x_i)) - \lambda P_\alpha(\beta) \right\}$$

where the elastic-net penalty is determined by the value of α :

$$P_\alpha(\beta) = \sum_{j=1}^{p} [\frac{1}{2}(1 - \alpha)\beta_j^2 + \alpha|\beta_j|]$$

$P_\alpha(\beta)$ is the elastic-net penalty term and is a compromise between the Ridge regression ($\alpha = 0$) and the Lasso penalty ($\alpha = 1$) and p is the number of paramaters. Whereas Lasso is indifferent to correlated predictors, the Ridge regression shrinks the coefficient of correlated predictors toward zero. Following Addo et al. (2018) and since there is a risk of correlation among our predictors, we pick $\alpha = 0.7$. We estimate the log-likelihood by applying the Newton Algorithm as in Friedman et al. (2010). We also estimate an optimal value of λ using 10-folds cross validation[49]. The folds are randomly selected and the results could face a variability issue. To reduce the randomness without increasing considerably the computation time, we run the cross-validation 50 times and average the error curves.

First introduced by Zou and Hastie (2005), the good performance of elastic-net penalty compared to other regularization methods has been confirmed in various applications (Mol et al. (2009); Mol et al. (2009); Destrero et al. (2009)). This is mainly due to the fact that, because it uses a penalty that is part ℓ_1 and part ℓ_2, this procedure works almost as well as the Lasso when the Lasso does best; but it also improves on the LASSO when the LASSO is dominated by the Ridge regression. This is usually the case if there exists high correlations among predictors, as in our case when we consider a large set of macroeconomic indicators (Tibshirani (1996)). As a consequence, the elastic-net penalty outperforms LASSO while preserving the sparse property (Zou and Hastie (2005); Mol et al. (2009)).

Five regularized regressions, called the "logit combination" experts, include variables which are selected on the batch sample thanks to the following procedure:

[49]To decrease the computation time, we use 5-folds cross validation for France QRT 2-years and France RT 3-years and 7-folds cross validation for UK QRT 2 years, UK QRT 3 years, Germany QRT 3 years.

1. The variables are selected thanks to an AUROC procedure performed on the batch sample, following Schularick and Taylor (2012) and Coudert and Idier (2016). We retain variables with an AUROC above 0.8.

2. Adding too many variables could decrease the forecasting ability. In our case, 3 to 12 variables are included (they correspond to 12 to 48 variables since we always include 1y, 2y and 3y transformations). If several variables have a large AUROC, i.e. superior to 0.8, more variables will be included in the logit combinations. For instance, for the case France 3y QRT, 65 variables have an AUROC greater than 0.8 (only 29 for the case France 2y RT). To decrease the risk of overfitting, we include one or two models with few variables (3 to 4).

3. There is only one pre-crisis to select variables. We do not include several similar variables (for instance GDP and its transformations) and apply the same PCA procedure used for the expert P2 if the AUROC procedure does not select one category of variables.

- The following experts are **Generic**:

  1. **Expert Lre** Logit real economy: GDP; GDP per person; GDP per hours work; unemployment rate; import, export, public debt.
  2. **Expert Lre2** Logit real economy 2: consumer prices; unemployment rate; GDP per person, GDP per hours work; GDP per capita; public debt; consumption; investment.
  3. **Expert Lval** Logit valuation: Share Price Index; Real Estate Price; Global Factor in Asset Prices; Short-term interest rate; Long-term interest rate; Dollar effective exchange rate.
  4. **Expert Lfor** Logit foreign: Cross Border Flows; Real Effective Exchange Rate; Dollar Effective Exchange Rate; Current Account; Terms of Trade.
  5. **Expert Lba** Logit bank: Risk Appetite; Share price Index; Equity holdings; Total Liquidity Index.
  6. **Expert Lcr** Logit credit: Total credit to non-financial sector; Banking Credit to non-financial sector; Total Credit to Households; Total Credit to non-financial corporations.
  7. **Expert Lbis** Logit BIS: Logit credit + DSR Households; DSR Non Financial corporations; DSR Total.
  8. **Expert Lm** Logit monetary: M3; Short-term interest rate; Long-term interest rate; Consumer Prices; Slope of the Yield Curve.

9. **Expert Lho** Logit housing: Price-to-rent; Price to income; Rent Price Index; Real Estate Price.
10. **Expert Lfgo** Logit Foreign Global: Logit Foreign + Global Factor in Asset Prices.
11. **Expert Lfgho** Logit Foreign Global + Housing.
12. **Expert Lhore** Logit housing + real economy.
13. **Expert Lbfo** Logit bank + foreign.
14. **Expert Lrisk** Logit Risk: VXO, Risk Appetite; Equity Holdings.

- We then have 5 Logits elastic net which are **country-specific** combinations. **Expert Lc1** to **Expert Lc5**. They are obtained by using the variables with the highest AUROC for a given country on the batch sample thanks to the procedure described above.

B.4 Variables for quasi-real time experts

Country-specific selected variables for each expert :

1. **France** :
   - P1 : Real Estate Price (2y), GDP per person (2y), Price-to-rent (2y), Banking Credit to private non-financial sector (2y).
   - P2 : Unemployment Rate, Rent Price Index, Loans, Dollar Effective Exchange Rate, Domestic Liquidity Stock
   - BMA : GDP (2y), Price-to-rent (2y), Banking Credit to private non-financial sector (2y)
   - GAM : Real Estate Price (2y)
   - Lc1 : Price-to-rent, Price-to-income, Real Estate Price, GDP, Oil Price (with 1y and 2y change).
   - Lc2 : Banking Credit to private non-financial sector (+ gap to trend), Total Credit to non-financial corporations (+ gap to trend), Total Credit to private non-financial sector (+ gap to trend), Total Credit to Households(+ gap to trend), Risk Appetite, EquityHoldings (with 1y and 2y change).
   - Lc3 : Risk Appetite, Cross Border Flows , Total Liquidity Index , Liquid Assets (with 1y and 2y change).
   - Lc4 : Real Estate Price, GDP, Total Credit to Households, Rent Price Index, Banking Credit to private non-financial sector, Price to income, Investment (with 1y and 2y change).

- Lc5 : Short-term interest rate, Price to rent, Terms of Trade, Housing 2 forecast , Total Credit to household , Total Credit to non-financial Corporation, Rent Price Index, Banking Credit to non-financial sector, Investment (with 1y and 2y change).

2. **UK** :

   - P1 : Price-to-rent, Total Credit to private non-financial sector (2y), Multifactor productivity (1y), GDP per hour worked (2y)
   - P2 : Loans (2y), Price-to-income, Banking Credit to private non-financial sector (2y), Total Credit to households (2y), Domestic Liquidity Stock (2y), Price-to-rent.
   - BMA : Price-to-rent, Total Credit to private non-financial sector (2y), Multifactor productivity (1y), loans (2y)
   - GAM : Long-term interest rate, Price-to-rent
   - Lc1 : Loans, Domestic Liquidity Stock, Liquid Assets, Total Credit to Households, Banking Credit to private non-financial sector, Total Credit to private non-financial sector.
   - Lc2 : Domestic Liquidity Stock, Dollar effective exchange rate, GDP, Multifactor Productivity, Slope of the yield curve.
   - Lc3 : Lc2 + Lfor.
   - Lc4 : Lc2 + Lho.
   - Lc5 : Lc2 + Lfgho.

3. **Germany** :

   - P1 : Public Debt, Equity Holdings, Banking Credit gap-to-trend, Long-term interest rate
   - P2 : Price-to-rent ratio, Rent Price Index, Loans, Banking Credit gap-to-trend, Banking Credit 2y change
   - BMA : Public Debt, Equity Holdings, Banking Credit gap-to-trend, Long-term interest rate
   - GAM : Public Debt
   - Lc1 : Price-to-rent, Total credit to non-financial sector, GDP per hour worked, Price-to-income, terms of trade, Risk Appetite .
   - Lc2 : Real Estate Price, Housing 1 survey of pro forecaster, Housing 2 survey of pro forecaster, Domestic Liquidity Stock , Short-term interest rate, Global Factor in Asset Prices, Total credit to Household.

- Lc3 : Housing 1 survey of pro forecaster, Housing 2 survey of pro forecaster, unemployment rate, Global Factor in Asset Prices, Real Estate Price;
- Lc4 : Price-to-rent, Investment, Housing 1 survey of pro forecaster, Housing 2 survey of pro forecaster, consumption, short term rate.
- Lc5 : Housing 1 survey of pro forecaster, Housing 2 survey of pro forecaster, total credit to households, unemployment rate, real estate price, banking credit to private non-financial sector.

4. **Italy**:

   - P1 : GDP (2y),Real Estate Price (1y), Price-to-rent (2y), Housing 2 forecast (2y)
   - P2 : Dollar effective exchange Rate, terms of trade, Rent Price Index , GDP, Public Debt.
   - BMA : GDP (2y), Price-to-rent (1y), Housing 2 forecast (2y), loan to income (2y).
   - GAM : GDP (2y).
   - Lc1 : Consumption, Investment, Housing 2, Total Credit to Households, Global Factor in Asset Prices.
   - Lc2 : Consumption, Investment, Housing 1,Housing 2, Total Credit to Households, Global Factor in Asset Prices.
   - Lc3 : GDP , Housing 1,Housing 2, Total Credit to Households, Global Factor in Asset Prices.
   - Lc4 : Consumption, Investment, Housing 1,Housing 2, Total Credit to Households, Global Factor in Asset Prices, Dollar Effective Exchange Rate, Real Effective Exchange Rate, Terms of Trace.
   - Lc5 : Price-to-rent, Housing 1,Housing 2, Total Credit to Households, Total Credit to private non-financial sector, Global Factor in Asset Prices, Dollar Effective Exchange Rate, Real Effective Exchange Rate, Terms of Trade.

5. **France** (2 years pre-crisis period) :

   - P1 : Real Estate Price (2y), GDP (2y), Short-term interest rate (2y), Cross Border Flows (1y).
   - P2 : Unemployment Rate, Terms of Trade, Dollar Effective Exchange Rate, Public Debt.
   - BMA : Real Estate Price (2y), GDP (2y), short term rate (2y), Cross border flows (1y).
   - GAM : Real Estate Price (2y).

- Lc1 : Price-to-rent, Price-to-income, Real Estate Price, GDP, Oil Price, current account, real effective exchange rate, equity holdings.
- Lc2 : Logit Housing + Logit real economy.
- Lc3 : Risk Appetite, Cross Border Flows, Total Liquidity Index, Liquid Assets.
- Lc4 : Real Estate Price, GDP, Total Credit to Households, Rent Price Index, loans, Banking Credit to private non-financial sector, Price to income, Investment, share price index, equity holdings.
- Lc5 : Short-term interest rate, Price to rent, Terms of Trade, Housing 2 forecast, Total Credit to household, Total Credit to non-financial Corporation, Rent Price Index, Investment, share price index, equity holdings.

6. **United Kingdom** (2 years pre-crisis period) :

- Logit combination 1 (Lc1) : Price-to-rent, Price-to-income, HOUSING 1, HOUSING 2, Loans.
- Logit combination 2 (Lc2) : Price-to-rent, Share Price Index, HOUSING 2.
- Logit combination 3 (Lc3) : Share price Index, Slope of the Yield Curve, Multifactor productivity.
- Logit combination 4 (Lc4) : GDP per capita, Total Credit to private non-financial sector, Multifactor Productiivty.
- Logit combination 5 (Lc5) : Price-to-rent, Price-to-income, Global Capital Factor, Short-term interest rate.
- P1 : Public Debt, Multifactor producitivty (2y), Long-term interest rate, Short-term interest rate.
- P2 : Loans (2y), Banking Credit to private non-financial sector (2y), Price to-income, Total Credit to households (2y), Domestic Liquidity Stock (2y), Price-to-rent.
- BMA : Public Debt, Terms of Trade (1y), Long-term interest rate, Short-term interest rate.
- GAM : Public debt.

C Experts: Real time data

Generic experts[50] :

[50] 1-year change 2-year and 3-year change are also included for each variable.

- **P3** : Private Sector Liquidity stock (1y), Domestic Sector Liquidity stock (1y), Share Price Index (1y), Private Sector Liquidity stock (gap-to-trend), Domestic Sector Liquidity stock (gap-to-trend) (global and country-specific variables.
- **Lli** : Total Liquidity Stock, Total Liquidity Flows, Domestic Liquidity Flows, Domestic Liquidity Stock, Domestic Liquidity Stock (local), Private Sector Liquidity Stock, Private Sector Liquidity Flows.
- **Lm** : Monetized Saving Index, Short-term interest rate , Long term interest rate.
- **Lrisk** : Share Price Index,Equity Exposure Index, Composite Risk Index, Financing risk Index, Risk Appetite.
- **Lfor** : Cross Border Flows, Dollar Effective Exchange Rate, Gross Capital Flows, Total Liquidity Flows.

Country-specific selected variables for each expert :

1. **France**
    - P1 : Short-term interest rate (2y), Private Sector Liquidity stock, Domestic Sector Liquidity stock (gap-to-trend), Total Liquidity Stock, Risk Appetite.
    - P2 : Quantity Liquidity Index, Total Liquidity Index, Financing RIsk Index, Quantity Liquidity Index (2y), Policy Liquidity Index.
    - BMA : Private Sector Liquidity stock (local), Private Sector Liquidity stock, Domestic liquidity stock (local), Domestic liquidity stock (gap-to-trend).
    - GAM : Domestic Sector Liquidity, Private Sector Liquidity stock, Domestic Sector Liquidity stock (gap-to-trend), Total Liquidity Stock, Risk.
    - Lc1 : Financial Condition index, Private Sector Liquidity Stock, Exposure Risk Index, Risk Appetite.
    - Lc2 : Financial Condition Index, Private Sector Liquidity Stock, Exposure Risk Index, Risk Appetite + Logit liquidity.
    - Lc3 : Monetized Saving, Short-term interest rate, Long-term interest rate, Private Sector Liquidity (local).
    - Lc4 : Monetized Saving, Short-term interest rate, Long-term interest rate, Cross border flows.
    - Lc5 : Monetized Saving, Short-term interest rate, Long-term interest rate, Private Sector Liquidity.

- Lc6 : Monetized Saving, Short-term interest rate, Long-term interest rate, Financing Risk Index.
- Lc7 : Monetized Saving, Short-term interest rate, Long-term interest rate, Domestic Sector Liquidity (gap), Private Sector Liquidity (gap).
- Lc8 : Financial Condition Index, Momentum, Private Sector Liquidity, Exposure Risk Index, Gross Capital Flows .

2. **UK**:

- Lc1 : Financial Condition index, Private Sector Liquidity Stock, Exposure Risk Index, Risk Appetite, Momentum.
- Lc2 : Private Sector Liquidity Stock, Domestic Liquidity Stock (local), Short-term interest rate, Long-term interest rate, Private Sector Liquidity Stock (local).
- Lc3 : Lc2 + logit risk.
- Lc4 : Financial Condition index, Private Sector Liquidity Stock, Exposure Risk Index, Total Liquidity Stock.
- Lc5 : Lc4 + Logit monetary.
- Lc6 : Logit Liquidity + Logit foreign.
- Lc7 : Lc3 + Lc4.
- P1 : Private Sector Liquidity Stock (2y), Domestic Liquidity Stock local (2y), Short-term interest rate, Long-term interest rate.
- P2 : Private Sector Liquidity (gap), Domestic Liquidity Sector (gap), Private Sector Liquidity, Domestic Liquidity Stock (local), Long-term interest rate, Short-temr interest ratre.
- BMA : Private Sector Liquidity Stock (2y), Domestic Liquidity Stock local (2y), Short-term interest rate, Long-term interest rate.
- GAM : Dollar effective exchange rate, Private Sector Liquidity Stock (2y), Domestic Liquidity Stock local (2Y).

D Aggregation rules

The fixed-share online aggregation rule[51] is similar to the EWA aggregation rule, except that we now consider a mixed rate $\alpha \in [0, 1]$. At each time instance, we include a small probability ffj

[51]Each aggregation is computed here with a delayed feedback and with a non-uniform weight vector.

Algorithm 1 Prediction with expert advice

1. The expert advice $\{f_{j,t} \in \mathcal{D} : j \in \mathcal{E}\}$ based on information until date t-1 is revealed to the forecaster.

2. The forecaster makes the prediction $\hat{y_t} \in \mathcal{D}$, based on information available at date t-1 and a sequential aggregation rule $\mathcal{S}$.

3. The t^{th} observation y_t is revealed.

4. The forecaster and each expert respectively incur loss $\ell(\hat{y_t}, y_t)$ and $\ell(f_{j,t}, y_t)$.

Algorithm 2 Prediction with expert advice with delayed feedback

1. The expert advice $\{f_{j,t} \in \mathcal{D} : j \in \mathcal{E}\}$ is revealed to the forecaster.

2. The forecaster makes the prediction $\hat{y_t} \in \mathcal{D}$.

3. The *t-12*th observation y_t is revealed.

4. The forecaster and each expert respectively incurs loss $\ell(y_{t}\hat{-}_{12}, y_{t-12})$ and $\ell(f_{j,t-12}, y_{t-12})$.

Algorithm 3 Gradient-based EWA

1. Parameter : Choose the learning rate $\eta_t > 0$.

2. Initialization : p_1 is the first uniform weight, $p_{j,1} = \frac{1}{N} \forall j \in \{1, ..N\}$.

3. For time instances $t = 2, 3, ..., T$ the weights vector p_t is defined by :

$$p_{j,t} = \frac{exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{j,s}})}{\sum_{k=1}^{N} exp(-\eta_t \sum_{s=1}^{t-1} \tilde{L_{k,s}})}$$

   where $\tilde{L_{j,s}} = \nabla \ell(\sum_{k=1}^{N} p_{k,s} f_{k,s}, y_s) \cdot f_{j,s}$

to have a m possibility of shifts in the sequence so that the best expert may change. We denote by $E_t \subset 1, ..., N$ the set of active experts at a given time instance t and assume that it is always non-empty. We define the fixed-share aggregation rule strategy $\mathcal{F}_{\eta,\alpha}$:

Algorithm 4 Fixed-share aggregation rule

1. *Parameter :* Choose the learning rate $\eta_t > 0$ and a mixing rate $\alpha \in [0, 1]$
2. *Initialization :* $(w_{1,0}, ..., w_{N,0}) = \frac{1}{|E_1|}(I_{1\in E_1}, ..., I_{N\in E_1})$.
3. For each round $t = 1, 2, ..., T$:
   (a) predict $\hat{y}_t = \frac{1}{\sum_{k=1}^{N} w_{k,t-1}} \sum_{j=1}^{N} w_{j,t-1} f_{j,t}$
   (b) (loss update) observe y_t and define for each $i = 1, ..., N$: $v_{i,t} = w_{i,t} e^{-\eta \tilde{L_{i,t}}}$
   (c) (share update) $w_{j,t} = \frac{1}{|E_{t+1}|} \sum_{i}^{N} v_{i,t} + \frac{\alpha}{|E_{t+1}|} \sum_{i\in E_t \cap E_{t+1}} v_{i,t}$ + (1 - α)$I_{j\in E_t\cap E_{t+1}} v_{j,t}$

Theorem 2 (Devaine et al. (2013)) *Consider the same assumptions than for the EWA aggregation rule. Then for all* $m \in \{0, ..., T-1\}$

$$\sup\{R_T(\mathcal{F}_{\eta,\alpha})\} \leq \frac{m+1}{\eta} ln(N) + \frac{1}{\eta} ln(\frac{1}{\alpha^m \alpha^{T-m-1}}) + \frac{\eta}{2} T \tag{1}$$

η is calibrated as in the EWA aggregation rule, α is calibrated online using the same methodology :

$$\alpha_t \in \arg\min_{\alpha > 0} \hat{L}_{t-1}(\mathcal{F}_{\eta,\alpha})$$

For the moment, we have restrained our analysis to convex aggregation rules, where the weight vector p_t is choosen in a simplex $\mathcal{P}$. These strategies, usually referred to as *Follow-the-leader*, aim at minimizing the cumulative loss on all past rounds. *Follow-the-Regularized-Leader* strategies add a slight modification. The forecaster minimizes the cumulative loss function plus a regularization term. The weights do not need to be choosen in a convex space since the regularization term stabilizes the solution.

Consider the case where the regularized term is a linear function. The aggregation rule $\mathcal{OGD}_\eta$, for Online Gradient Descent (OGD), was firstly introduced by Zinkevich (2003). It updates parameters by taking a step in the direction of the gradient. Define $||x|| = \sqrt{x \cdot x}$ and $d(x, y) = ||x - y||$. The weight vector p_{t+1} is selected according to :

$$p_{j,t+1} = P_j(p_{j,t} - \eta_t \partial \ell(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t))$$

where $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t ||$

Algorithm 5 Online-Gradient Descent aggregation rule

1. *Parameter :* Choose the learning rate $\eta_t > 0$
2. *Initialization :* an arbitrary vector p_1.
3. For each round $t = 1, 2, ..., T$, the vector p_{t+1} is selected according to :

$$p_{j,t+1} = P_j(p_{j,t} - \eta_t \partial \ell(\sum_{j=1}^{N} p_{j,t} f_{j,t}, y_t))$$

where $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t ||$

As for the strategy $\mathcal{E}_\eta^{grad}$, the strategy $\mathcal{OGD}_\eta$ satisfies our robustness requirement. The following bound was first established by Zinkevich (2003) :

Theorem 2. *If $\eta_t = t^{-\frac{1}{2}}$, the regret is bounded by:*

$$\sup\{R_T(\mathcal{OGD}_\eta)\} \leq \frac{1}{2}(3\sqrt{T} - 1) \tag{2}$$

Consider now the case where the regularized term is the square-ℓ_2-norm regularization, often called the Ridge aggregation rule $\mathcal{R}_\eta$. The Ridge aggregation rule minimizes at each time instance a penalized criterion. Hence this aggregation rule can be useful if the experts are correlated, which is probably the case in our exercise. For this aggregation rule, only the square loss is considered. Note that the Ridge aggregation rule is theoretically the most robust strategies for the forecaster. Indeed, it competes not only with the best expert or the best combination of experts, but with the best combination of experts with some sub-linear shifts.

The weight vector $p_t = (p_{1,t}, ..., p_{N,t})$ is given by :

$$p_t \in \arg\min_{v \in \mathbb{R}^N} \left\{ \lambda ||v||_2^2 + \sum_{s=1}^{t=1} (y_s - \sum_{j=1}^{N} v_j f_{j,s})^2 \right\}$$

where the tuning parameter λ is calibrated online, as the learning rate η

As for strategies $\mathcal{E}_\eta^{grad}$ and $\mathcal{OGD}_\eta$, the strategy $\mathcal{R}_\eta$ satisfies our robustness requirement. This theorem is stated by Cesa-Bianchi and Lugosi (2006) and Stoltz (2010) :

Algorithm 6 Ridge aggregation rule

1. *Parameter :* Choose the learning rate $\eta_t > 0$
2. *Initialization :* an uniform vector p_1.
3. For each round $t = 2, ..., T$, the vector p_t is selected according to :

$$p_t \in \arg\min_{v \in \mathbb{R}^N} \left\{ \lambda ||v||_2^2 + \sum_{s=1}^{t=1} (y_s - \sum_{j=1}^{N} v_j f_{j,s})^2 \right\}$$

Theorem 3. *Since* $\hat{y}_t \in [0, 1]$:

$$\sup\{R_T(\mathcal{E}_\eta^{grad})\} \leq \frac{ln(N)}{\eta} + \eta \frac{T}{2} \tag{3}$$

For every aggregation rule; the learning rate calibration is critical to guarantee algorithm's adaptation capacity, especially in a small sample. Theoretically, Cesa-Bianchi and Lugosi (2006) show that the following choice of η_t :

$$\eta_t = \frac{8ln(N)}{t-1}$$

gives the optimal bound $\sqrt{2Tln(N)} + \frac{lnN}{8}$. However, Stoltz (2010) shows that theoretical values for the learning rate give bad performance. Excet for the OGD aggregation, which is theoretically calibrated following Zinkevich (2003), the forecaster chooses the parameter η_t with the best past performance :

$$\eta_t \in \arg\min_{\eta > 0} \hat{L}_{t-1}(\mathcal{E}_\eta)$$

More precisely, following Amat et al. (2018), at each round, the forecaster chooses $\eta_t > 0$ in the grid $\frac{\mathcal{G}}{\sqrt{t}}$ where :

$$\mathcal{G} = \left\{ m \times 10^k, m \in \{1, 2, 5\} \text{ and } k \in \{-4, -3, -2, -1\} \right\} \cup \{1\}$$

D.1 Aggregation rules with delayed feedback

We modify the standard set up to account for the fact that the forecaster learns about a pre-crisis period with a 12 quarter delay. Experts have to learn on a first crisis episode so for each country, we start the exercise at the end of a first crisis. The robustness theorems (finite bounds on the regret) for the EWA described above hold with uniform initial weights (OGD can start with any initial weights). When we start to train experts on a first crisis episode, we have information on experts' in-sample performances. It can be valuable to use this information to decrease the estimation error to increase experts' performances. But this could jeopardise the forecaster's capacity to converge towards the best combination of experts. We face the classic dilemma between estimation error and approximation error. Consider a vector of arbitrary initial weight $w_{1,0},...,w_{N,0} > 0$ and the EWA forecaster. Cesa-Bianchi and Lugosi (2006) state the following theorem:

Theorem 3. *Under the same conditions as in Theorem 1* :

$$R_T(\mathcal{E}_\eta^{grad}) \leq \min_{j=1,...N} \left\{ ln(\frac{1}{w_{j,0}})\frac{1}{\eta_t} \right\} + \frac{lnW_0}{\eta_t} + \eta_t \frac{T}{8} \tag{4}$$

For our EWA aggregation rules, weights are chosen in a simplex so that $W_0 = 1$ and $ln(\frac{1}{w_{j,0}})$=lnN. The increase in the approximation error due to non uniform weights seems in many relevant cases negligible compared to the decrease in the estimation error. Each aggregation rule is therefore performed under delayed feedback with non-uniform initial weights.