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

Jérémy Fouliard
London Business School

Michael Howell
CrossBorder Capital

Hélène Rey
London Business School

First draft: November 2018. Preliminary. This Version: October 16, 2019

Abstract

Financial crises cause economic, social and political havoc. 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 averaging and is "meta-statistic"since we can incorporate any predictive model of crises in our set of experts and test its ability to add information. We are able to predict systemic financial crises 12 quarters ahead in quasi-real time with very high signal to noise ratio. We also analyse which models and variables provide the most information for our predictions at each point in time, allowing us to gain some insights into economic mechanisms underlying the building of risk in economies. Finally we also show that our methodology is also able to predict systemic crises in France using real time data.

[*]We are highly indebted to Pierre Alquier for his generous help. We thank Gilles Stolz for extremely useful comments and multiple discussions. We also thank Nicolas Vayatis for his help. We are also very grateful to our discussants Fabrice Collard, Franck Diebold, Christian Julliard, Sebnem Kalemli Ozcan and Luc Laeven and to seminar participants at ENSAE, KTO-GREDEQ-OFCE, Nowcasting, the LBS AQR academic symposium, the Annual BIS Conference and the Women in Economics Chicago Conference for comments. Our thanks also go to Virginie Coudert, Julien Idier, Mikael Juselius, Marco Lo Duca, to Benoit Mojon and to Katja Taipalus for help with the database. Hélène Rey is grateful for the funding of the European Research Council (grant number 695722). 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 England 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, it took several months before the British Academy wrote a three-page missive to Her Majesty blaming the lack of foresight on the crisis on the "failure of the collective imagination of many bright people" and also pointing to 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". The letter mentions that "Everyone seemed to be doing their own job properly on its own merit. And according to standard measures of success, they were often doing it well. The failure was to see how collectively this added up to a series of interconnected imbalances over which no single authority had jurisdiction." There are many different models in macroeconomics and in finance which are used to understand financial crises. Some emphasise runs (Diamon Dybvig, Gorton). Very few analyze the boom phase of the financial cycle, the few who do emphasise limited liability and asset overvaluations due to risk-shifting (Coimbra and Rey), or deviations from rational expectations and financial constraints (Shleifer and Gennaioli). Many models focus on the bust phase of the crisis and on amplification mechanisms (Kiyotaki Gertler, etc...). From an empirical point of view, a number of variables have been used to predict financial crises (mostly in sample). From the classic paper of Kaminsky and Reinhart, numerous papers have very usefully described the behaviour of a number of key variables around crisis episodes (see e.g. Gourinchas and Obstfeld (2012) ....) More recently the work of Shularick and Taylor and Borio et al. emphasising the role of credit growth or credit to GDP gaps and the work of Mian and Sufi (2018) underlining the importance of household debt have been very influential in shaping our understanding of financial crises. Some recent attempts to introduce new forecasting methods such as decision tree and random forest can be found in Ward and Bluwstein et al. (2019). From a general econometric point of view Rossi discusses in detail in her Handbook Chapter the importance of accounting for instabilities in time series data when performing out-of-sample forecasting exercises. She also underlines the problem of over-fitting.

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) is still very limited. Turning points and non linear phenomena such as crises have been notoriously difficult to predict in quasi real time. Price based early warning indicators tend to be more coincident indicators than good predictors. Predicting pre-crisis periods (12 quarters before the crisis) in order to give macroprudential and other authorities the time to act proves to be extremely difficult. Yet financial stability policies need this type of input. From a theoretical point of view, there is no agreement on a workhorse model of crises; this may be a reflection of the fact that although crises have some common factors or symptoms -crises are often "credit booms gone bust" as described by Minsky and Kindleberger, they also display some differences in their mechanics. 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. We would like to forecast financial crises without knowing the ”true” model of the economy, using as much information as possible (in our case that means many possible models of the economy or ”experts”) in a way which is flexible enough to do dynamic evolving forecasting (weights put on different ”experts” should vary over time). Our contribution is to adapt the *framework of sequential prediction or online machine learning* to overcome some of these difficulties. This framework is perfectly appropriated for our problem. Since there is no consensus on a unique theory of financial crisis, it does not make assumptions on how the data are generated. Indeed online machine learning is specifically geared at quasi-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. It is then measured by how well the forecaster uses the available information to make his own prediction. This available information is composed by reference forecasters, also called

experts. This is why this approach can be described as "meta-statistic" since the aimis to make the best prediction by aggregating **experts**' predictions. Then, the forecaster's error is the sum of two errors : an estimation error, known expost, measured by the error of the best combination of experts, representing the best prediction the forecaster can make using the available information and an approximation error, relative to the difficulty to approach ex ante the ex post combination of experts. Though based on model averaging with time varying weights, on line learning is more general than bayesian model averaging; importantly it does not make any strong assumption on the data generating processes and allows for one learning rate (or more). In some cases, even very simple ones, (see Grunwald and Van Rommen (2017)) Bayesian Model averaging does not converge due to heteroskedasticity. We emphasize that our approach is not approximated Bayesian Model Averaging: it is more general. To our knowledge online machine learning has never been applied to economics (one exception is Stolz et al. for exchange rates) though it has been used in a number of applications outside economics, for example to predict French electricity consumption. 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 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 causes of the crisis.

We present our database on systemic crisis dates as well as the different variables which will be used to build our "experts" (predictive models) in section 2. In section 3, we describe the general methodology of sequential prediction 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 12 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 models (random forests, classification trees) to illustrate the power of our methodology. We assess predictive ability using four model aggregation rules and we present AUROC results. In all cases we uncover a time varying subset of models and variables which carry most of the information to predict financial crises. The quasi real time forecast of our online aggregators is usually very high and provides well-behaved signals for policy makers. Section 5 concludes.

2 Data on systemic crises and macroeconomic indicators

We need two types of data : the datation of systemic crisis episodes and a dataset of economic indicators for a range of countries in order to construct forecasting models (”experts”). Due to data availability the period under consideration is 1985q1 to 2018q1. We consider seven countries : France, Germany, Italy, Spain, Sweden UK and US. They include the largest eurozone economies, a small open economy and the two largest financial centres (US and UK).

2.1 Data on Systemic Crisis Episodes

We borrow the definition and the dates of systemic crises from the Official European database constructed by the ECB [Marco Lo Duca et al., 2017]. The date is partly based on quantitative indicators but is ultimately based on the expert judgement of the relevant national authorities. The methodology used is a two-step approach. Following Duprey et al. [2015], it aims at firstly identifying historical episodes of elevated financial stress which were also associated with real economic slowdowns using a quantitative analysis. The financial stress is measured by a financial stress indicator which captures three financial market segments : i) equity market : stock price index, ii) bond market : 10-year government yields and iii) foreign exchange market : real effective exchange rate (see more details in Appendix). Industrial production growth is used as measure of real economic activity. 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 financial stress period is drawn. The second step aims at using a qualitative approach. Each national authority distinguishes between systemic crisis 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). National authorities are also asked whether they want to complement the list of events or disagree with the timing of events already flagged. The database of crisis episodes is already available for European countries. We replicate the exact same methodology for the US.
We focus on predicting systemic crises twelve quarters ahead (pre-crises periods) in quasi real time.

We denote the characteristic function $C_{n,t}$:

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

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 Indicators

We consider a large set of macroeconomic indicators X_k. We take into account the main risks on financial markets, real estate markets, credit market, interest rates and macroeconomic conditions. Our database contains commonly used Early Warning Indicators (n=144) with transformations (1-y, 2-y, 3-y change and gap-to-trend). Whenever we detrend a variable we make sure we use only data of the estimation sample (and no future data). We make use of OECD's Main Econonomic indicators and National Accounts databases, the BIS and of the database of Cross Border Capital (CBC) which contains monthly data series on liquidity aggregates (public and private), capital flows and risk indices. Importantly these variables are available in revised format as well as in real time (see more details in Appendix).

- **Macroeconomic indicators** : Consumer Price Index, Unemployment rate, GDP, GDP per person employed, GDP per capita, GDP per hour worked, M3, Unemployment rate, Cross-border flows, Total Liquidity Index, Current account, General Goverrnment Debt, Golden rule (gap of real long term interest rate to real GDP), Economic Political Uncertainty Index

- **Credit 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), Banking credit to private sector, skewness of leverage, Household Debt, Bank assets, Bank equity.

- **Interest rates indicators** : 3-month rate, 10 years rate, slope of the yield curve (10y-3m).

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

- **Market indicators**: Real effective exchange rate, Share prices, Financial Conditions Index, Risk Appetite Index, oil price, Equity holdings, Financial assets.

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 $\mathcal{D}$.

This framework has two main specificities. First, the observations $y_1, y_2, ...$, are revealed in a sequential order. At each step $t = 1, 2, ..$, the forecaster makes a prediction $\hat{y}_t$ before the *t*th observation is revealed on the basis of the previous $t-1$ observations. This is why this approach is said to be "online" since the forecaster sequentially receives information. The 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, 1996, 2003 and Rossi]. Second, in contrast to the stochastic modelling approach, we do not assume that $y_1, y_2, ...$ are the product of a stationary stochastic process. The sequence $y_1, y_2, ...$ could be the result of any unknown mechanism which is in line with the fact that there is no consensus on a theory of financial crises.

The forecaster predicts the sequence $y_1, y_2, ...$ 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). We consider here a set of experts where each expert $j = 1, ..., N \in \mathcal{E}$ 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 dependant on the quality of our set of experts. If we put "garbage in", we will get "garbage out". The methodology of *online learning* is therefore extremely flexible and general as any forecasting model can be used to contribute to the optimal forecast. But of course there is no magic, if all forecasting models are bad, he optimal forecast will also be bad.

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

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.
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 $\ell(f_{j,t}, y_t)$.

How do we measure the sequential aggregation rule's performance ? If the sequence $y_1, y_2, ...$ were the realisation of a stationary stochastic process, it would be possible to estimate the risk 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 forecaster's loss over time 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 thus 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 convex 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 (date t-1) is close to the best combination of experts, which is known at the end of the round (date t). To get a robust aggregation rule, the forecaster wants, in addition of having the smallest bound possible for the regret, to obtain a "vanishing per-round regret" so that when T goes to infinity:

$$lim R(\mathcal{S})/T = 0$$

In this case, the forecaster's cumulative loss will converge to the loss of the best linear combination of experts known ex-post. Note that the evaluation of an aggregation rule is always relative to experts performances : if each expert makes a bad prediction, the forecaster's prediction will be bad. In other words, the famous "garbage in, garbage out" proposition applies. This approach can be described as "meta-statistic" 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 (known ex post)[1].

Whereas this approach is very popular in machine learning, most statistical and econometric research uses a "batch" framework, where one starts from estimating a model on a complete

[1]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 always fixed over time whereas forecasters strategy includes time-varying weights. Forecasters 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.

sample. For model averaging problems, one of the most popular ”batch” methodology in econometrics is the Bayesian Model Averaging (BMA) framework which uses Bayesian decision theory. It would be wrong to say that there is no link between Bayesian decision theory and the theory of sequential predictions [2]. 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 is a specific 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 [2017] 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 Deblasi 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; McAllester, 2003; Barron and Cover, 1991; Walker and Hjort, 2002; Zhang, 2006a]. We note that since online learning can be seen as a ”meta-statistic approach” (or a ”meta-algorithmic approach”), it can incorporate Bayes 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 classic framework of sequential predictions. In the classic framework previously described, the forecaster knows the true observation y_t at the end of the period t. After that, he incurs a loss and can update his weights.

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

In our case, this assumption is not valid anymore. Indeed, the pre-crisis period is an ex-post definition. When a 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 here 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 Wintenberg and Ordentlich[2002] and of Joulani and 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 $\hat{y}_t$. When $\tau_t = 0$, the general framework of sequential predictions does not change. In our case, τ is a constant which is equal to 12.

We define $R'(\mathcal{S})$ as the regret of the strategy $\mathcal{S}$ in a delayed setting. Wintenberg and Ordentlich show :

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

Introducing a delayed feedback increases the bound of the regret - the approximation error - but does not violate our robustness requirement.

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 $\ell(\hat{y_{t-12}}, y_{t-12})$ and $\ell(f_{j,t-12}, y_{t-12})$.

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.

3.3 Selecting aggregation rules

We only select robust aggregation rules, which compete with the best combination of experts (ex post). We consider several aggregation rules.

3.3.1 Exponentially weighted average aggregation rule

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

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

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 his past prediction. The forecaster predicts at each time t :

$$\hat{y}_t = \frac{\sum_{j=1}^{N} e^{-\eta_t L_{j,t-1}} f_{j,t}}{\sum_{i=1}^{N} e^{-\eta_t L_{i,t-1}}}$$

where η_t is the learning rate, the speed at which weights are updated.

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

An important advantage of the gradient-based version of the EWA aggregation rule is that 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.

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

The strategy $\mathcal{E}_\eta^{grad}$ competes with the best convex combination of experts. The following theorem is stated in Stoltz [2010]:

Theorem 1. *If $\mathcal{D} = [0, 1]$ is convex, $\mathcal{L}(\cdot, y)$ are differentiable on $\mathcal{D}$* and $\tilde{\mathcal{L}_{j,t}}$ *are in* $[0, 1]$, *for all* $\eta_t > 0$:

$$\sup\{R_T(\mathcal{E}_\eta^{grad})\} \leq \frac{ln(N)}{\eta_t} + \eta_t \frac{T}{2} \tag{1}$$

The strategy $\mathcal{E}_\eta^{grad}$ satisfies our robsutness requirement :

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

The bound of the regret depends on three parameters, two exogeneous and one endogenous. The number of experts N and the number of time instances T differ according to the way we design experts and the pre-crisis period we want to predict. An interesting property of the theorem is that the bound does not directly depend on the number of experts, but on the log of it. A large number of experts will not drastically increase the difference between forecaster's cumulative loss and the cumulative loss of the best combination of experts. This goes in favour of choosing a large number of experts.

The last parameter of the bound η_t is the learning rate. For the gradient-based EWA aggregation rule, the forecaster chooses the parameter η_t with the best past performance :

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

3.3.2 Online Gradient Descent aggregation rule

For the moment, we have restrained our analysis to convex aggregation rules, where the weight vector p_t is chosen in a simplex $\mathcal{P}$. These strategies, usually referred to as *Follow-the-leader*, aim at minimising the cumulative loss on all past rounds. *Follow-the-Regularized-Leader* strategies add a slight modification. The forecaster minimises the cumulative loss function plus a regularization term. The weights do not need to be chosen in a convex space since the regularization term stabilises 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 first 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 $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t||$

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:*

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

We note that the learning parameter η_t is calibrated (and not optimized upon as for the pre-

Algorithm 4 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 $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t ||$

vious EWA rule). For more details on online gradient descent and other aggregation rules such as Ridge, the reader is referred to the Appendix.

3.4 Aggregation rules with delayed feedback

As previously mentioned we have to modify the standard set up to account for the fact that the forecaster learns about a pre-crisis period with a 12 quarter delay. As we predict a binary variable, we cannot start the forecasting exercise at the beginning of the sample. Indeed, experts have to learn on a first crisis episode. 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. To what extent starting with non-uniform initial weights increases the 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* :

$$R_T(\mathcal{E}_{\eta}^{grad}) \leq \min_{j=1,...N} \left\{ ln(\frac{1}{w_{j,0}})\frac{1}{\eta_t} \right\} + \frac{lnW_0}{\eta_t} + \eta_t \frac{T}{8} \quad (3)$$

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. Note that this bound only holds for EWA, and not for the gradiend based version or other aggregation rules such as OGD or Ridge. We are working on another version of this theorem which will be available soon..

3.5 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 number of experts. Nevertheless, it does not directly increase with N but with $ln(N)$, so that it is often better to use a large number of experts.

We will pick different sets of experts in section 4: we pick both"off-the shelf" experts used in the literature and in Central Banks to predict financial crises as well as bayesian averaging models and machine learning models such as classification trees and random forests. The beauty of our approach is that we can include *any* type of experts and therefore be very oecumenical in terms of methodology.

4 Horse race between financial crisis models

We compare the out-of-sample performance of several models used by academics and by central banks in their effort to construct a set of early warning indicators for macro prudential policies. Many of the models were summarised by the Macro-prudential Research Network of the ECB. Some models are estimated on a panel, others are estimated country by country. We also make use of global averages. Therefore our experts incorporate informaion from the entire set of countries.

We add a series of models that we constructed ourselves (classification trees, random forests and logits with elastic net penalties). Note that our models incorporate various horizons of changes so that inflexion points can be captured.The models have been re-estimated with our variables on our sample. In a small number of cases we could not include one variable of the model as it was not publicly available. Our estimates can therefore in some case differ from the original estimates. We give here a list and very brief description of the 22 models we use in our forecasting exercise and refer the reader to the appendix for more details. We note that we could consider many more variables and models if we could imporve on data availability. We could also extend the country sample. The methodology is flexible enough to incorporate all these improvements.

- **Expert P1**: Dynamic Probit Model, Panel 1.

  Variables selected with a country-specific AUROC on the batch sample (see Appendix).

- **Expert P2**: Panel logit fixed effect, Panel 2.

  Variables selected with a country-specific PCA Analysis on the batch sample (see Appendix)

- **Expert P3**: Panel logit fixed effect, Panel 3.

  Banking credit to private sector gap-to-trend [3]; 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[4]); 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).

- **Expert P4**: Random Coefficient Logit, Panel 4. Share Price 1y change; Real GDP 1y change; Banking credit to private sector 1y change; Rent Price Index 1y change; Total Liquidity

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

[4]Global variable are a simple average of this variable for each country

Index 1y change; Financial Condition Index 1y change.

- **Expert B1**: Bayesian Model Averaging, BMA 1 (panel)

  To reduce the number of possible combinations, we pre-select indicators with a Panel Auroc analysis on the batch sample (see Appendix)

- **Expert B2**: Bayesian Model Averaging, BMA 2 (country-specific)

  To reduce the number of possible combinations, we pre-select indicators with a country-specific Auroc analysis on the batch sample. (see Appendix)

- **Expert T1**: Binary Classification Tree, (panel)

- **Expert T2**: Binary Classification Tree, (country-specific)

- **Expert R1**: Random Forests, (panel)

- **Expert R2**: Random Forests, (country-specific)

  We also introduce some logit with elastic-net penalty which are country-specific.

- **Expert Lh**: Logit with elastic-net penalty, Logit housing. Price-to-rent; Price to income; Price to rent 1y change; Price to income 1y change; Real estate price 1y change; Real estate price 2y change.

- **Expert Lr**: Logit with elastic-net penalty, Logit real economy. GDP (millions dollars 1y change 2 y change); GDP (Per capita per person per hour); Multifactor productivity; Oil price 1y change; Oil price 2 y change.

- **Expert Lcr**: Logit with elastic-net penalty, Logit credit Every credit variable with every transformation (1y, 2y, 3y change and gap to trend)

- **Expert Lba**: Logit with elastic-net penalty, Logit risk taking

  Risk Appetite; Financial Condition Index; Share price Index; Equity holding; Liquid Assets with two transformations: 1y change and 2y change.

- **Expert Lm**: Logit with elastic-net penalty, Logit monetary.

  M3; Short term interest rate (nominal); Short term interest rate (real); Consumer prices.

- **Expert Lbu**: Logit with elastic-net penalty, Logit bubble. Bubble indicators (Taipalus et al).

- **Expert Lc**: Logit with elastic-net penalty, Logit CrossBorder Capital (Risk appetite; financial condition index; total liquidity index (including capital flows); share price index.).

- **Expert B**: Lazy logit regression, Bashful

  14 best variables selected with a country-specific AUROC Analysis

- **Expert G**: Lazy logit regression, Grumpy

  All variables available in 1987q3.

- **Expert Lc1**: Logit with elastic-net penalty, Logit Combination 1

  Housing + Real Economy

- **Expert Lc2**: Logit with elastic-net penalty, Logit Combination 2

  Credit + Risk taking

- **Expert Lc3**: Logit with elastic-net penalty, Logit Combination 3

  Monetary + Crossborder capital variables: Risk appetite; financial condition index; total liquidity index (including capital flows); share price index.

We now have a total of 22 experts of all stripes and shapes including some models with common components, Baysian averaging, classification trees and random forests. Our models contain all the variables that have been shown to be important in the literature: credit (Shularick and Taylor; BIS credit to GDP gap); household debt (Mian and Sufi); many spreads and financial condition index variables (Krishnamurthy and Muir and Adrian et al) in particular. Our oecumenical approach can accommodate many more. Our only restriction is data availability (length of the time series). For example although it would be desirable to test the information content of a number of variables based on individual banks balance sheets, the timing of the first crisis and the 12 quarter lags means that in practice those variables cannot be incorporated in the analysis (yet).

5 Results

We present a series of results for France, Germany and the UK. Most of the literature focuses on in-sample results and attempts to predict crises (not pre crisis). We present results for out-of-sample pre-crisis prediction. Our exercise is a quasi real time exercise since we do not have the different data vintages to do real time forecasts. 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 exercise as an early warning indicator. If the signal tends to be monotonously increasing it is likely to be very useful as an early warning indicator. For each country we present in the main text our estimated probability of pre-crisis using the EWA aggregating rule. We show in Appendix results for the other rules (ML, OGD, Ridge). 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 as well as diagnostics of fit of our model (mean squared errors and AUROCs) both in the text and in appendix for the different aggregation rules.

5.1 France

Figure 1 presents the timing of pre-crises in France in light blue (12 quarters before the beginning of the crisis). The systemic crises are in dark grey. There are 2 systemic crises during the period 1985q1 to 2017q2 (the first one from 1991 q2 till1995 q1 and the second one from 2008 q1 to 2009 q4). There is also one residual event which we call the sovereign debt crisis (Figure 2) from 2011 q1 till 2013 q4. We estimate the expert models on sample 1987Q3-2000Q2 (or shorter depending on availability of the experts on the whole sample; 1987Q3 is the earliest possible date we can start because of data availability) a period during which France experienced the first systemic crisis.

The algorithm learns therefore from the first systemic crisis episode. The 1991q2-1995 q1 French systemic crisis was linked to real estate in particular. France experienced a period of a high GDP growth and deregulation from 1987 to 1990, fuelling an increase in both residential and commercial real estate prices. The sharp increase in oil prices triggered a severe macroeconomic slowdown from 1990 Q2 and a significant decrease in real estate prices. The French banks saw an increase in non-performing loans, a fall in value of real estate property assets in portfolios and substantially reduced the supply of loans to property developers and sellers. The sharp decrease in commercial real estate prices, used a s collateral had a negative impact on the financial position of borrowers and led to some defaults. The activity was then dampened by the ERM crisis of 1992 and the fragility of the banking sector with the near bankruptcy of the Crdit Lyonnais (related to the real estate market downturn and a risky business model strategy). The trough of the recession was reached in 1993 Q1. The sluggishness of activity in the construction sector, following the drop in real estate prices in 1990, lasted until 1998.

We present results for out-of-sample pre-crisis prediction for 2000Q3 to 2017Q4. This includes the period of the second systemic crisis (2008 q1 to 2009 q4) which is predicted out-of sample. That systemic crisis followed the collapse of Lehman Brothers after an era of growing GDP, falling unemployment, excessive credit growth and booming real estate prices. A confidence crisis triggered a recessionary loop with a fall in investment and consumption, as private agents tried to deleverage in front of a deteriorating economic environment.France entered recession in Q3 2008, for four quarters. In the meantime, unemployment rate rose from 7.5% to 9.5%. This also resulted in a 10% fall in residential real estate prices after a boom in the 1995-2007 period. There was in particular a restructuring and capital injection into Dexia, 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 (May2009). 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. That period however saw spillovers from the crises in some euro area countries both in terms of real activity and via expo-

Figure 1: France

sure of French banks to the periphery.

Out-of-sample prediction of systemic crises: France.

Figure 2 presents the results for the EWA aggregation rule. It shows that the probability of being in a pre-crisis in 2001-2004 was non zero but low with a sharp increase starting in 2005q2. Since the probability increases over time, the model provides a very good early warning system. The 12 quarter ahead crisis probability reaches 1 and remains very high till 2008q1. The model performs very well as the crisis starts in 2008 Q1and accordingly the probability drops like a stone. For the 3 other aggregation methods (shown in appendix D) the results are very similar. For all the aggregation riles there is a short-lived spike in the probability of pre-crisis which occurs during the pre-euro area crisis period (a residual event) and then dies out. 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 ML and the Ridge it is optimised upon whereas for the OGD the theoretically calibrated value of the learning rate is used. As a result we usually find very similar results for EWA, ML and Ridge and somewhat different time series of weights for the

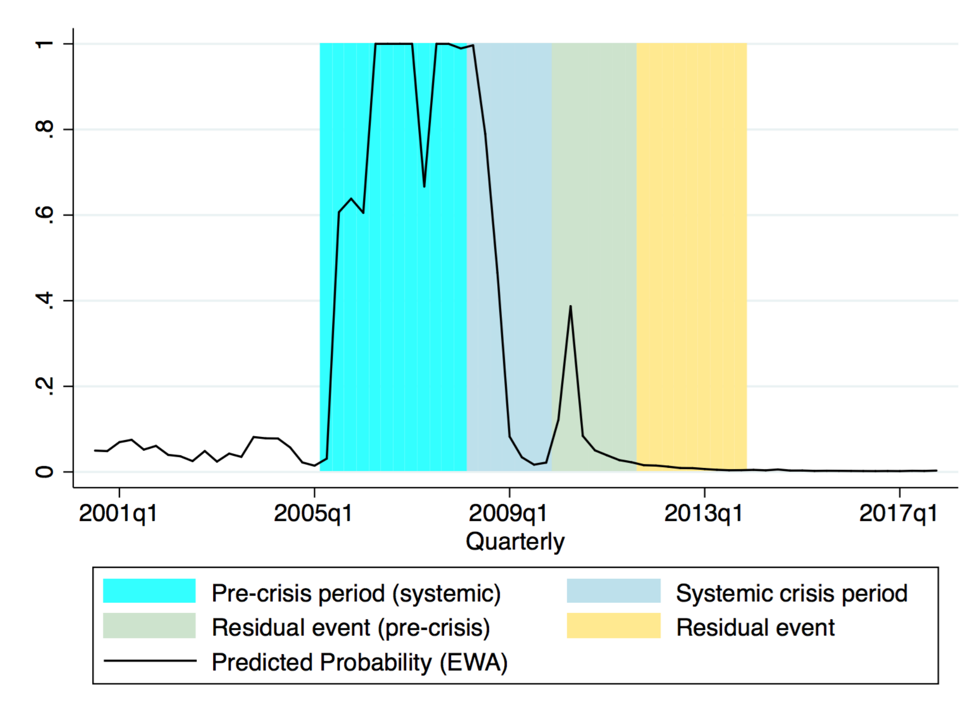

Figure 2: France: Predicted probability - EWA

OGD (different speeds of learning, the OGD being less reactive). This said, the results across the 4 aggregation rules are very consistent.

Table 1 presents the Root Mean Squared Errors (RMSE) of our different aggregation rules. We note that the EWA RMSE is close to its theoretical asymptotic value of the best convex combination of experts (0.23 versus 0.24 for the best convex combination (known ex post with fixed weights). EWA, ML, OGD, Ridge all do a lot better than uniform weights. The slightly higher RMSE for OGD compared to EWA, ML and Ridge may be due to the calibrated weights). 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.

Figure 3 shows the time varying weights associated to each of our 22 experts for the EWA aggregation rule and Figure 4 presents the contribution of the experts to the forecast. First on the batch sample from 1990q2 till 2000q1, then on the online sample with fixed weights for 12 quarters (2000q2 till 2003q2) because of delayed information revelation and then with time varying weights from 2003 q3 till 2017q4. We see that there is some updating of weights when we go from in-

Online Aggregation Rule	RMSE
EWA	0.23
ML	0.25
OGD	0.30
Ridge	0.23
Best fixed convex combination	0.24
Uniform	0.41

Table 1: RMSE of different aggregation rules and expert. France: quasi-real time

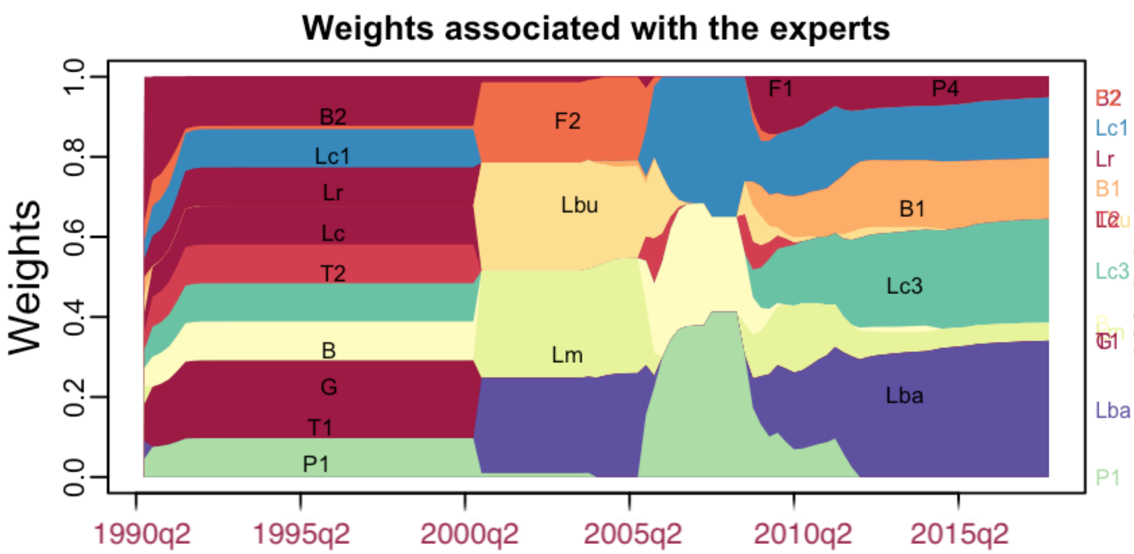

Figure 3: France: Weights. quasi-real time. EWA. 2000q3-2017q4

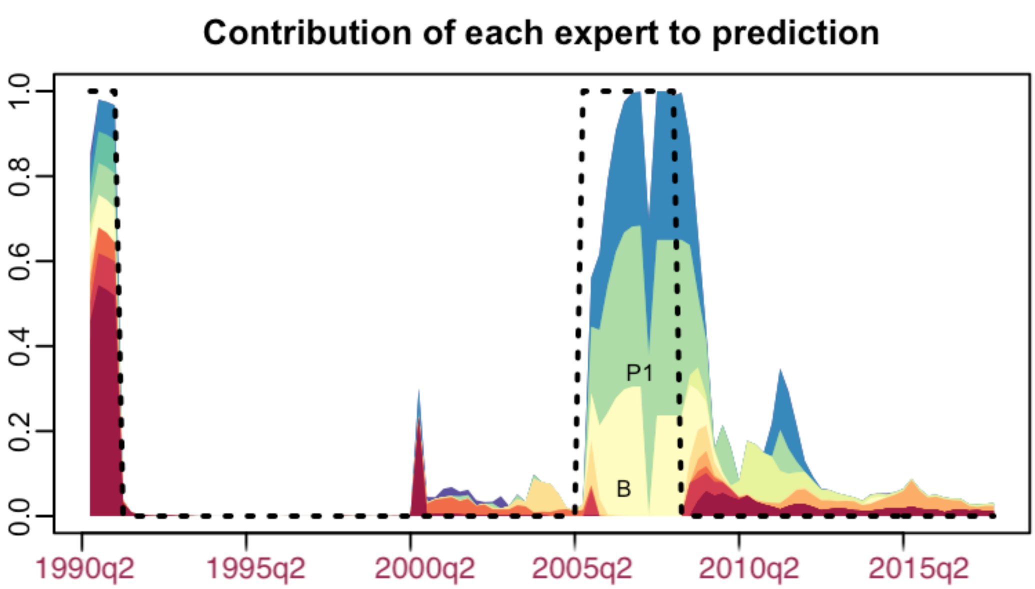

Figure 4: France: Experts. quasi-real time. Contribution to forecast. EWA. 2002q3-2017q4

sample to out-of-sample forecasting and when information is revealed. The optimal forecast for each of our rules puts some positive weights on most of our models on the batch sample while towards the end of the sample fewer experts tend to dominate. Towards the beginning of the sample the more 'general' models tend to be included (classification trees, bayesian averaging, grumpy, bashful) and there is probably some overfitting in sample. As the out of sample prediction and evaluation starts, the models picked up change. In the intermediate period, the random forest and some models with risk taking, monetary and bubble variables become important. In Figure 4, we see that when the crisis probability peaks, Lc1 (elastic net with housing and the real economy), P1 (credit to non financial sector, real GDP and the rent price index) and bashful [5] are the largest contributors. Subsequently, during the sovereign debt crisis event, the monetary variables logit (Lm) becomes important for the prediction as well as the variables of B1 (credit, rent price index, total credit to non financial sector, unemployment rate) but the probability of crisis remains low. The weights assign on the risk taking elastic net (Lba) and on the monetary and risk taking elastic net (lc3) are also high in the latter part of the sample but these variables (unlike Lc1 and Lm) do not contribute much to an increase in crisis probability during the sovereign debt crisis event.

This is not totally surprising. Housing, real economy, credit and risk taking variables played an important role in the Great Financial Crisis in the US (Mian and Sufi, ...). They seem to have done so as well in France with 3 main models incorporating rent-to-price index, house price to income, real estate price, credit to non financial sector (total and bank credit), real GDP, productivity, oil price, financial conditions, bank equity and risk appetite giving a sharp signal as early as 2005q2. We note that these models incorporate variables in levels, first and second difference, sometimes third, etc... so that non linearities and inflexion points in growth can be captured.

The elastic net models tend to be picked. The performance of elastic-net penalty, firstly introduced by Zou and Hastie [2005] uses a penalty that is part ℓ_1, part ℓ_2, and has the advantage to work as well as the loss whenever the lasso does the best and can fix the problems when lasso is

[5](Price-to-rent ratio; Price-to-income ratio; Unemployment ratio; Total Credit to non-financial corporations; %GDP; Total Credit to non-financial corporations gap to trend; Total Credit to non-financial corporations 1y change; Total Credit to private non-financial sector 1y change; Banking credit to private non-financial sector 1y change; GDP 1y change; M3 1y change; Financial Conditions Index 2y change; Financial Conditions Index 3y change; Total Liquidity Index 3y change; Bank equity)

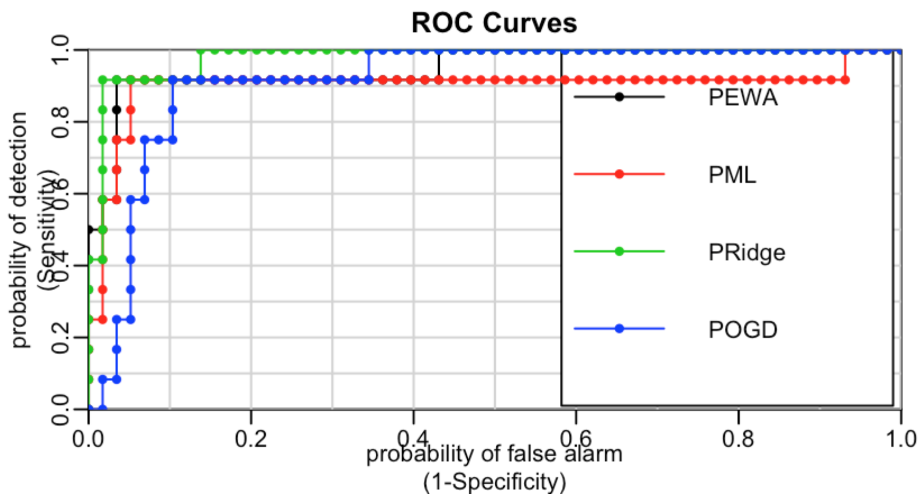

Figure 5: France: AUROCs. EWA=0.95; ML=0.90; OGD=0.98; Ridge=0.92

dominated by ridge regression. This is usually the case if there exist high correlations among predictors, as in our situation [Tibshirani, 1996]. The elastic net outperforms LASSO and preserves the sparse property [Zou and Hastie, 2005; Mol, Vitto and Rossasco, 2009]. Its performance have been confirmed in various applications [Zou and Hastie, 2005; Mol, Vitto and Rossasco, 2009; Mol et al., 2009; Destrero et al.,2009]

We present the results for some of the other aggregation strategies in Appendix D. They are remarkably consistent. Because the Ridge allows negative weights, its weighting scheme is different.

A very commonly used diagnostic of quality of early warning indicators is the ROC curve (Receiving Operator Curve). The ROC curve represents the ability of a binary classifier by plotting the true positive rate against the false positive rate for all thresholds. Figure 5 plots the ROC curves for France for the four aggregation rules. If the model made a perfect prediction the area under the curve would be equal to 1. We see that the performance of the model is exceptionally good as the AUROC is very close to 1 (above 0.9) in all 4 cases. Furthermore the time profile of the signal is such that it would have been very valuable for policy makers. Figure 6 plots the ROC curves for France for the Ridge and, the best fixed convex combination (known ex post) and the uniform aggregation rule. It is clear that the on-line learning rules win hand down.

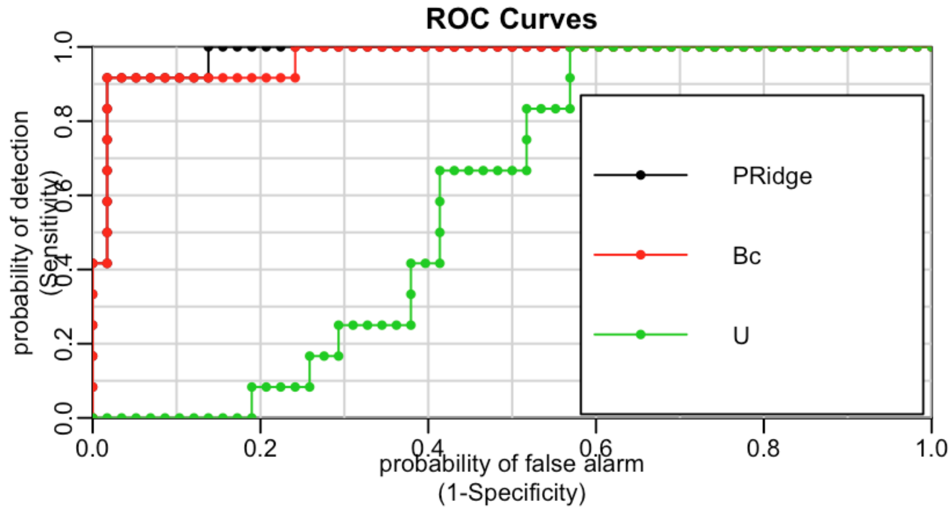

Figure 6: France: AUROCs. Ridge=0.92; Best fixed convex combination (ex post)=0.95; Uniform=0.59

5.2 Germany

We now turn to Germany. We estimate the expert models on sample 1987Q1-2000Q2. 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. Figure 7 presents the timing of pre-crises in Germany in light blue. The systemic crises are is in grey. The algorithm learns on the systemic crisis 2001q1- 2003 q4. That 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 finew economy" stock market bubble, put significant stress on a structurally vulnerable German financial sector with low profitability. Some of the largest institutions, had to adjust their balance sheets and to correct lending standards in a pro-cyclical way with negative feedbacks effects.

The second systemic crisis 2007 q2 till 2013 q2 is predicted out-of-sample. During the years preceding the Lehman Brother down fall, German financial institutions became strongly interconnected with the international financial system and involved in the build-up of systemic risks in this system, while domestic credit demand expanded only moderately. The drying up of market and funding liquidity was a key element during the early stages of the crisis (Brunnermeier). In addition to securitizations, some banks in Germany had noticeable exposures to commercial real

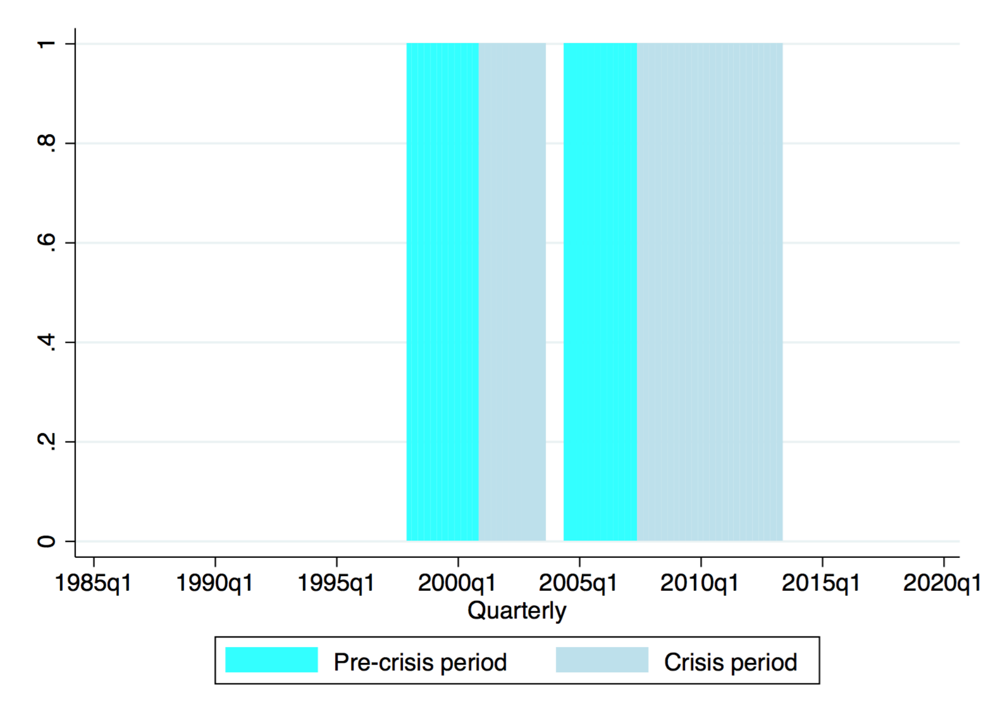

Figure 7: Timing of systemic crises and pre-crises: Germany

estate and the shipping industry (Hellwig). High leverage increased the risk of pro-cyclical fire sales and a credit crunch. In the later stage of the crisis exposures to stressed EMU sovereigns and banking systems affected the financial sector in Germany.

We present results for out-of-sample prediction for 2000Q3 to 2017Q4. Unlike France, there are no residual events during that out-of-sample forecast period but a longer systemic crisis and fewer periods in between the in sample and the out of sample systemic crises.

Out-of-sample prediction of systemic crises: Germany.

Figure 8 (see also Appendix D) presents the predicted probability for the EWA aggregation rule. The out-of sample forecast is very good although not as good as the one for France. This could be due to teh performance of our individual experts, may be not as well suited for the German data, in which case we could imporve by expanding the set of experts. Or it coudl be due to the fact that the two crises are closer to one another with a smaller scope for learning. Nevertheless, the out-of sample forecasting power is still very high. The probability of being in a

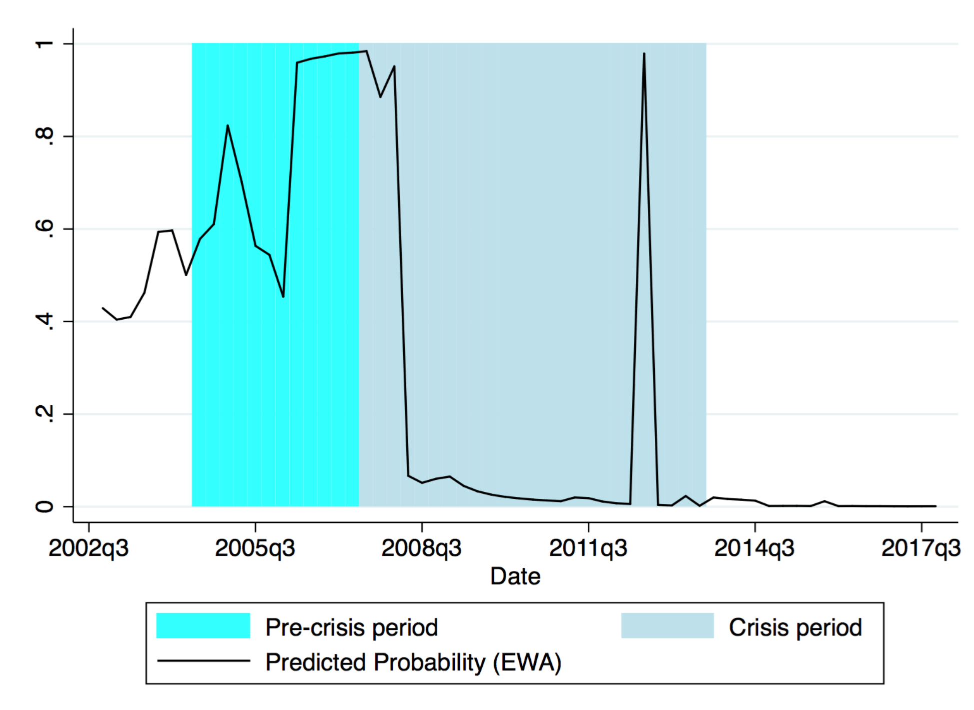

Figure 8: Germany: Predicted probability - EWA (quasi-real time)

pre-crisis in 2003 is already elevated according to all four rules but there was a sharp increase in 2004. The model shows a good ability to pick up the turning points albeit not as well as for France. The model also performs well as the crisis starts: the probability drops quickly for the EWA rule, but more gradually for the other rules. There is an unexplained peak in pre-crisis probability in 2013 at the time of the euro area sovereign debt crisis.

Figure 9 shows the time varying weights associated to each of our 22 experts for the EWA aggregation rule and Figure 10 presents the contribution of the experts to the forecast. First on the batch sample from then on the online sample with fixed weights for 12 quarters because of delayed information revelation and then with time varying weights from 2002 q3 till 2017q4. The optimal forecast for each of our rules puts some positive weights on mqny of our models on the batch sample while towards the end of the sample fewer experts tend to dominate. Towards the beginning of the sample the more 'general' models tend to be included (classification tree T2, bayesian averaging B2, grumpy G, bashful B but also P1 and Lcr and Lc3) and there is probably some overfitting in sample. As the out of sample prediction and evaluation starts, the models picked up change. In the intermediate period, most of the logit models with elastic net penalty

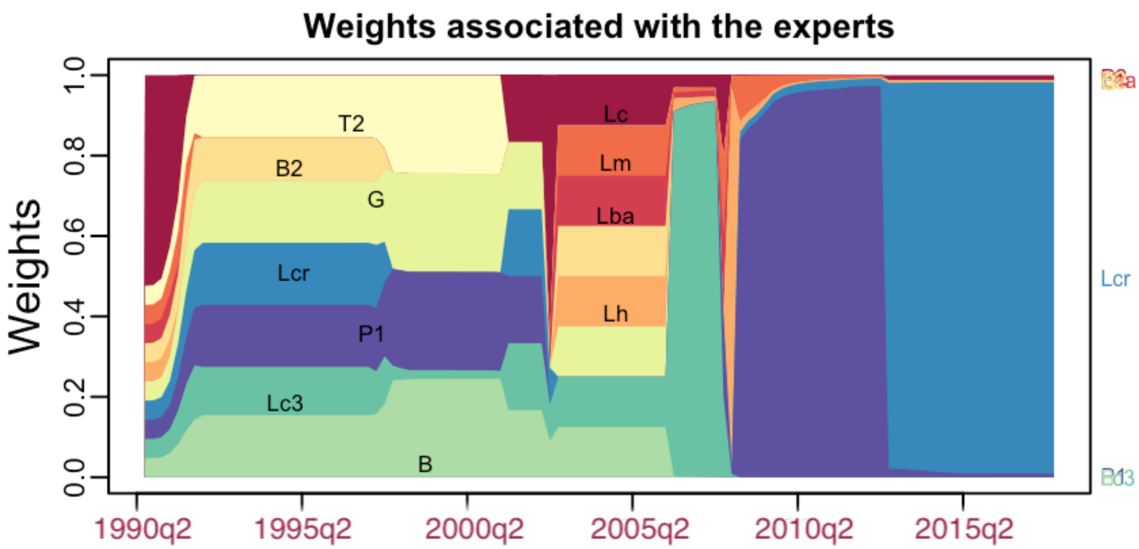

Figure 9: Germany: Weights. quasi-real time - EWA

become important while towards the end of the out of sample period the models P1 (housing, credit to non financial, real economy) and then Lcr (real economy and credit) become very dominant. In Figure 10, we see that when the pre-crisis probability peaks, the monetary logit (Lm), the housing (Lh), the risk taking logit (Lba), the monetary and risk taking and cross border variables (Lc3), the risk taking cross border capital logit (Lc) all give strong signals. It is then the monetary and risk taking (and cross border variables) Lc3 that peaks towards 1 for an extended part of the pre-crisis period. The peak in 2013 during the sovereign debt crisis is due to P1 (for Germany: share price index, rent price index, credit to non financial corporations; equity holdings). This is interesting. The same variables as in France tend to play a role but the monetary and stock market varaibles seem to be a bit more important in giving pre-crisis signals. We present the results for some of the other aggregation strategies in Appendix D.

Table 2 presents the RMSE of our different aggregation rules. We note that the EWA and the Ridge RMSE are very close to zero and so is the best convex combination of experts (known ex post with fixed weights). EWA, ML, OGD, Ridge all do a lot better than uniform weights. EWA does better than the best uniform convex combination as the EWA weights can change over time.

Figure 11 and 12 plots the ROC curves for Germany. Just like for France the 4 aggregation strategies achieve very high AUROcs (larger than 0.9). They are clearly superior to the uniform

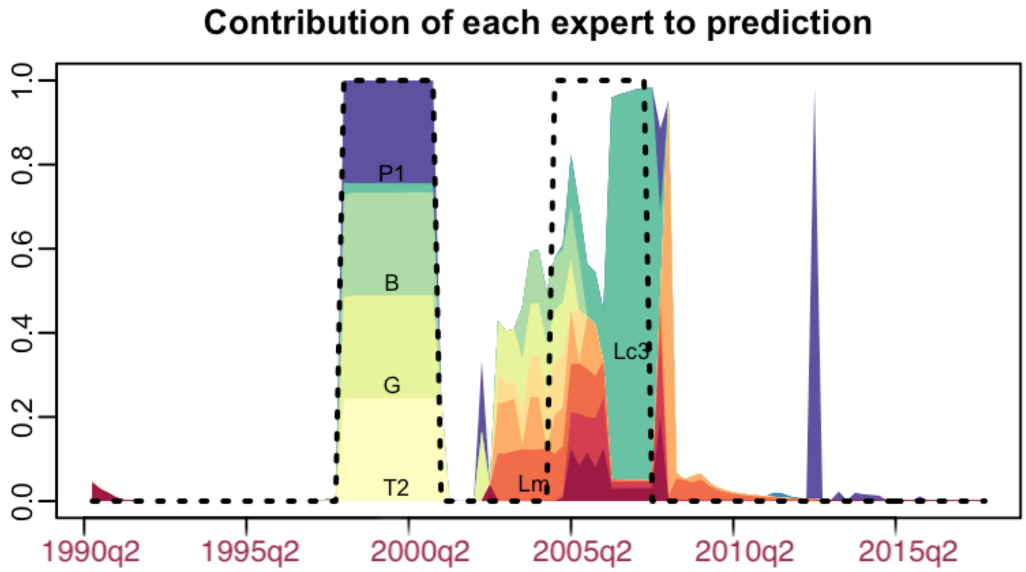

Figure 10: Germany: Contribution of experts to forecasts. Quasi-real time - EWA

Online Aggregation Rule	RMSE
EWA	0.33
ML	0.33
OGD	0.36
Ridge	0.33
Best convex combination	0.35
Uniform	0.44

Table 2: RMSE of different aggregation rules and expert. Germany: quasi-real time from 2002Q3 to 2017Q4

strategy or to the best linear convext combination of experst (computed using ex post information)

5.3 UK

For the UK, the crisis started in 2007 Q2 and ended in 2010 Q1. The previous crisis was 1991 Q2 till 1994 Q2.

The algorithm learns on the systemic crisis 1991q1- 1994 q4. This crisis was linked to excessive credit growth, high real estate prices and leverage.There was rapid credit expansion that 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

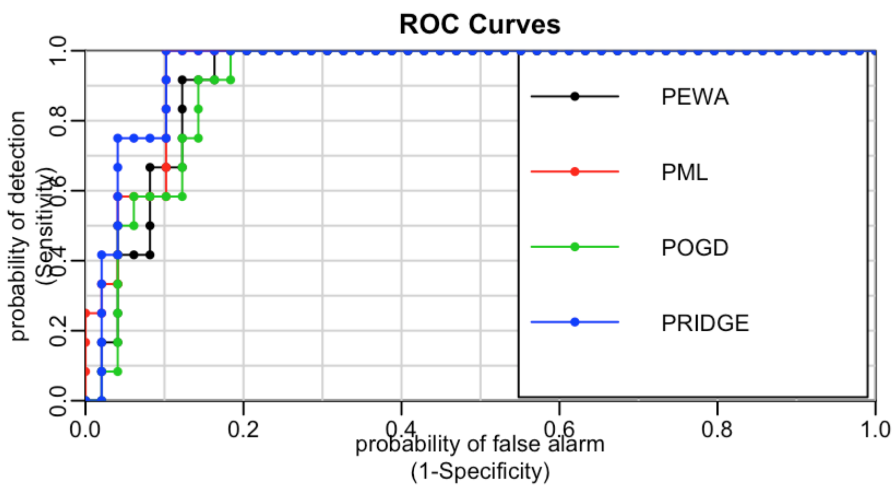

Figure 11: Germany: AUROCs EWA= 0.92; ML=0.95; OGD=0.92; Ridge=0.98. Quasi-real time.

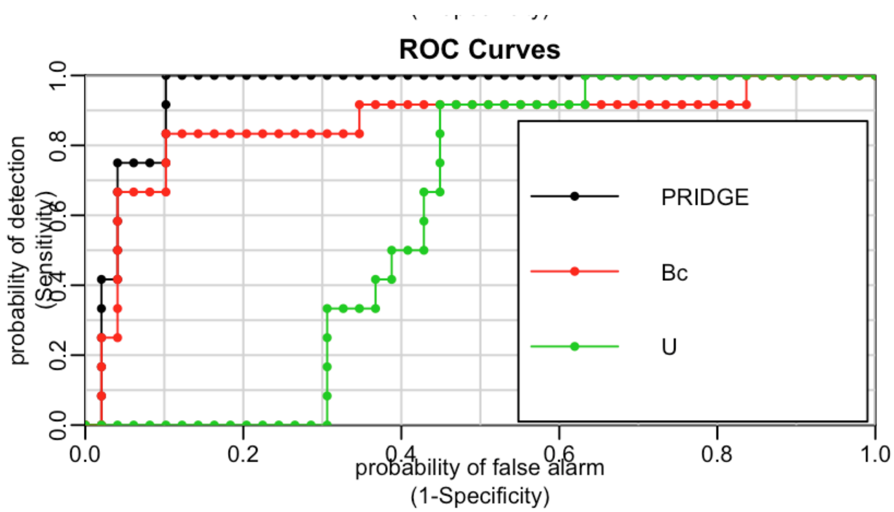

Figure 12: Germany: AUROCs Ridge=0.98; Best fixed convex combination (ex post)=0.86; Uniform=0.6 . Quasi-real time.

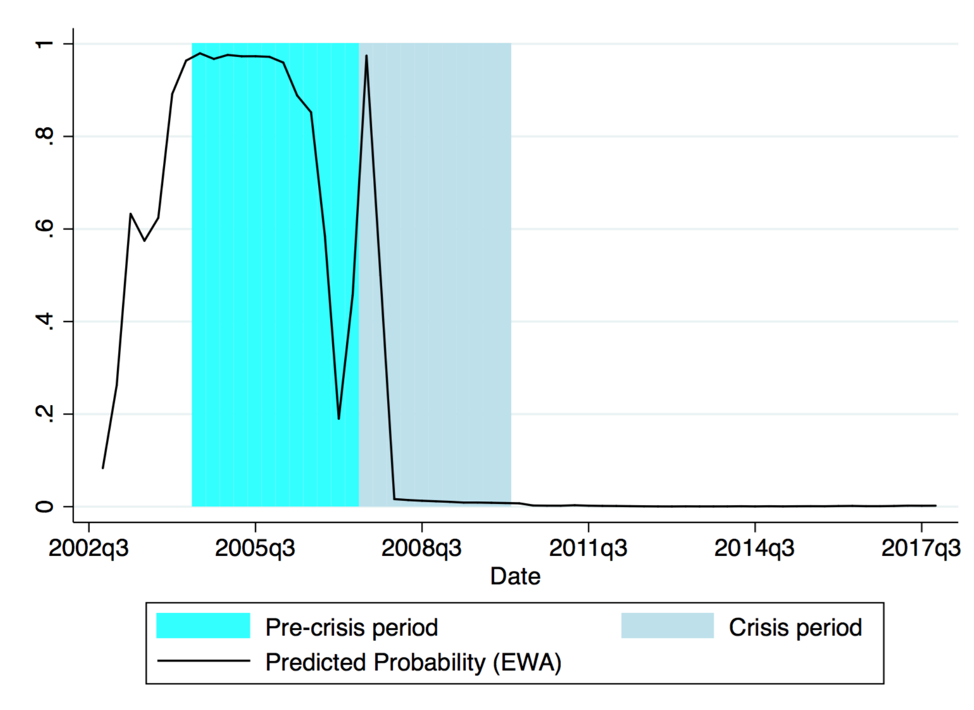

Figure 13: UK: Probability of pre-crisis quasi real time.

unable to access their funds at the BCCI (Bank of Credit and Commerce International). The event generated panic and moved their money to larger institutions. The European Rate Mechanism (ERM) forced the Bank 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 was rooted in weaknesses within the financial system that developed during an extended global credit boom: rapid balance sheet expansion; the creation of assets whose liquidity and credit quality were uncertain in less benign conditions; and fragilities in funding structures" (Financial Stability Report, October 2008, Bank of England). Although, the problem was not solely confined to the financial sector.

We present results for out-of-sample prediction for 2002Q3 to 2017Q4. Unlike France, there are no residual events during that out-of-sample forecast period.

Out-of-sample prediction of systemic crises: UK.

Figure 13 (see also Appendix D) presents the predicted probability for the EWA aggregation

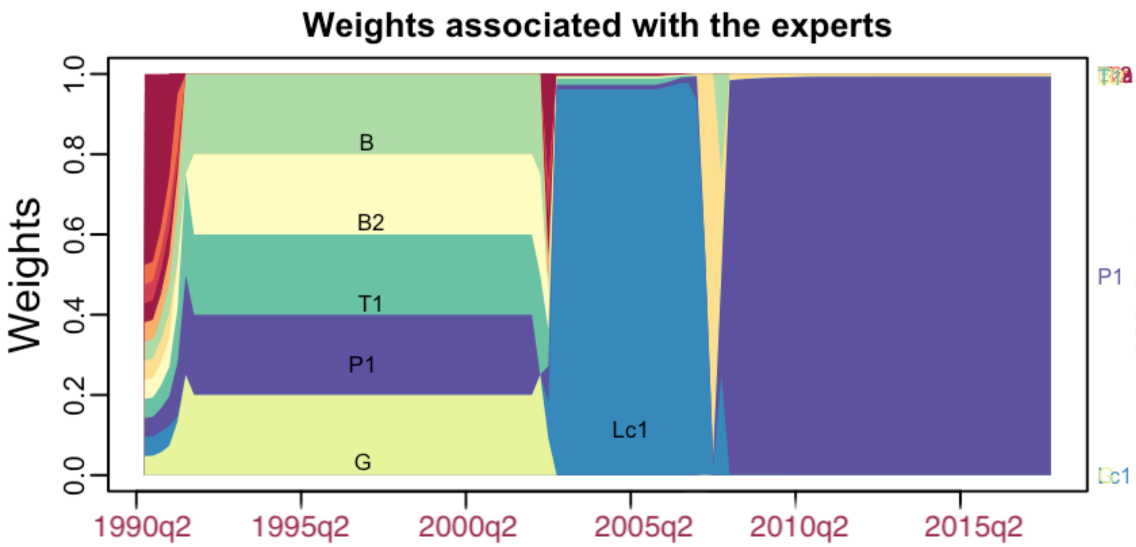

Figure 14: UK: Weights. quasi-real time - EWA

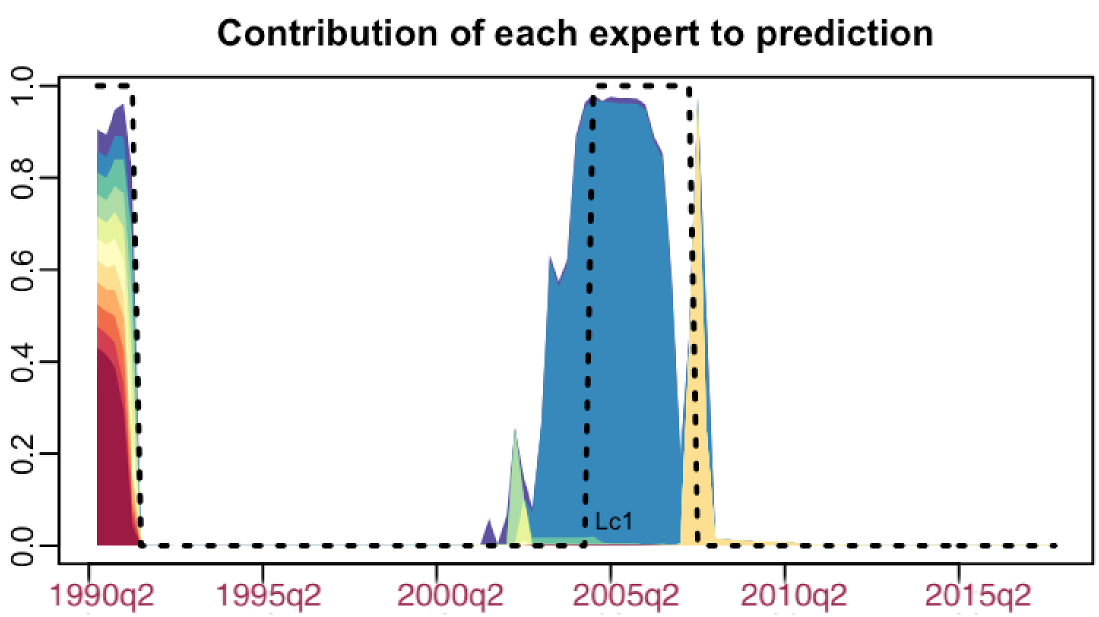

Figure 15: UK: Contribution of experts to forecasts. Quasi-real time - EWA

Online Aggregation Rule	RMSE
EWA	0.29
ML	0.32
OGD	0.35
Ridge	0.35
Best convex combination	0.34
Uniform	0.47

Table 3: RMSE of different aggregation rules and expert. UK: quasi-real time from 2002Q3 to 2017Q4

rule. The out-of sample forecast is very good. The probability of being in a pre-crisis in 2003 is already elevated according to all four rules but there was a sharp increase in 2004 exactly as it should be. The model shows a good ability to pick up the turning points. The model also performs well as the crisis starts: the probability drops quickly for the EWA rule, but more gradually for the other rules.

Figure 14 shows the time varying weights associated to each of our 22 experts for the EWA aggregation rule and Figure 15 presents the contribution of the experts to the forecast. First on the batch sample from then on the online sample with fixed weights for 12 quarters because of delayed information revelation and then with time varying weights from 2002 q3 till 2017q4. The optimal forecast for each of our rules puts some positive weights on many of our models on the batch sample while towards the end of the sample fewer experts tend to dominate. Towards the beginning of the sample the more 'general' models tend to be included (classification tree T1, bayesian averaging B2, grumpy G, bashful B but also P1) and there is probably some overfitting in sample. As the out of sample prediction and evaluation starts, the models picked up change. In the intermediate period, the logit models with elastic net penalty Lc1 (real economy and housing) become dominant while towards the end of the out of sample period the model P1 (Price-to-rent ratio, Price-to-income ratio, Total Credit to non-financial corporations, Consumer prices) becomes very dominant.

In Figure 15, we see that when the pre-crisis probability peaks, the logit model with elastic net penalty Lc1 (real economy and housing) is the one giving the signal. We present the results for some of the other aggregation strategies in Appendix D.

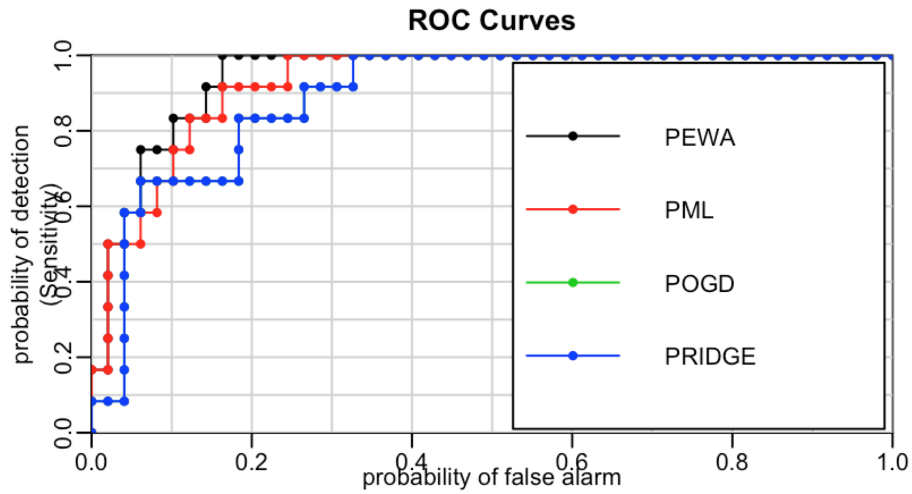

Figure 16: UK: AUROCs; EWA=0.95; ML=0.93; OGD= 0.89; Ridge=0.89. quasi-real time.

Figures 16 and 17 plot the ROC curves for the UK.

5.4 Spain

We present the parallel set of results for Spain. Spain had two systemic crises, 1978q1-1985q3 (too early to be in our sample) and 2009q1-2013q4. It also suffered a 'residual event' from 1993q3 to 1994q3 on which the algorithm will be able to learn.

In 1993q3 to 1994q3, the bank Banesto suffered problems related to its exposure to the industrial sector, which had undergone a severe crisis. The European Monetary System crisis (1992) also impacted the Spanish economy.

In 2009q1-2013q4, there was an economic crisis due to the collapse of the real estate market. This was coupled with a banking crisis. In addition the distress of other euro area economies caused severe sovereign tensions starting in 2011. While the fiscal position of Spain was relatively sound at the beginning of the crisis (debt over GDP was below 40% at the beginning of the crisis), the economic recession and the banking crisis exhausted the absorbing capacity of government debt. The main trigger of the crisis was the bursting of the housing bubble. We present results for out-of-sample prediction for 2002Q3 to 2017Q4.

Out-of-sample prediction of systemic crises: Spain.

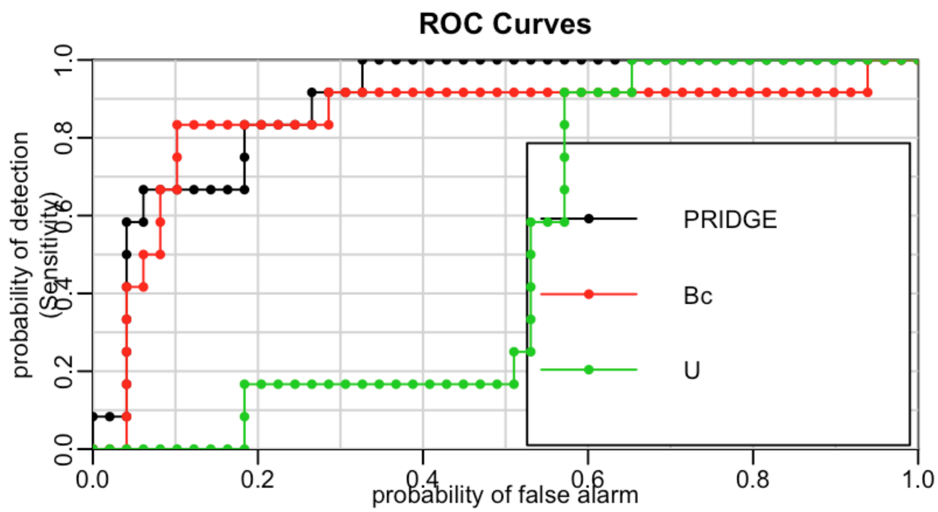

Figure 17: UK: AUROCs; Best convex (ex post)= 0.85; Uniform= 0.5; Ridge=0.89. quasi-real time.

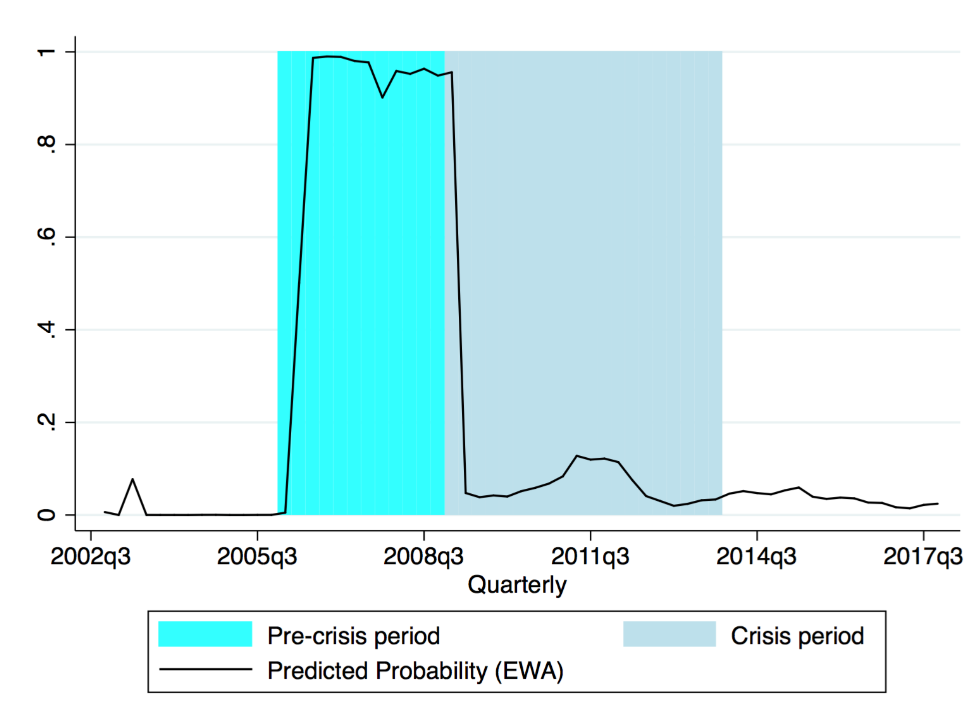

Figure 18: Spain: Probability of pre-crisis quasi real time. EWA

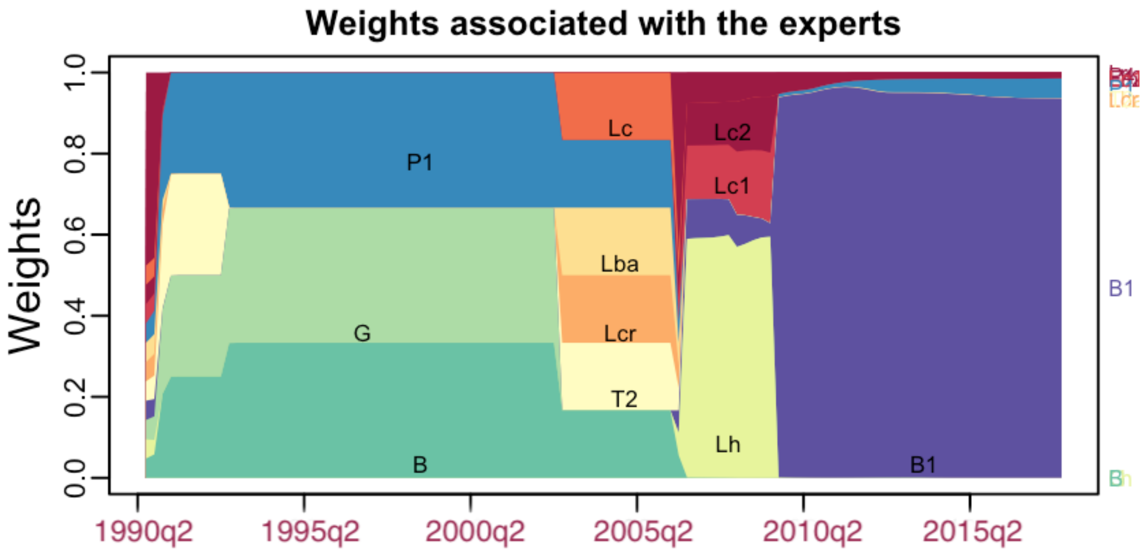

Figure 19: Spain: Weights. quasi-real time - EWA

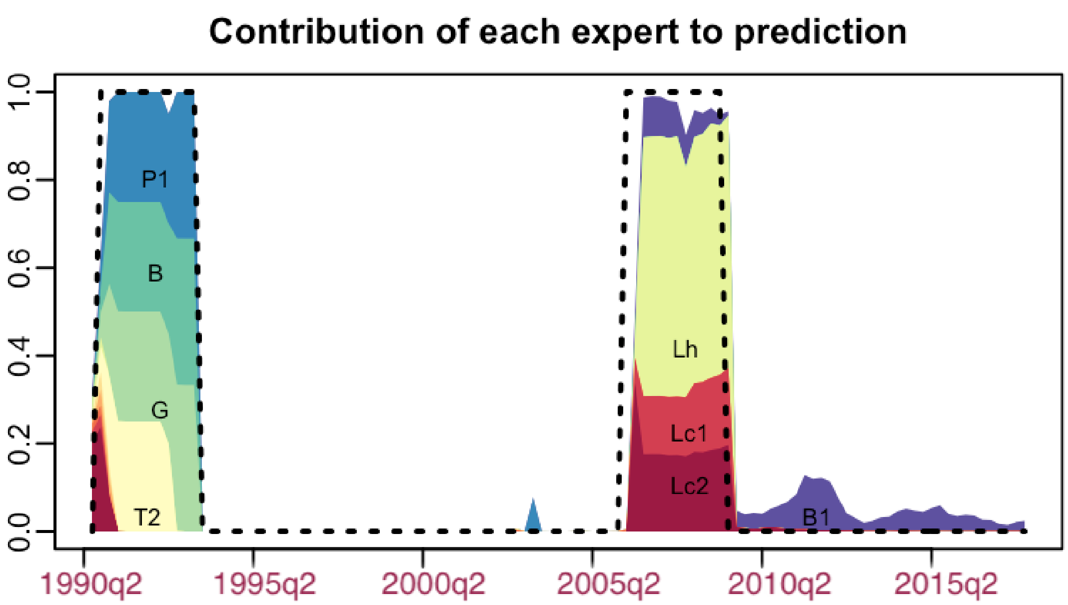

Figure 20: Spain: Contribution of experts to forecasts. Quasi-real time - EWA

Figure 18 (see also Appendix D) presents the predicted probability for the EWA aggregation rule. The out-of sample forecast is very good. The probability of being in a pre-crisis in 2003-2004 is low according to all four rules but there was a sharp increase in 2005. The probability of pre-crisis is very close to 1 from 2005 onwards which would have given time to act to policymakers. The model shows a very good ability to pick up the turning points. The model also performs well as the crisis starts: the probability drops quickly for the EWA and the Ridge rules, but more gradually for the other 2 rules.

Figure 19 shows the time varying weights associated to each of our 22 experts for the EWA aggregation rule and Figure 20 presents the contribution of the experts to the forecast. First on the batch sample, then on the online sample with fixed weights for 12 quarters because of delayed information revelation and then with time varying weights from 2002 q3 till 2017q4. The optimal forecast for each of our rules puts some positive weights on grumpy G, bashful B but also P1 and there is probably some overfitting in sample. As the out of sample prediction and evaluation starts, the models picked up change. In the intermediate period, the logit models with elastic net penalty Lc1 (housing and real economy), Lh (housing), Lc2 (credit and risk taking) and bayesian averaging B1 become dominant while towards the end of the out of sample period the model B1 (Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector, 1y change, Total Credit to non-financial corporations 2y change, Total credit to non-financial sector gap to trend, Consumer prices 1y change and 3y change, M3 : 1y change and 2y change) becomes very dominant.

In Figure 20, we see that when the pre-crisis probability peaks, the logit model with elastic net penalty Lh (housing) is the one giving the signal, helped by Lc1 (housing and real economy) and Lc2 (credit and risk taking). B1 gives some signal during the euro area cris. We present the results for the other aggregation strategies in Appendix D.

Figures 21 and 22 plot the ROC curves for Spain.

Online Aggregation Rule	RMSE
EWA	0.19
ML	0.23
OGD	0.32
Ridge	0.26
Best convex combination	0.26
Uniform	0.38

Table 4: RMSE of different aggregation rules and expert. Spain: quasi-real time from 2002Q3 to 2017Q4

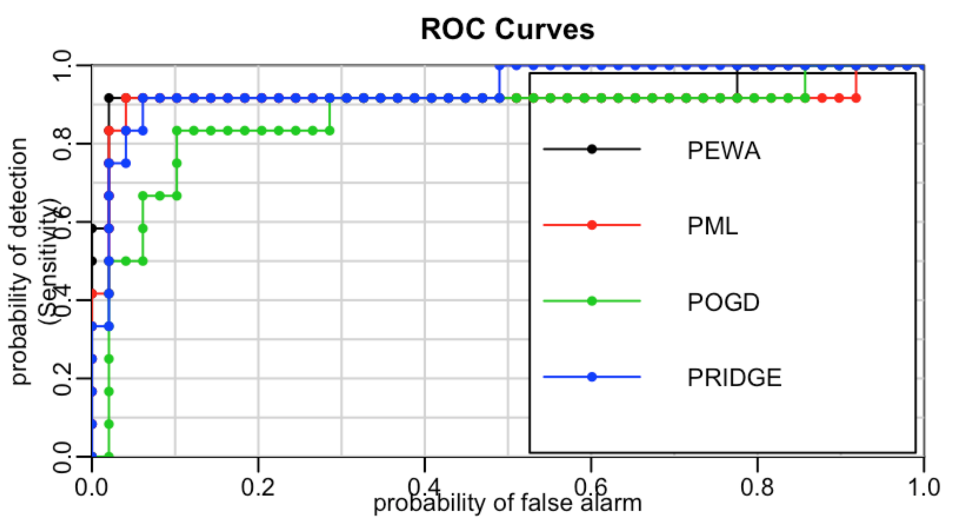

Figure 21: Spain: AUROCs; EWA=0.93; ML=0.91; OGD= 0.86; Ridge=0.94. quasi-real time.

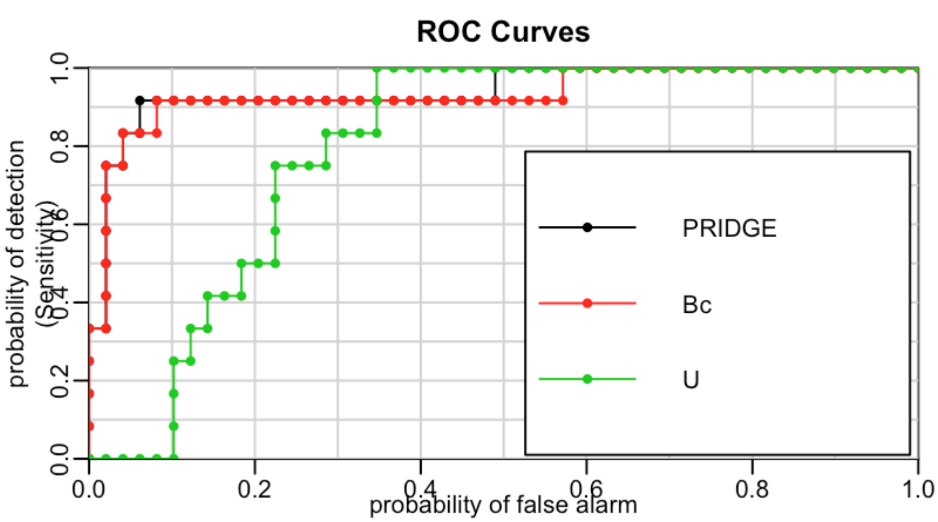

Figure 22: Spain: AUROCs; Best convex (ex post)= 0.93; Uniform= 0.38; Ridge=0.94. quasi-real time.

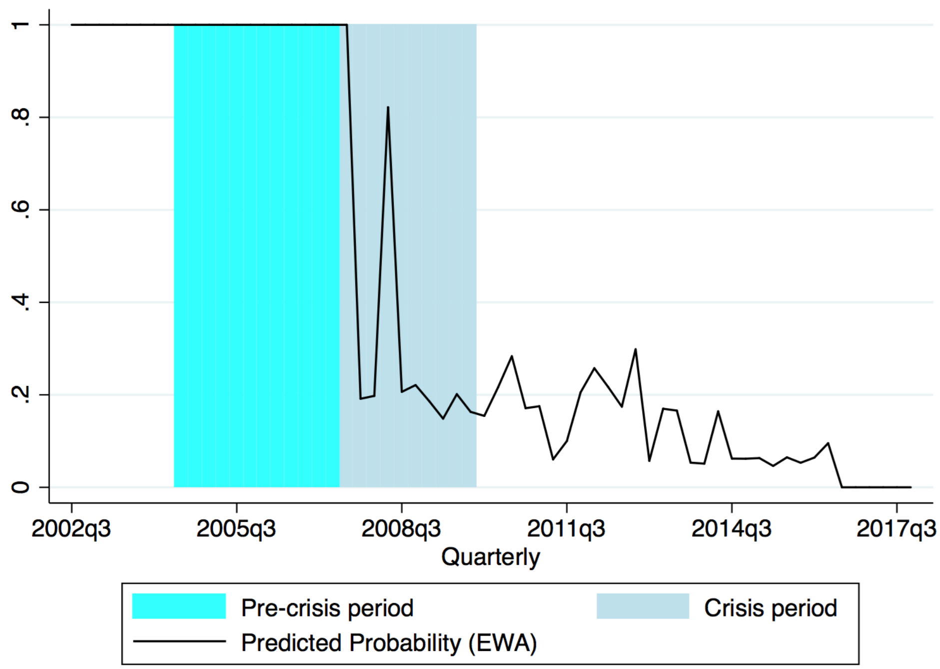

Figure 23: US: Probability of pre-crisis quasi real time. EWA

5.5 United States

We present the parallel set of results for the United States. The US had only one systemic crisis in our sample. It also suffered a 'residual event' on which the algorithm will be able to learn.

Out-of-sample prediction of systemic crises: United States.

Figure 23 (see also Appendix D) presents the predicted probability for the EWA aggregation rule. The probability of being in a pre-crisis in 2002-2004 was already at 1 however for the EWA and high for the other aggregation rules as well.

Figure 24 shows the time varying weights associated to each of our 22 experts for the EWA aggregation rule and Figure 25 presents the contribution of the experts to the forecast. First on the batch sample, then on the online sample with fixed weights for 12 quarters because of delayed information revelation and then with time varying weights from 2002 q3 till 2017q4. The optimal forecast for each of our rules puts a very high weight on bashful B which is giving also the signal of the pre-crisis early on. Bashful is a logit regression with the following variables: GDP per

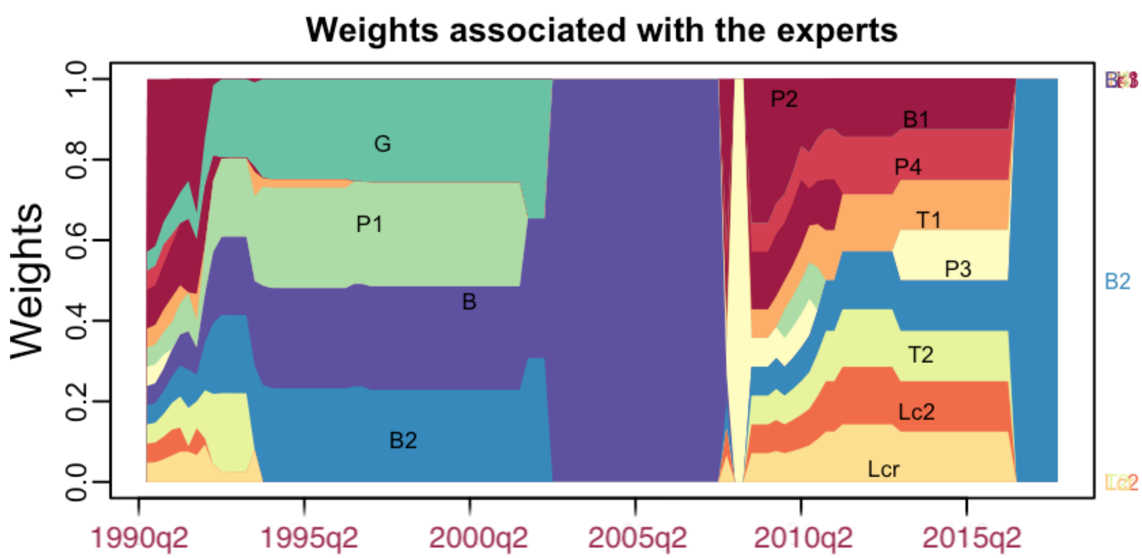

Figure 24: US: Weights. quasi-real time - EWA

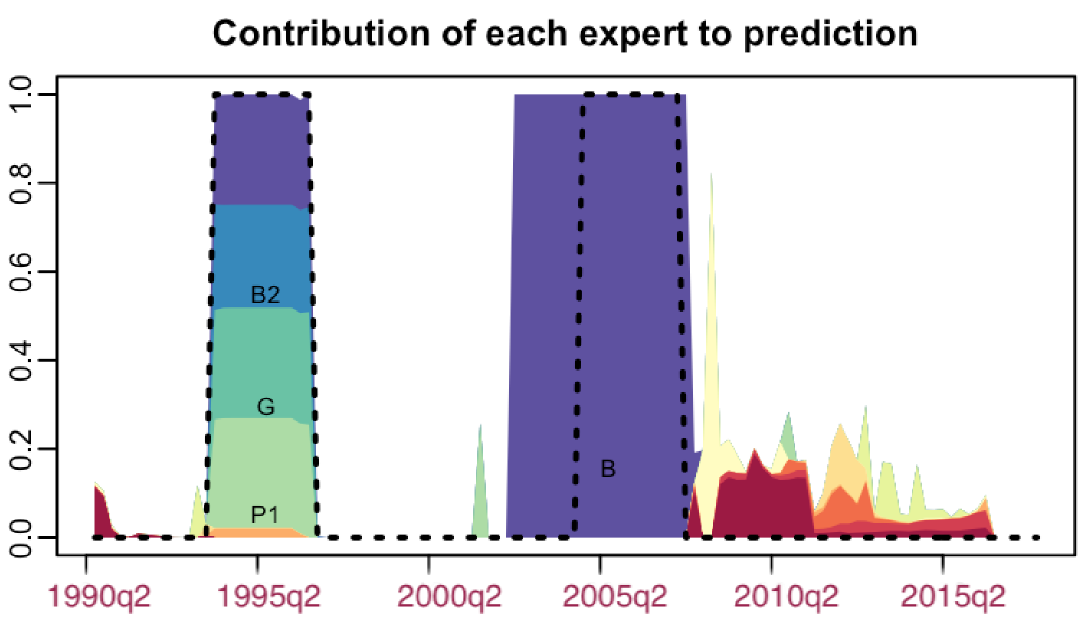

Figure 25: US: Contribution of experts to forecasts. Quasi-real time - EWA

Online Aggregation Rule	RMSE
EWA	0.41
ML	0.33
OGD	0.38
Ridge	0.38
Best convex combination	0.39
Uniform	0.47

Table 5: RMSE of different aggregation rules and expert. US: quasi-real time from 2002Q3 to 2017Q4

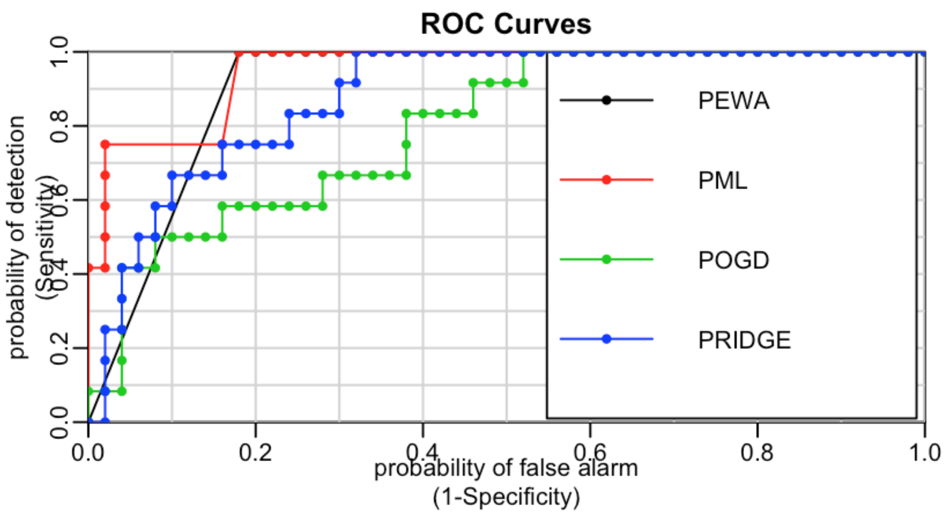

Figure 26: US: AUROCs; EWA=0.91; ML=0.95; OGD= 0.80; Ridge=0.88. quasi-real time.

person Index; Share price; Unemployment; GDP; Cross border flows; Slope of the yield curve; Golden rule 2y change; Golden rule 3y change; Consumer price 1y change and 2y change; Price-to-income ratio 1y change; Equity Holdings; Equity holdings 1y change,2y change,3y change; Financial assets; Financial assets 1y change, 2y change, 3y change; Fixed Holding 1y change, 2y change, 3y change; Liquid Assets 1y change,3y change; Bank assets 1y change; Bank equity 1y change ; Leverage 1y change; Economic Uncertainty Policy Index 1y change; Oil price 1y change. We note that for the US more variables than usual are available as some time series are longer (in particular bank assets and leverage).

We present the results for the other aggregation strategies in Appendix D.

Figures 26 and 27 plot the ROC curves for the US.

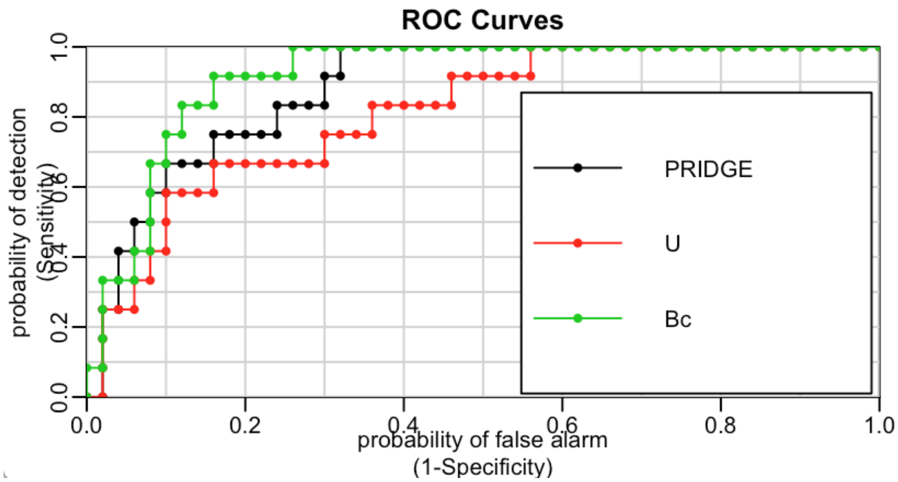

Figure 27: US: AUROCs; Best convex (ex post)= 0.91; Uniform= 0.81; Ridge=0.88. quasi-real time.

6 Real time data

We test our methodology using vintage data for France. 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 data, credit data and housing market related variables. We mostly have cross border capital data for the whole panel (liquidity indices built on real time flow data as well as risk taking indices built on asset price data) as well as exchange rates, and specifically for France M3 and inflation data which in France are not revised. We reestimate all our experts on the 1990q2-2002q2 sample using *only vintage data* and we use also only vintage data for the out- of-sample exercise. Despite the strong data limitations, we get very good results as shown in Figure 28[6].The probability of pre-crisis goes up now a bit before the pre crisis period but reaches 1 during the pre-crisis period. There is a similar spike as before during the pre crisis of the residual event. 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. As shown in Figure 29, The models given a high weight in the out-of sample prediction in real time are T1 (classification tree with panel data), B2 Bayesian model averaging, country specific), containing the following variables: M3 (1 year growth), Private Sector Liquidity Stock, Private

[6]Our real time out-of-sample exercise is too strict. Indeed, we even estimate our experts on the batch sample using vintage time series. We could instead use revised data for the batch estimation and use vintage data for the out of sample real time estimation. The strength of our results is therefore impressive.

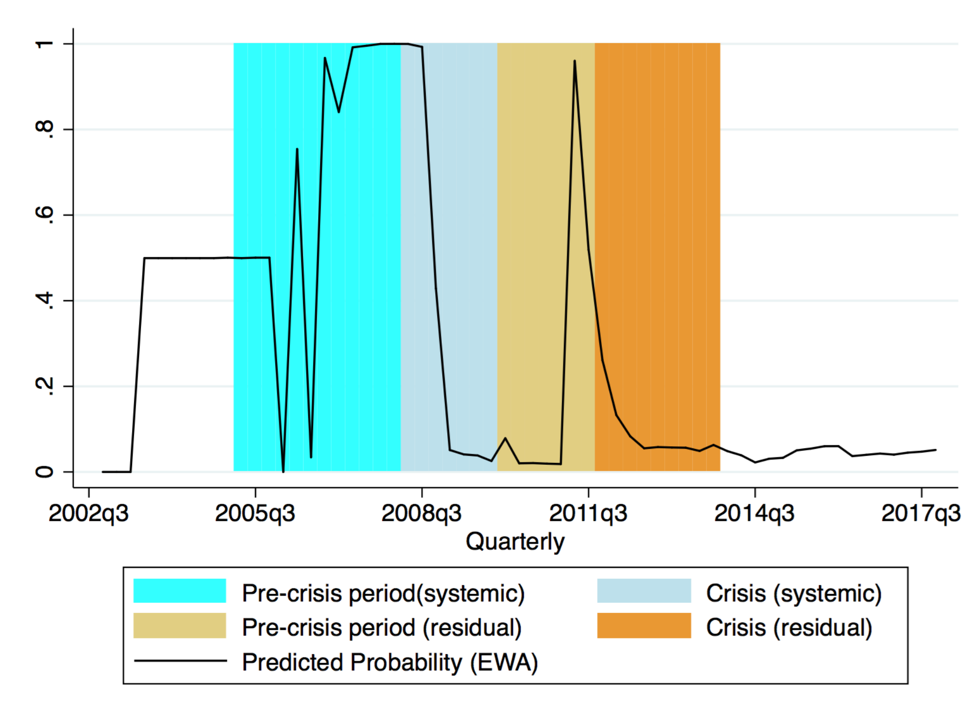

Figure 28: France: Probability of pre-crisis in **real time**. EWA

Sector Liquidity Stock (2 year growth), Domestic Liquidity Stock Local (2 year growth) and P1 (dynamic probit model with the same variables as B2), G (logit with all variables available) and Lf (Logit with elastic-net penalty with cross border flows, total liquidity flows, gross capital flows, in levels, and 1 year and 2 years growth rates).

Figure 30 shows that he models giving the signal is now mostly B2 and T1 which gives the early rise in probability.

Figures 31 and 32 plot the ROC curves for France in **real time**.

Online Aggregation Rule	RMSE
EWA	0.35
ML	0.83
OGD	0.72
Ridge	0.80
Best convex combination	0.75
Uniform	0.54

Table 6: RMSE of different aggregation rules and expert. France: **real time** from 2002Q3 to 2017Q4

Figure 29: France: Weights. **Real time** - EWA

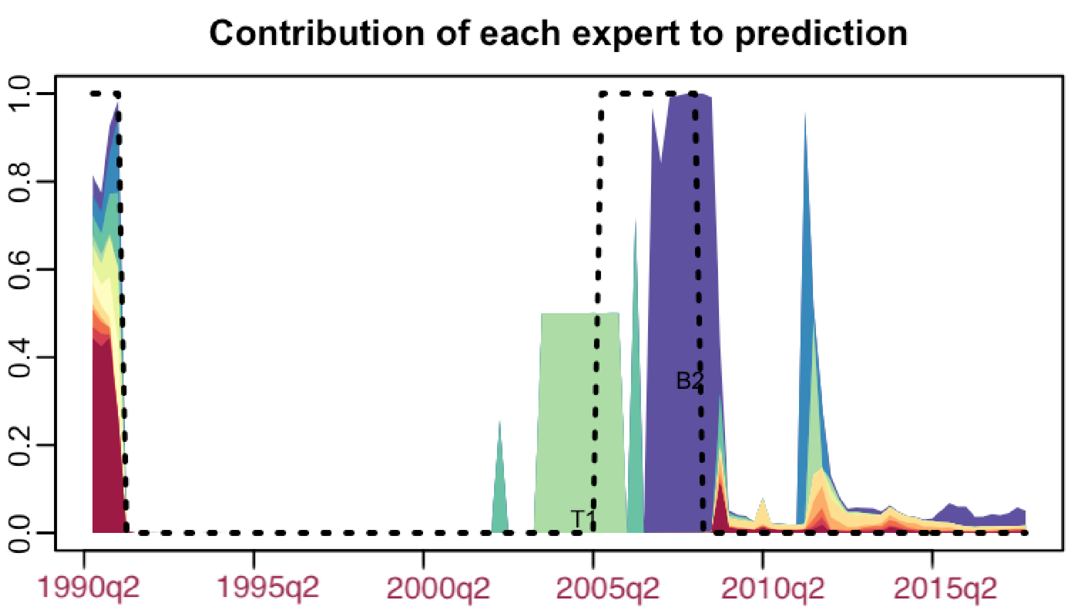

Figure 30: France: Contribution of experts to forecasts. **Real time** - EWA

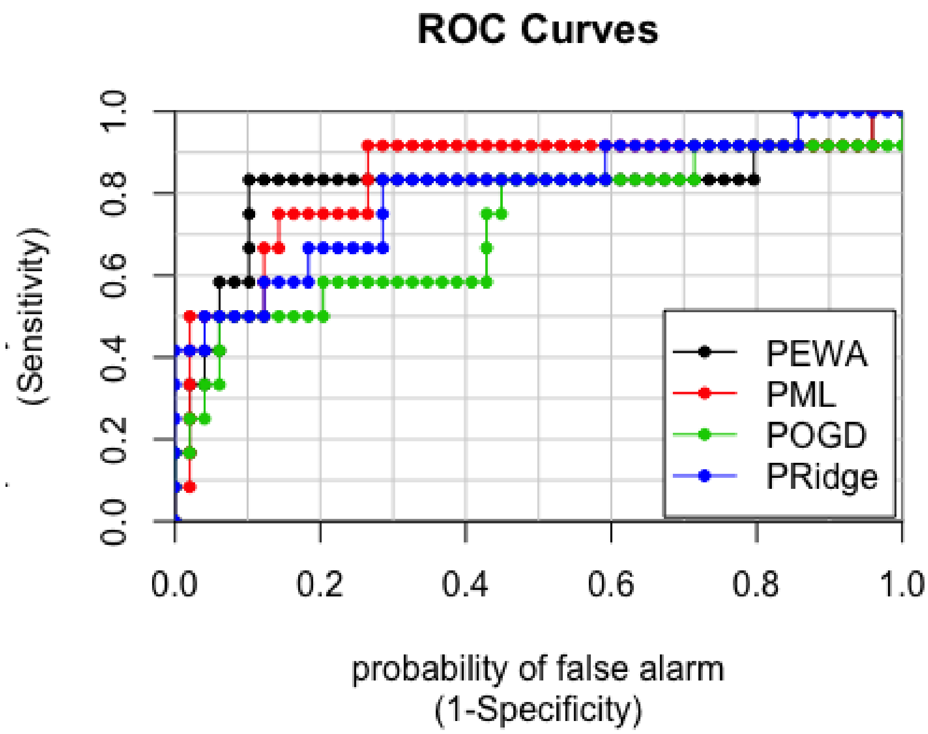

Figure 31: France: AUROCs; EWA=0.81; ML=0.83; OGD= 0.72; Ridge=0.80. **real time**.

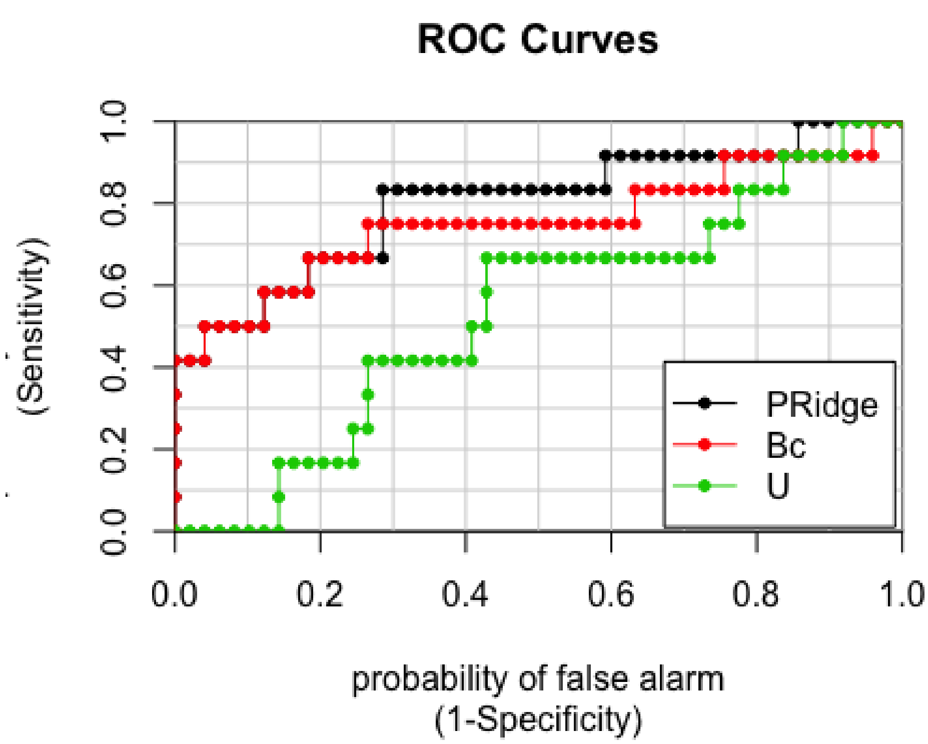

Figure 32: France: AUROCs; Best convex (ex post)= 0.75; Uniform= 0.54; Ridge=0.80. **real time**.

7 Conclusions

Our method has a unique ability to run a horse race between very eclectic experts to assess their performing ability and aggregate them in order to produce an optimal forecast. Using a mix of 22 experts, some of them being central bank financial crises models, some of them being machine learning models, we find that for France, Germany , UK, Spain, US we are able to predict systemic financial crisis 3 year ahead out of sample with AUROC close to 1. For the out-of-sample prediction, aggregation rules put a high weight on a model with credit, housing, monetary and risk taking variables but those weights are very heterogenous depending on the countries. For France, credit, real estate and real variables all contribute to give a signal. For GErmany it is the monetary and risk taking variables. For Spain it is the housing and credit variables. For the UK it is the housing and real variables. 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 is where the online nature of our algorithm is of course key as standard methodologies would not be able to extract enough information from the sample. Our method is very flexible: we could incorporate many more experts (deep learning, subjective judgement) and potentially increase further the performance of our model. Strikingly the performance of our model is also very good when we use real time data for France despite the restricted set of variables which are available in vintage time series. In a companion paper we use our methodology of online learning on historical data (Jorda Schularick and Taylor dataset) to predict the Great Recession in quasi real time. An obvious other application of our methodology is to predict recessions, something we will tackle in future work.

Appendix

A Data

B Expert

Country-specific selected variables for each expert :

1. France :
   - P1 : Total Credit to private non-financial sector 1y change, Total credit to non-financial firms 2y change, Real GDP 2y change, Rent Price Index 2y change.
   - P2 : M3,GDP per person index, Multifactor Productivity index, Price-to-income ratio.
   - B1 : Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector 1y change, Banking credit to private sector 2y change, Rent Price Index 2y change.
   - B2 : Total Credit to private non-financial sector 1y change, Total credit to non-financial firms 2y change, Real GDP 2y change, Rent Price Index 2y change, Total credit to non-financial firms 2y change, Unemployment rate.
   - B: Price-to-rent ratio; Price-to-income ratio; Unemployment ratio; Total Credit to non-financial corporations; %GDP; Total Credit to non-financial corporations gap to trend; Total Credit to non-financial corporations 1y change; Total Credit to private non-financial sector 1y change; Banking credit to private non-financial sector 1y change; GDP1y change; M3 1y change; Financial Conditions Index 2y change; Financial Conditions Index 3y change; Total Liquidity Index 3y change; Bank equity.
2. Germany :
   - P1 : Share Price Indices, Rent Price Index 1y change, Total Credit to non-financial corporations 2y change, Equity Holdings.
   - P2 : Total Credit to Households gap to trend, Total Credit to non-financial corporations gap to trend, GDP per hour worked index, M3.
   - B1 : Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector 1y change, Banking credit to private sector 2y change, Rent Price Index 2y change, Total Credit to non-financial corporations fi?? 2y change.
   - B2 : Share Price Indices, Rent Price Index 1y change, Total Credit to non-financial corporations 2y change, Equity Holdings.
   - B: Price to rent; Price to income; Banking credit to private non-financial sector; TC for Households; Long-term interest rate (nominal); Share Price Indices; Real effective exchange rate 3y change; Rent Price Index 1y change; Rent Price Index) 2y change; Equity holdings.
3. UK :

- P1 : Price-to-rent ratio, Price-to-income ratio, Total Credit to non-financial corporations 2y change, Consumer prices 2y change.
- P2 : M3, GDP per hour worked index, Total credit to household gap-to-trend,Banking credit to private sector gap-to-trend.
- B1 : Consumer Prices, Short-term interest rates (nominal), Banking credit to private sector 1y change, Banking credit to private sector 2y change, Rent Price Index 2y change, Total Credit to non-financial corporations 2y change.
- B2 : Price-to-rent ratio, Price-to-income ratio, Total Credit to non-financial corporations 2y change, Consumer prices 2y change, Multifactor productivity index,Short-term interest rate (nominal);
- B: GDP Per Capita Index; GDP Per Hour Index; Price-to-income; Total Credit (Household, private non-financial sector, non-financial corporations, banking credit) % GDP and gap to trend; Consumer prices; Short-term interest rate (nominal); Real estate prices; TC Households 1y change and 3y change; Banking credit 3y change; TC for private non-financial sector 1y change; GDP 1y change; Consumer prices 2y change; Risk Appetite 3y change; Financial Condition Index 1y change and 2y change; Total Liquidity Index 1y change and 2y change; Oil prices 1y change, 2y change, 3y change.

4. Spain:
    - P1: Price-to-rent ratio; Share price Indices; Rent Price Index 2y change; M3 2y change.
    - P2: GDP per capita Index ; Total Credit to households gap to trend; Total Credit to households % GDP; Equity Holdings.
    - B1: Consumer Prices; Short-term interest rates (nominal); Banking credit to private sector 1y change; Total Credit to non-financial corporations 2y change; Total credit to non-financial sector fi?? gap to trend; Consumer prices : 1y change and 3y change; M3 : 1y change and 2y change.
    - B2: GDP per Capita Index; GDP per Hours Index; Long-term interest rate; Unemployment rate; Total Credit to households gap to trend; Total credit to non-financial sector gap to trend; Consumer prices : 1y change and 3y change; M3 : 1y change and 2y change; Equity holdings; Oil prices : 1y change, 2y change and 3y change.
    - B: GDP per Capita Index; GDP per Hours Index; Total Credit to households; Consumer prices; Long-term interest rate; Unemployment rate; Total Credit to households gap to trend; Total credit to non-financial sector gap to trend; Total Credit to private non-financials ector 1y change; Total Credit to Households 3y change; Total Credit to non-financial sector; Consumer prices : 1y change and 3y change; M3 : 1y change and 2y change; Equity holdings; Oil prices : 1y change, 2y change and 3y change.

C Aggregation rules

We now consider another aggregation rule firstly introduced by Stoltz[] as an extension of the Prod Algorithm of Cesa-Bianchi[]. There are two main differences compared to other aggregation rules. First, there is no one unique learning rate for each expert anymore. In the Polynomially weighted averages with multiple learning rates (ML-pol) aggregation rule, each expert j

is associated to its own learning rate $\eta_{j,t}$. This aggregation rule is well calibrated for theoretical values. This is why it is complementary to other aggregation rules. Secondly, the forecaster still wants to control his cumulative loss, but he do that by directly controlling his regret $R_{j,t}$ against each expert j. For the notation, define the weight vector $p_t = (p_{0,t}, ..., p_{N,t})$ and the mixture $w_t = (w_{0,t}, ..., w_{N,t})$. Then the loss vector is defined by $\hat{\ell}_t = w_t^T \ell_t$ and each weight by :

$$p_{j,t} = (p_{j,t-1}(1 + \eta_{j,t-1}(\hat{\ell}_t - \ell_{j,t})))^{\frac{\eta_{j,t}}{\eta_{j,t-1}}}$$

Algorithm 5 Polynomially weighted averages with multiple learning rates (ML-Poly)

Parameter : a rule to sequentally pick the learning rates $(\eta_{1,t}, ..., \eta_{N,t})$
Initialization : the vector of regrets $(R_0 = (0, ..., 0))$
For each round $t = 1, ..., T$

1. pick the learning rates $(\eta_{1,t-1}, ..., \eta_{N,t-1})$
2. form the vector w_t defined compnent-wise : $w_{j,t} = \eta_{j,t-1} w_{k,t-1} / \eta_{t-1}^T w_{t-1}$
3. observe the loss vector ℓ_t and incurr loss $\hat{\ell}_t = w_t^T \ell_t$
4. for each expert j perform the update :

$$p_{j,t} = (p_{j,t-1}(1 + \eta_{j,t-1}(\hat{\ell}_t - \ell_{j,t})))^{\frac{\eta_{j,t}}{\eta_{j,t-1}}}$$

As in Stoltz[], we calibrate the learning rates following this rule :

$$\eta_{j,t-1} = \frac{1}{1 + \sum_{s=1}^{t-1} (\hat{\ell}_s - \ell_{j,s})^2}$$

With these learning rates, Stoltz et al.[] prooved the following bound for the cumulative loss :

Theorem 4. For all sequences of loss vectors $\ell_t \in [0, 1]^K$,

$$\sum_{t=1}^{T} \hat{\ell}_t \leq \min_{1 \leq j \leq N} \left\{ \sum_{t=1}^{T} \ell_{j,t} + \sqrt{N(1 + ln(1 + T))(1 + \sum_{t=1}^{T} (\hat{\ell}_t - \ell_{j,t})^2} \right\} \tag{4}$$

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

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 :

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

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] :

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

$$R(\mathcal{R}_\eta) \leq \inf_{v \in \mathbb{R}^N} \left\{ \lambda ||v_2^2|| \right\} + N \times ln(1 + \frac{T}{\lambda N}) \tag{5}$$

D Experts

The logits with elastic net penalty are constructed following Friedman [2010] as:

$\min_{\beta_0, \beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} F(\beta_0 + x_i \beta) - \lambda P_\alpha(\beta) \right\}$ and $P_\alpha(\beta) = \sum_{j=1}^{p} [\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|]$

Since there is a risk of correlation, we pick $\alpha = 0.7$.

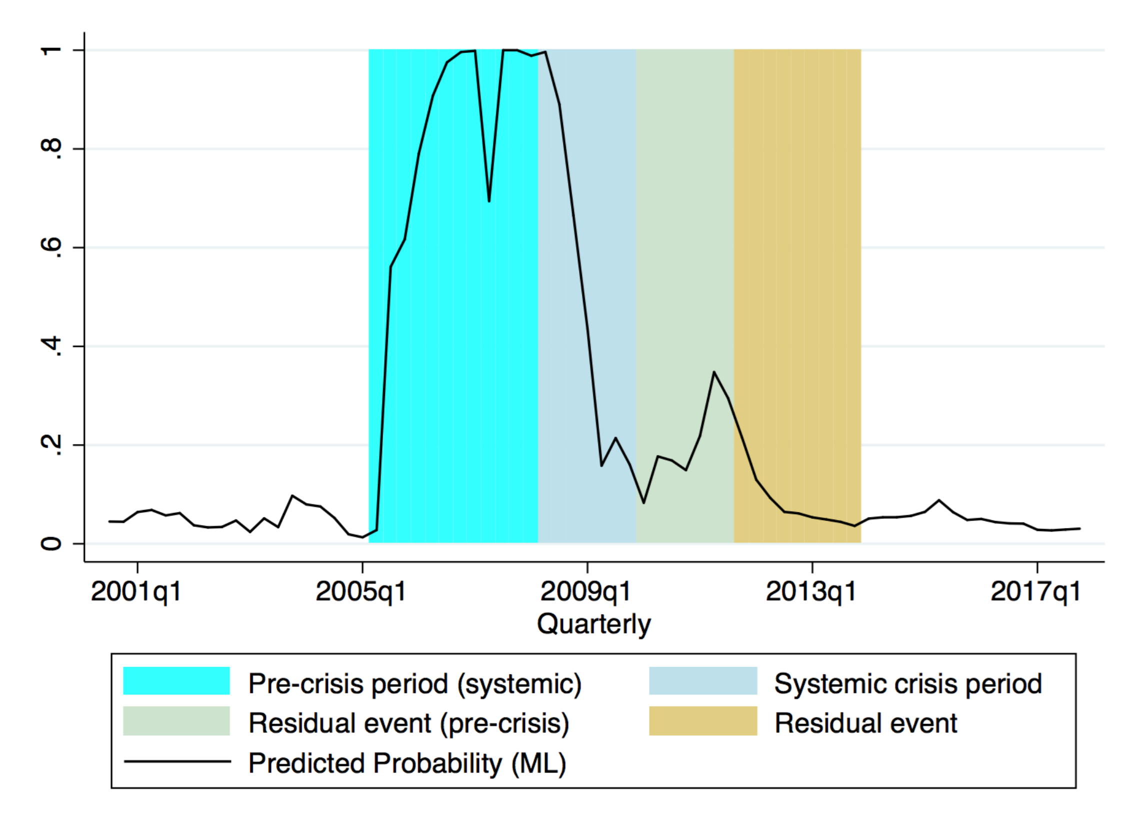

Figure 33: France: Predicted probability - PML (quasi-real time)

E Results: France

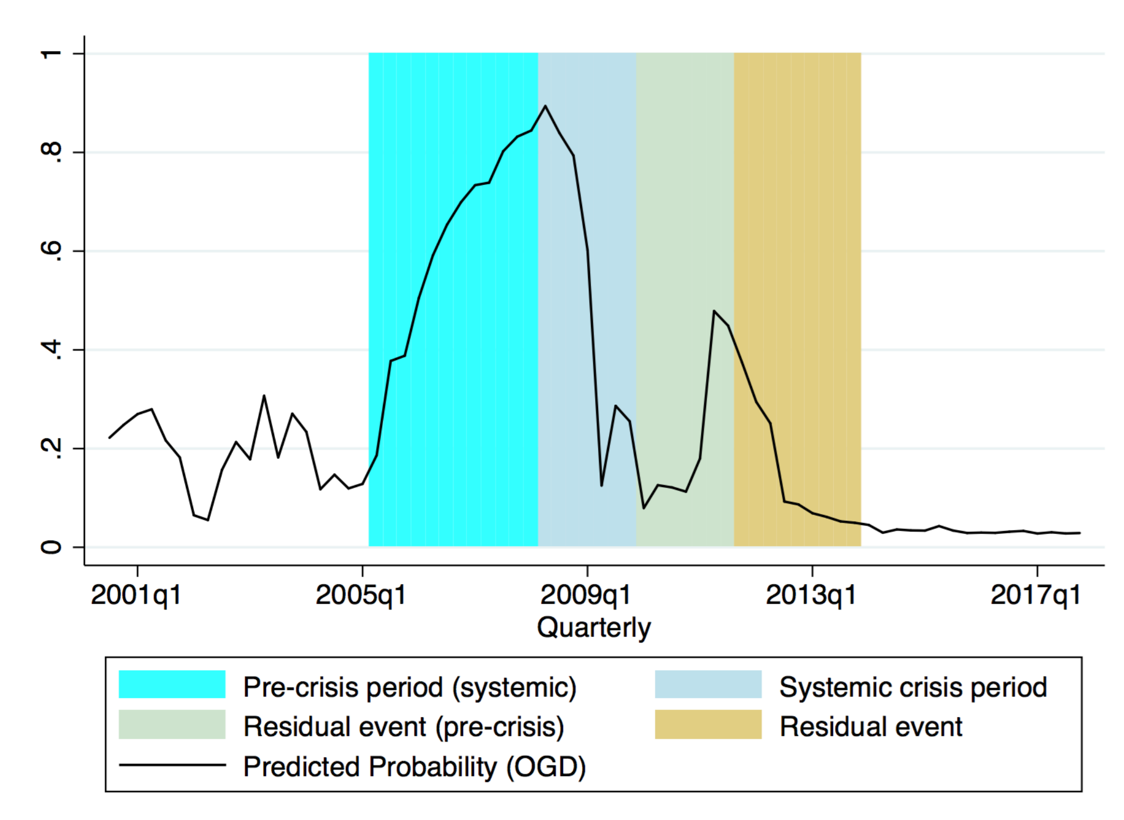

Figure 34: France: Predicted probability - OGD (quasi-real time)

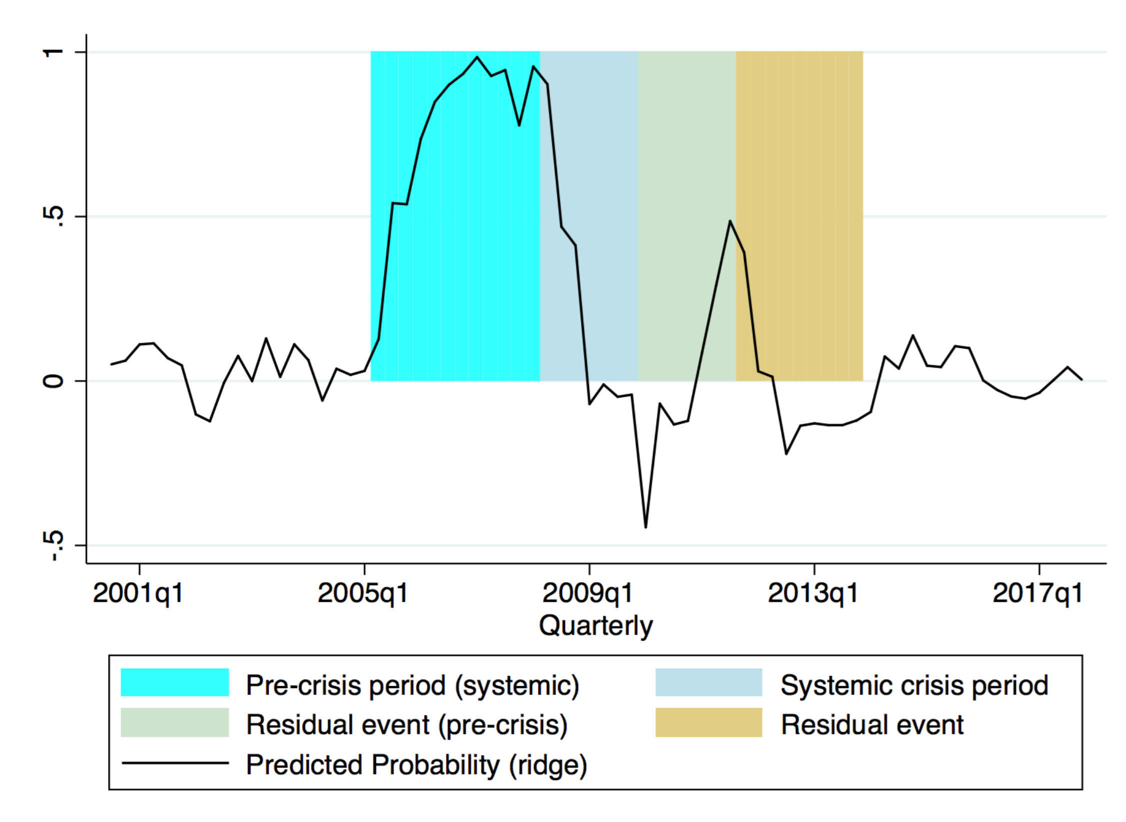

Figure 35: France: Predicted probability - Ridge (quasi-real time)

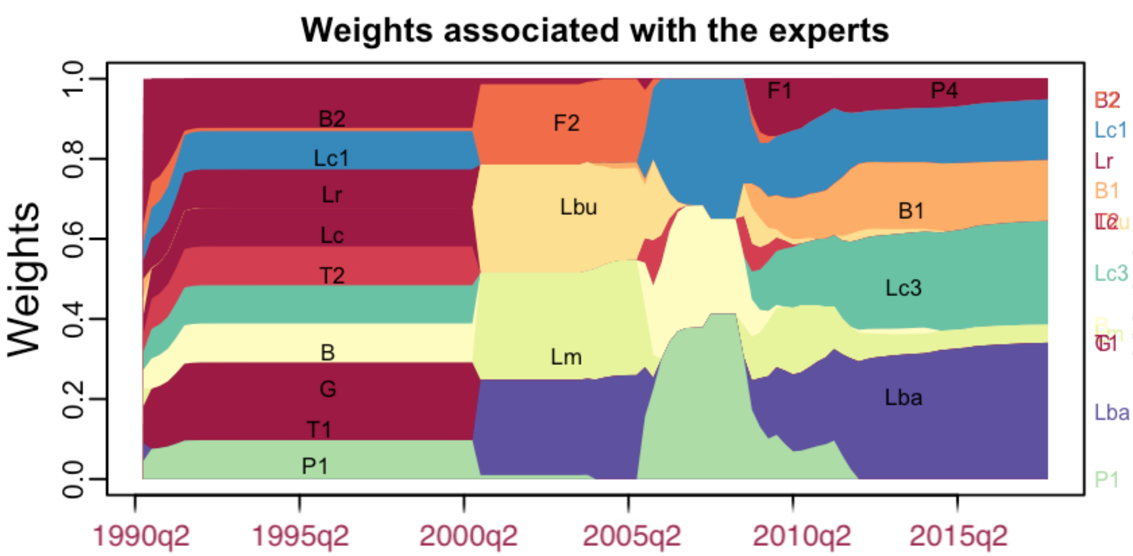

Figure 36: France: Weights. quasi-real time. ML

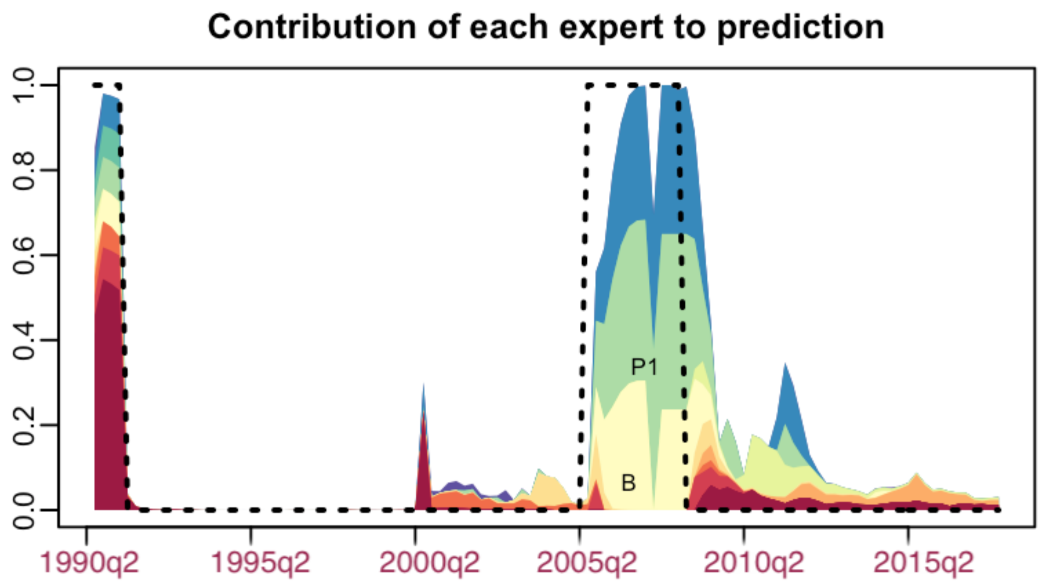

Figure 37: France: Experts contribution to forecast. quasi-real time. ML

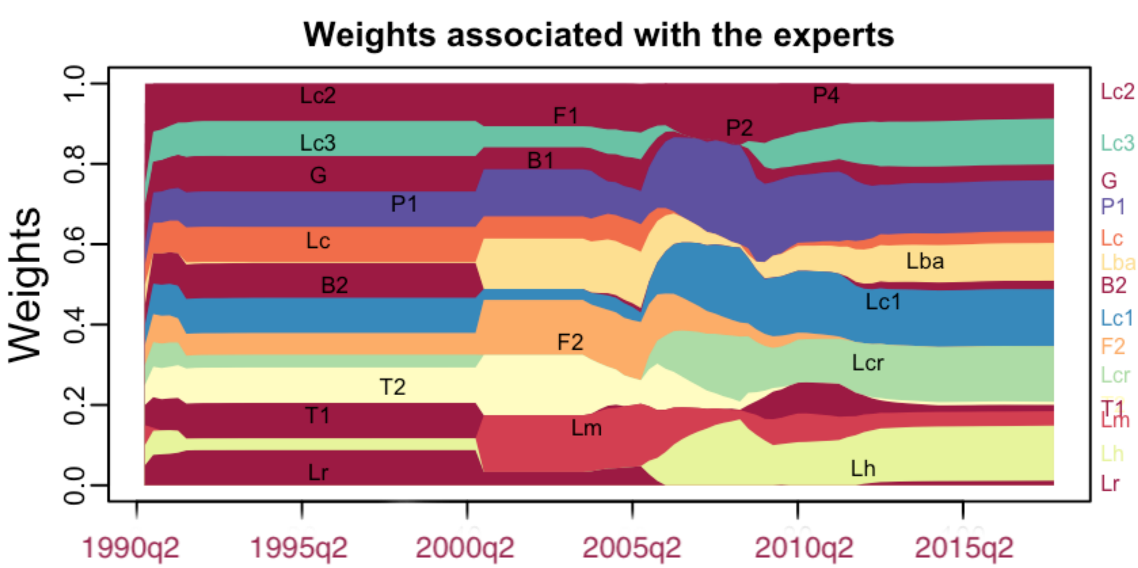

Figure 38: France: Weights. quasi-real time. OGD

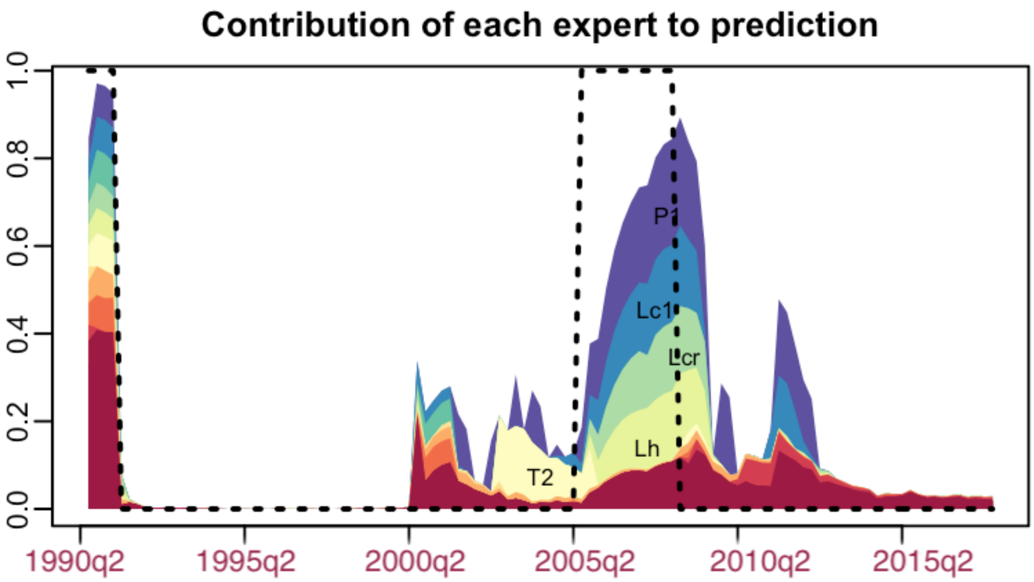

Figure 39: France: Experts contribution to forecast. quasi-real time. OGD

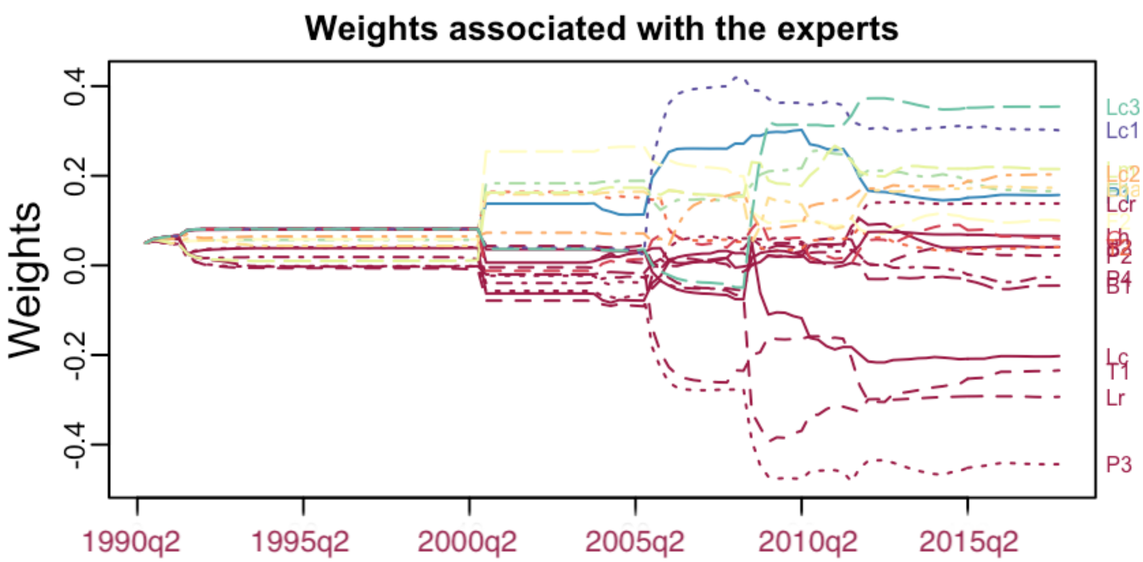

Figure 40: France: Weights. quasi-real time. Ridge

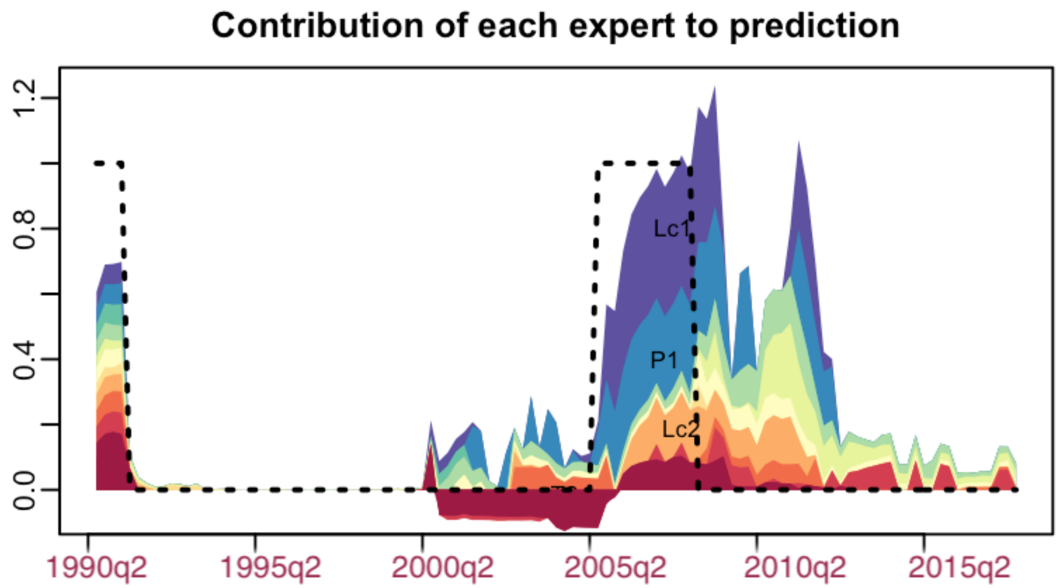

Figure 41: France: Experts contribution to forecast. quasi-real time. Ridge

F Results: Germany

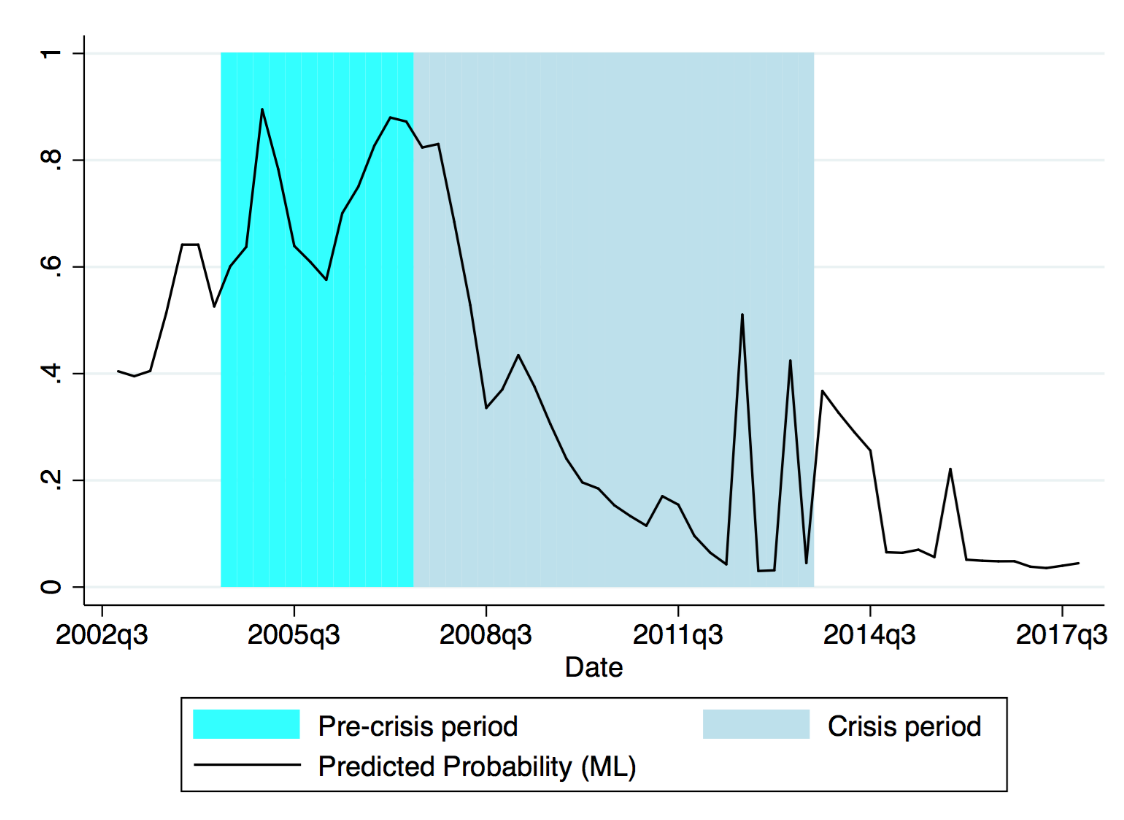

Figure 42: Germany: Predicted probability - ML (quasi-real time from 2002Q3 to 2017Q4)

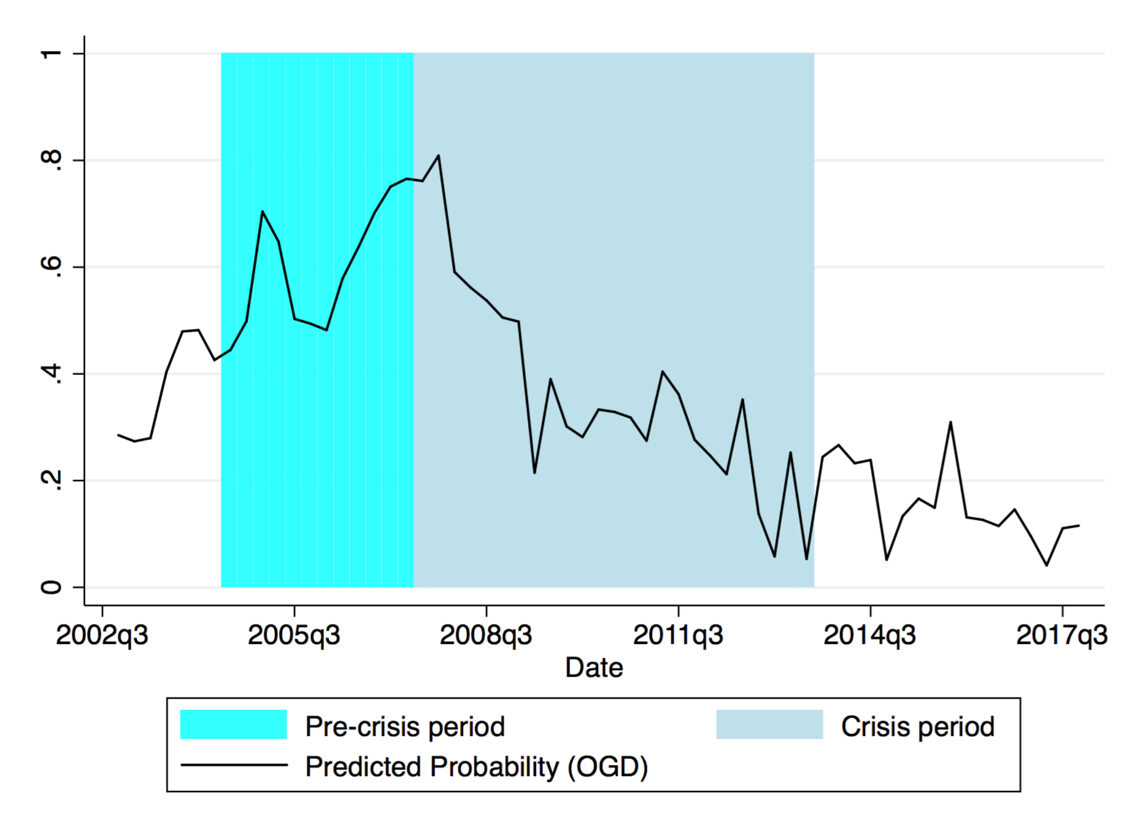

Figure 43: Germany: Predicted probability - OGD (quasi-real time from 2002Q3 to 2017Q4)

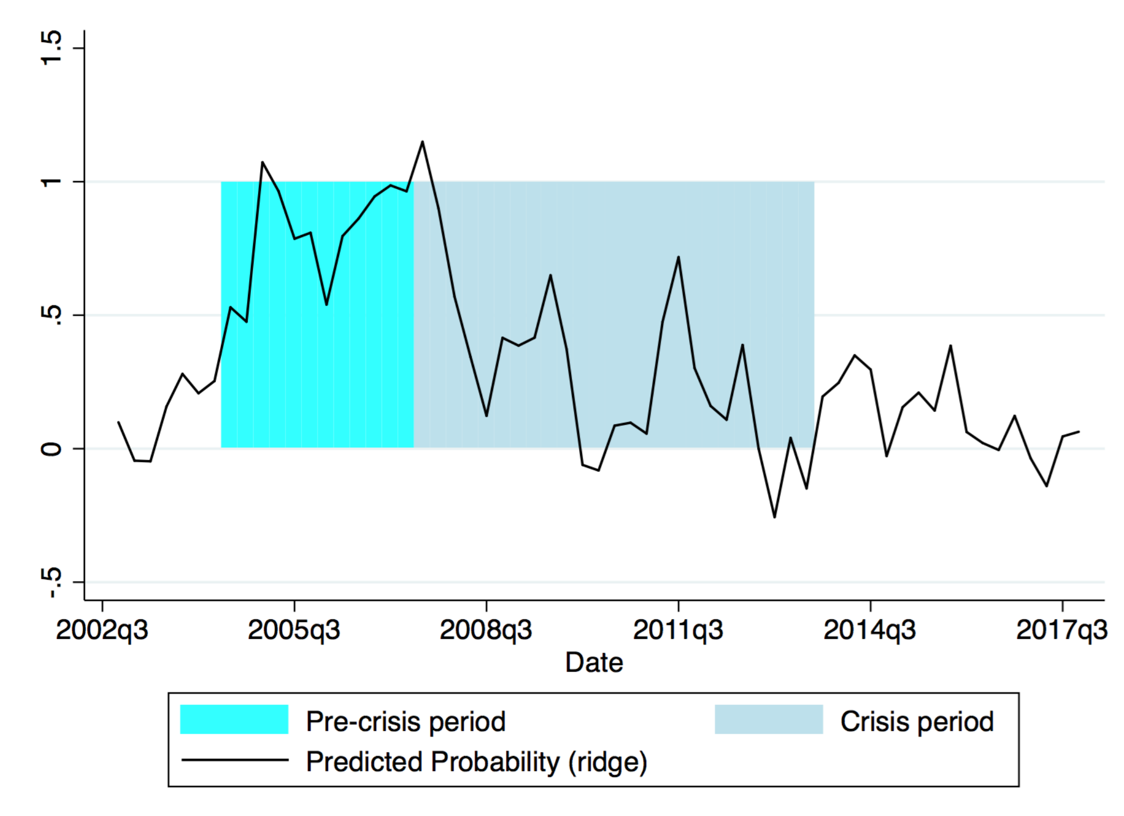

Figure 44: Germany: Predicted probability - Ridge (quasi-real time from 2002Q3 to 2017Q4)

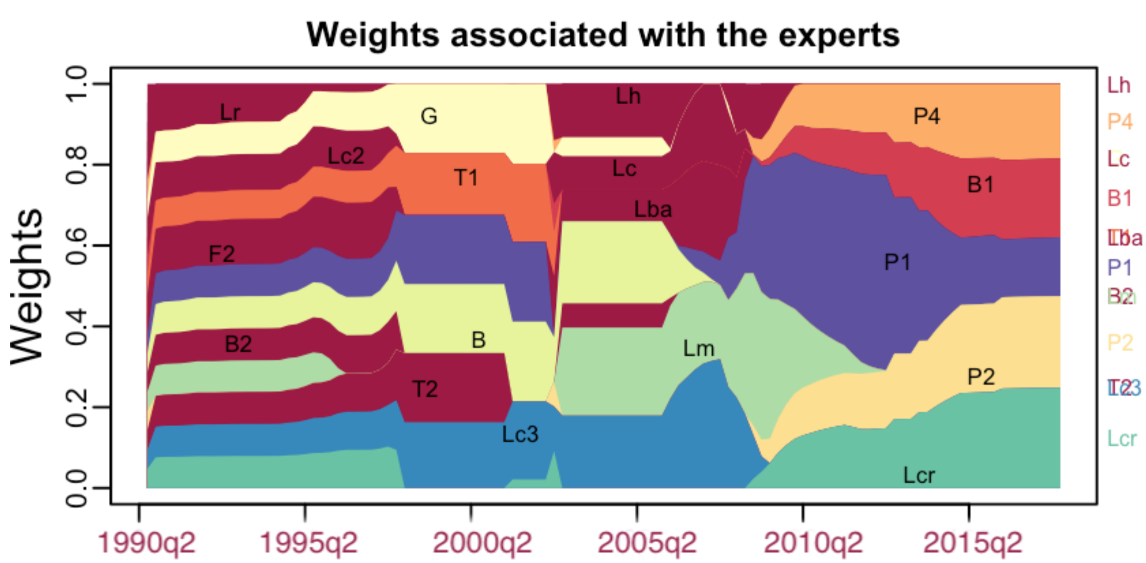

Figure 45: Germany: Weights. quasi-real time. ML

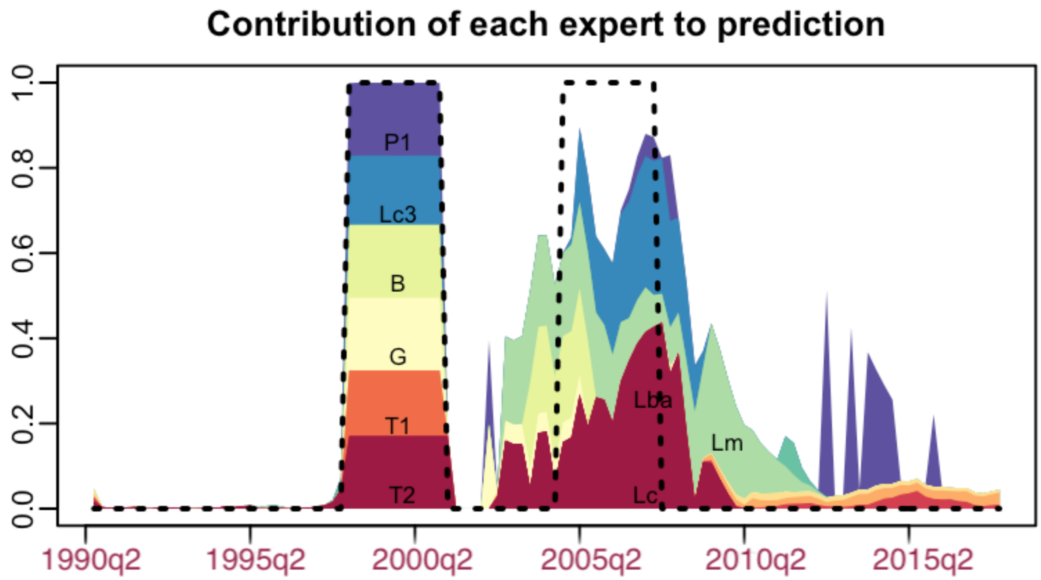

Figure 46: Germany: Experts contribution to forecast. quasi-real time. ML

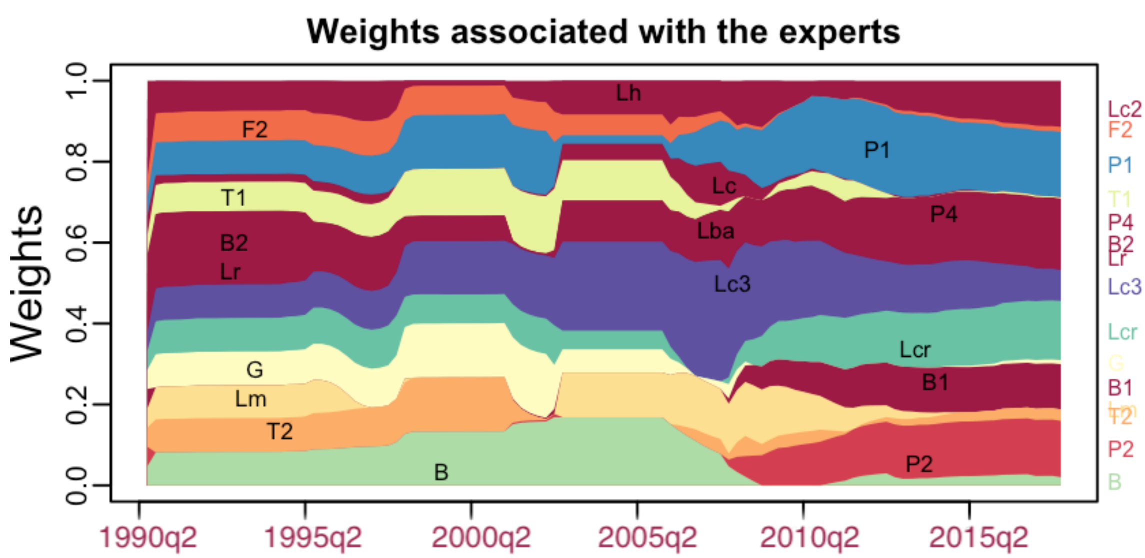

Figure 47: Germany: Weights. quasi-real time. OGD

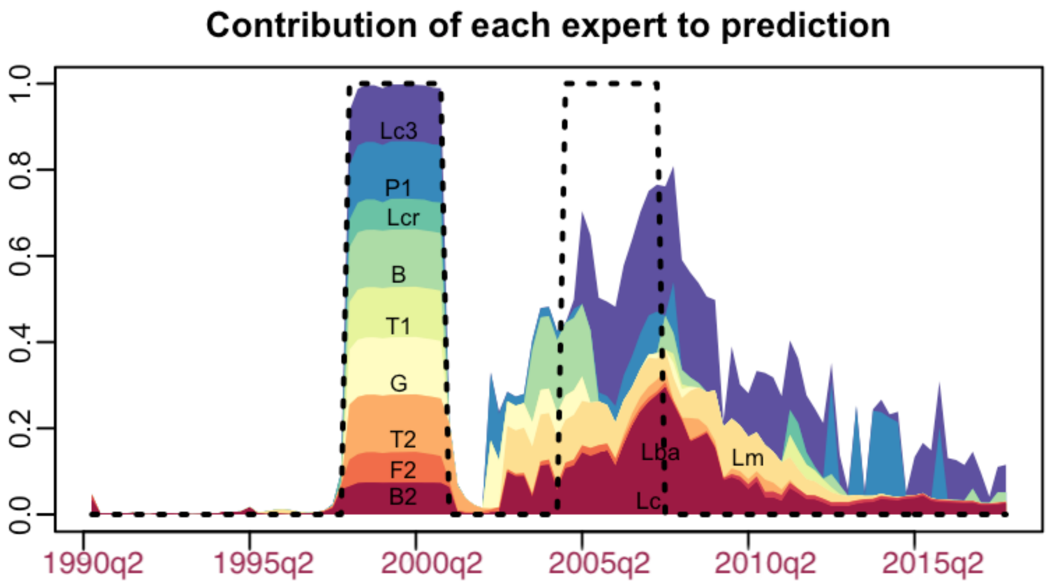

Figure 48: Germany: Experts contribution to forecast. quasi-real time. OGD

Figure 49: Germany: Weights. quasi-real time. Ridge

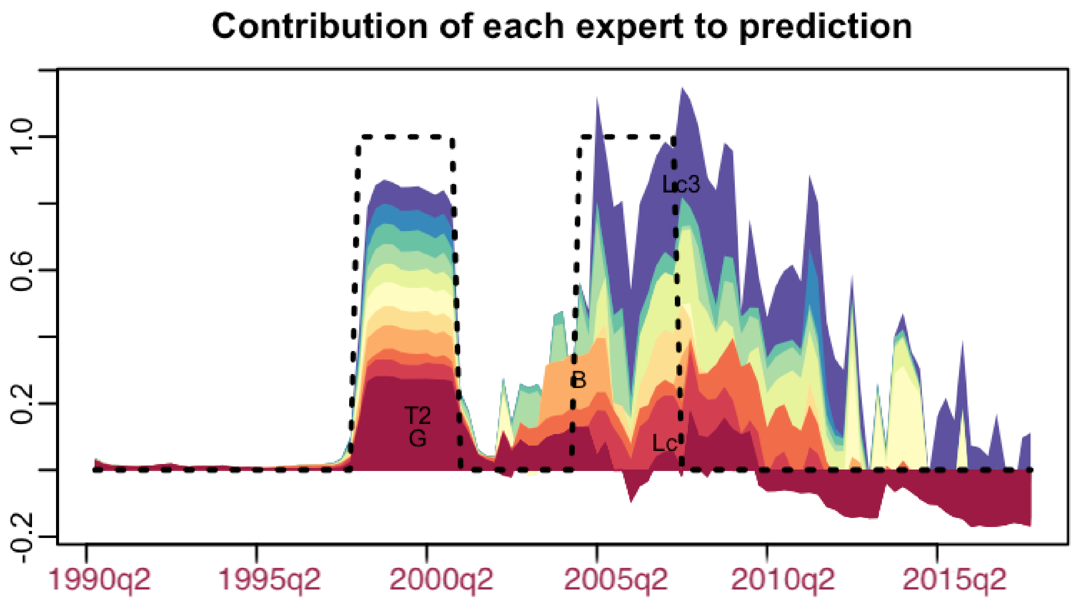

Figure 50: Germany: Experts contribution to forecast. quasi-real time. Ridge

G Results: UK

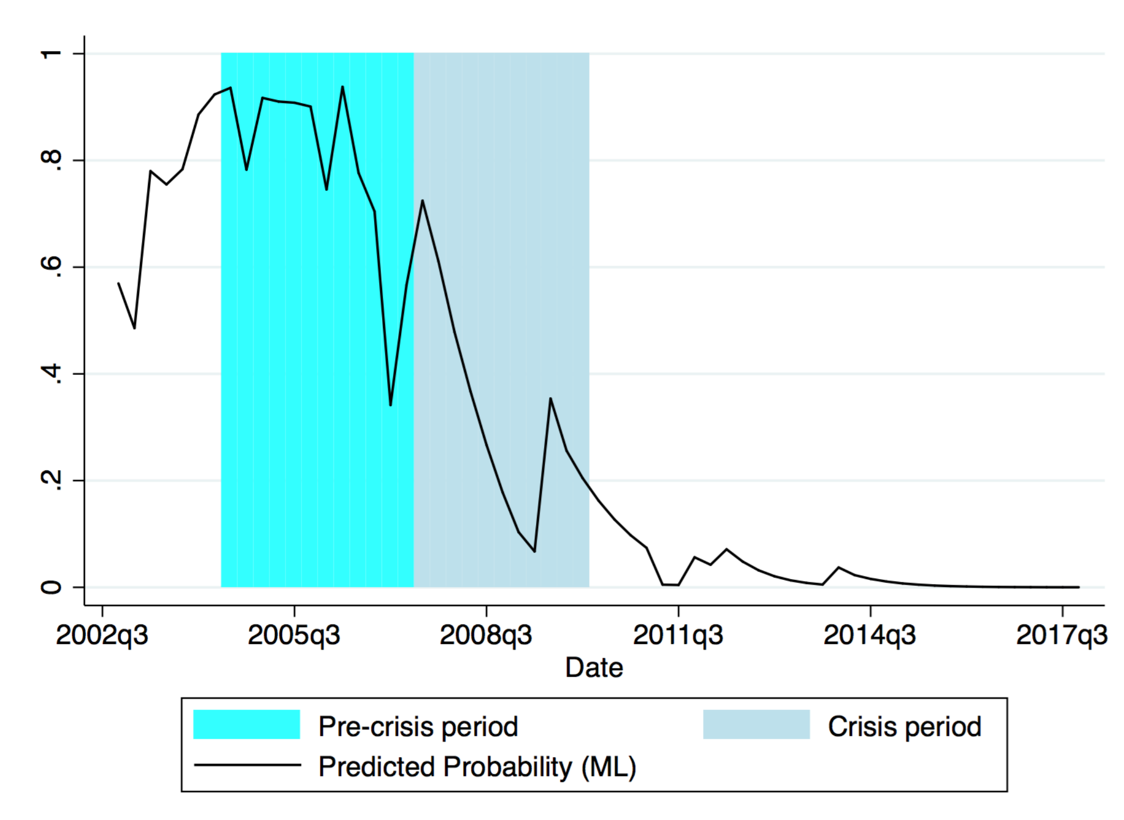

Figure 51: UK. Predicted probability - ML (quasi-real time from 2002Q3 to 2017Q4)

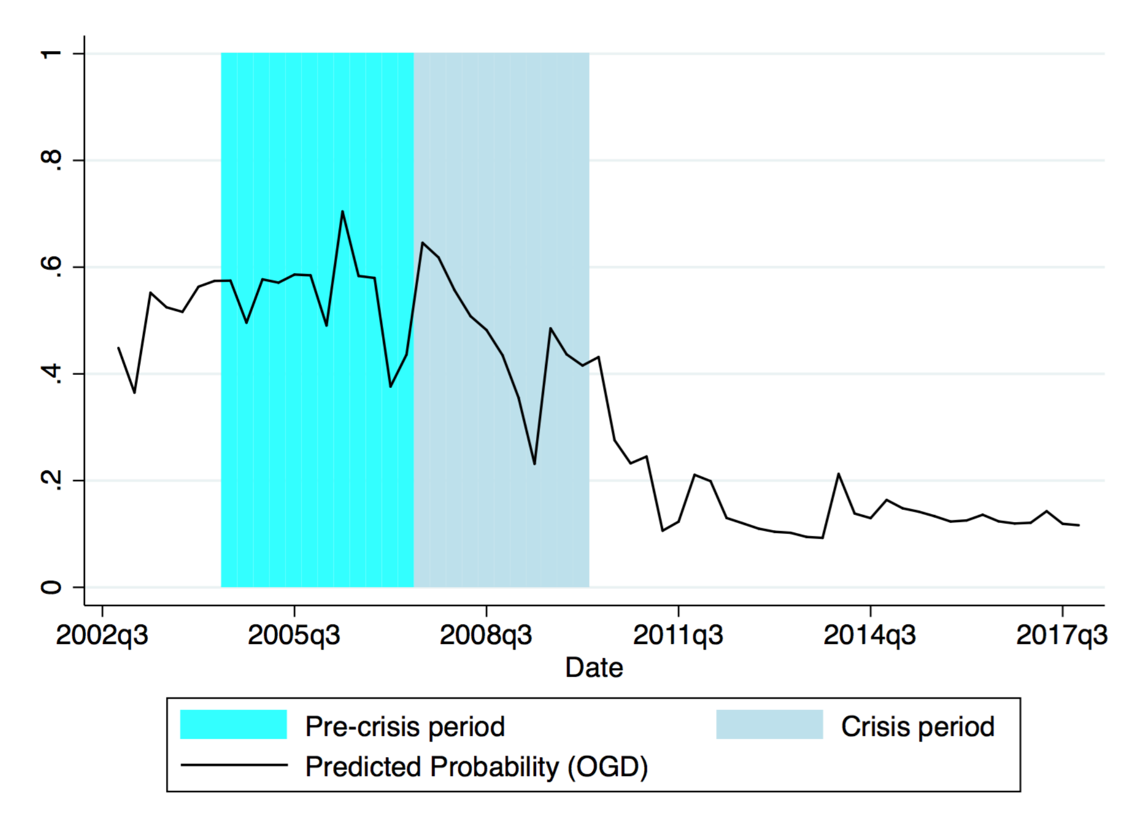

Figure 52: UK. Predicted probability - OGD (quasi-real time from 2002Q3 to 2017Q4)

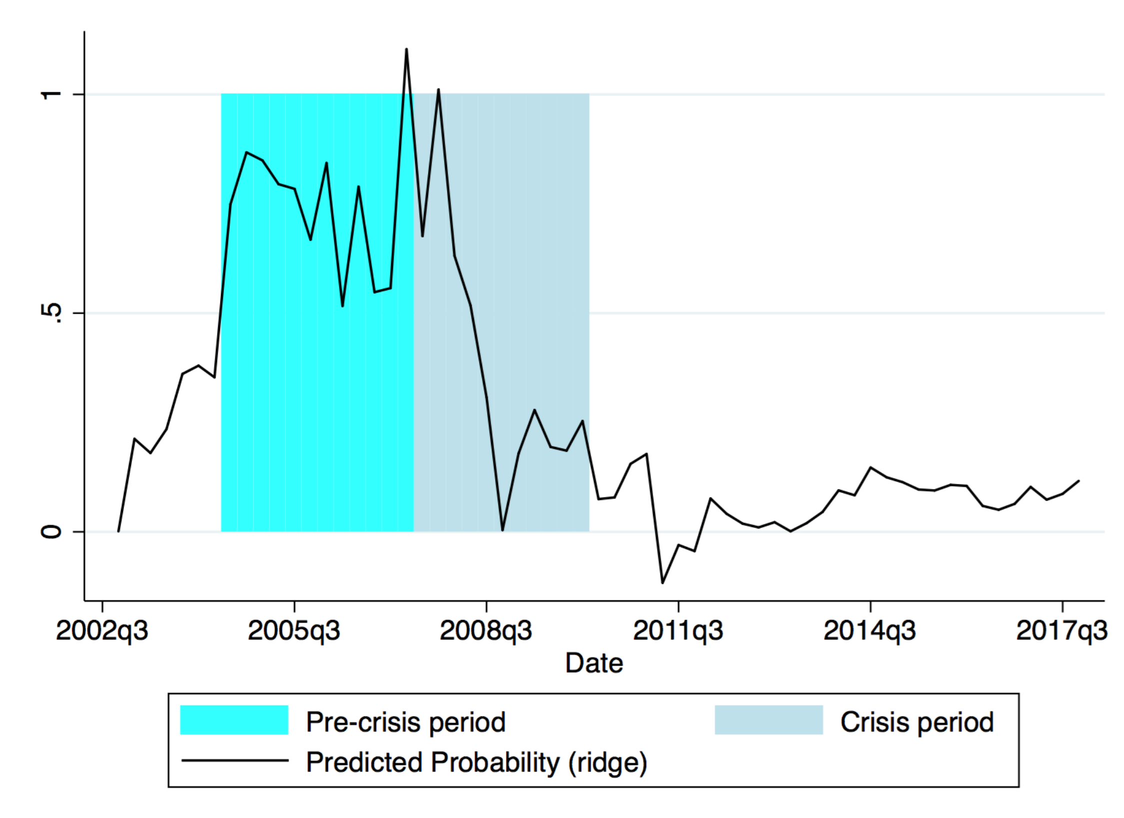

Figure 53: UK. Predicted probability - Ridge (quasi-real time from 2002Q3 to 2017Q4)

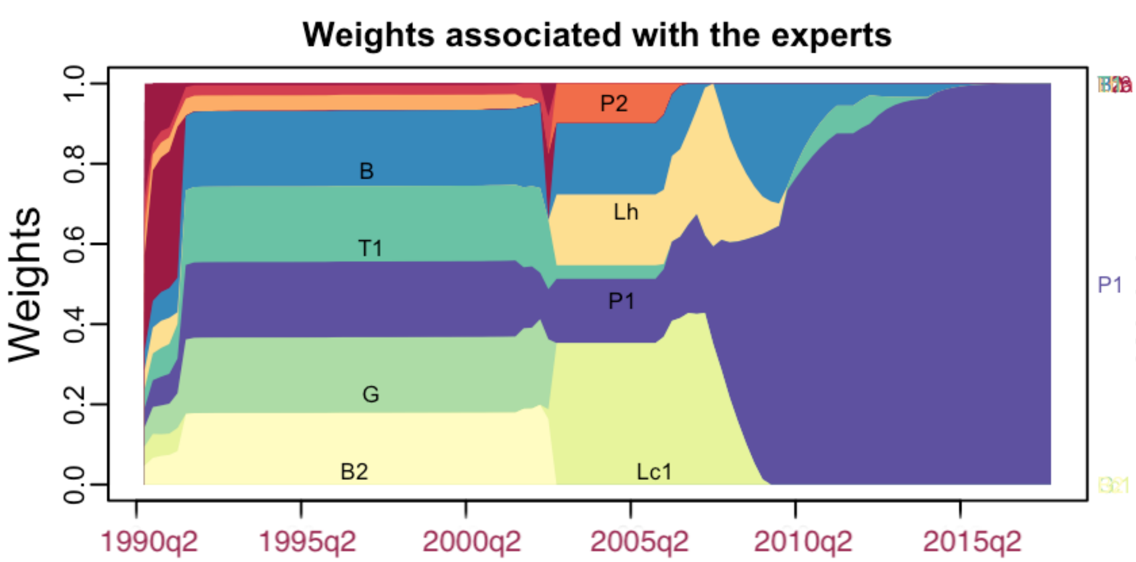

Figure 54: UK: Weights. quasi-real time. ML

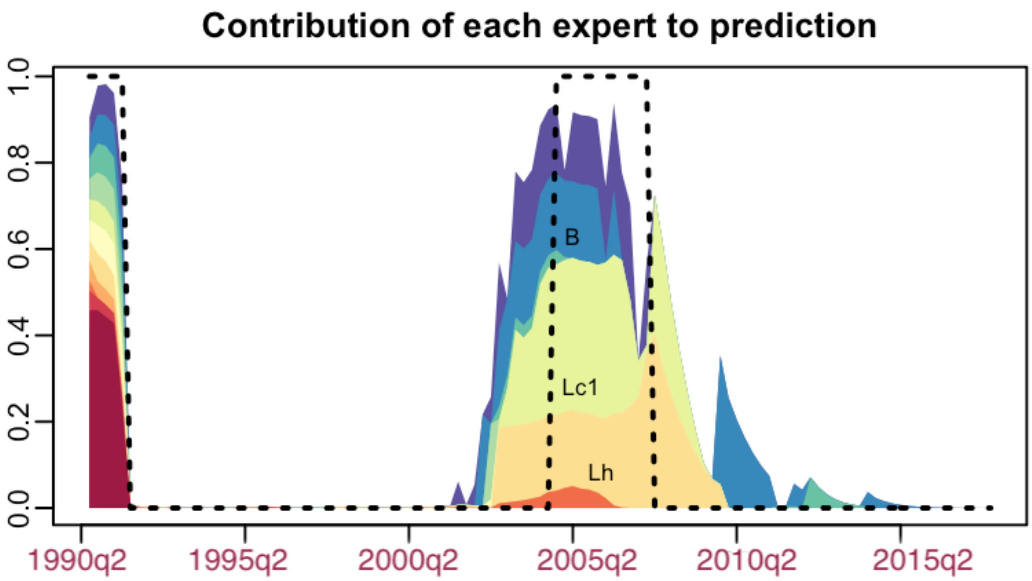

Figure 55: UK: Experts contribution to forecast. quasi-real time. ML

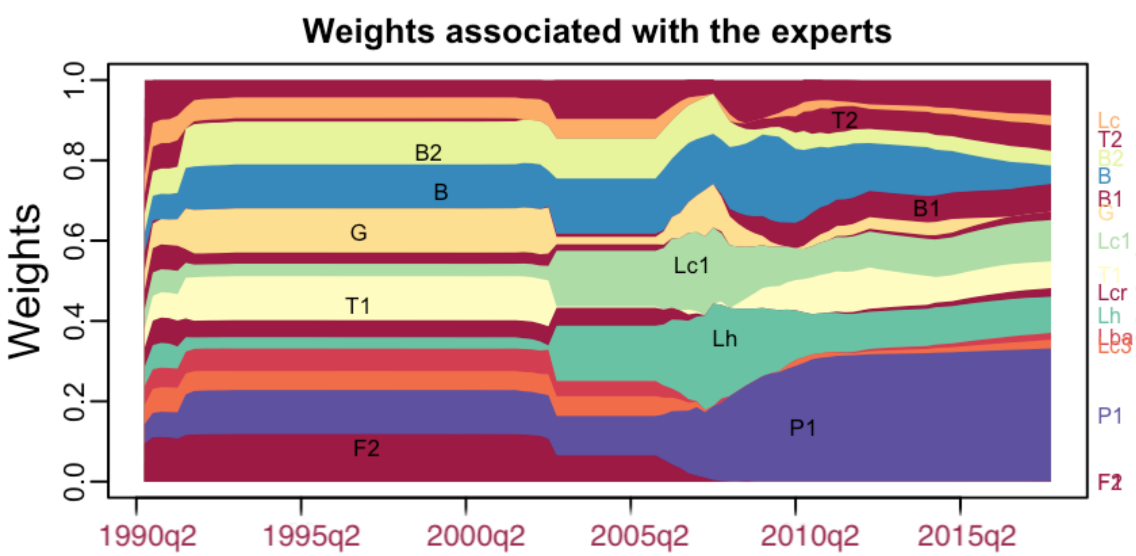

Figure 56: UK: Weights. quasi-real time. OGD

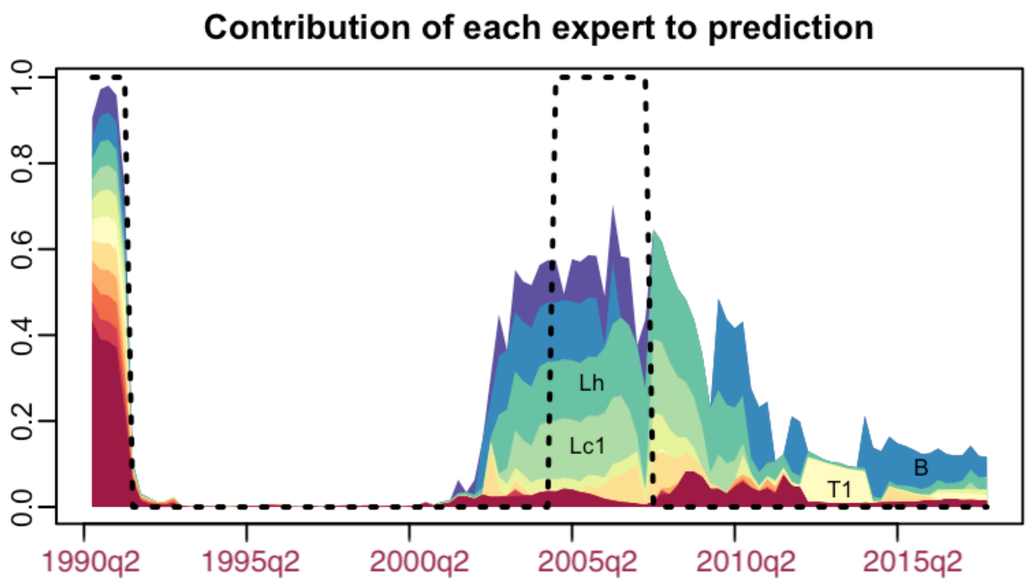

Figure 57: UK: Experts contribution to forecast. quasi-real time. OGD

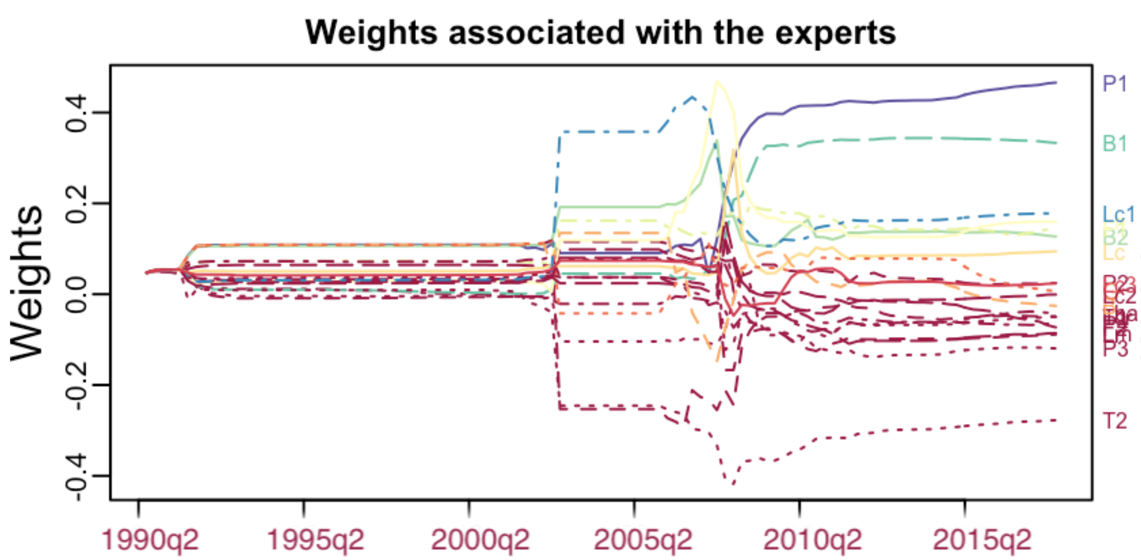

Figure 58: UK: Weights. quasi-real time. Ridge

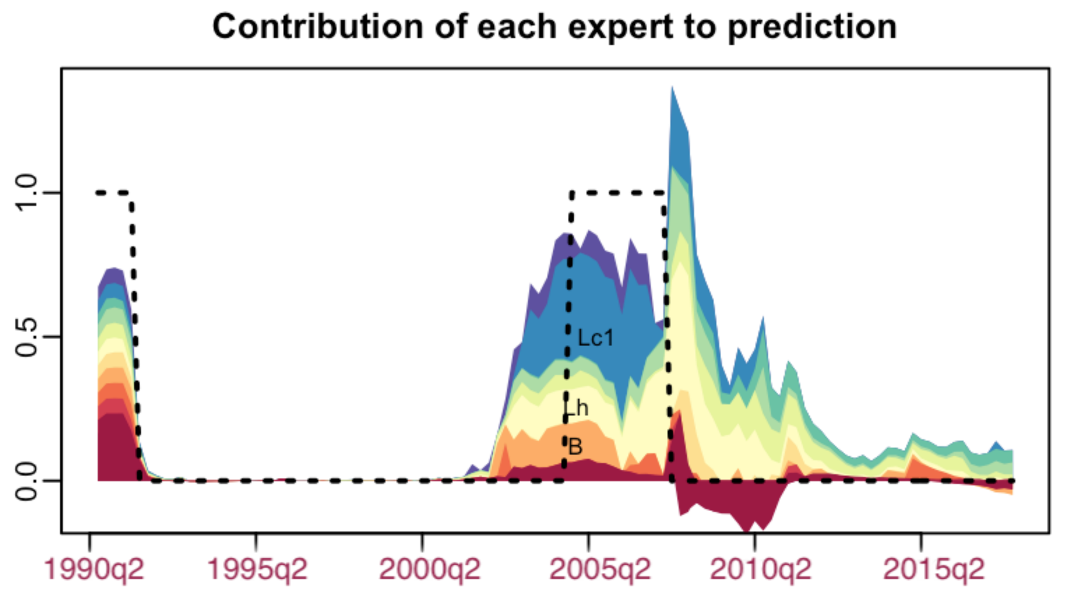

Figure 59: UK: Experts contribution to forecast. quasi-real time. Ridge

H Results: Spain

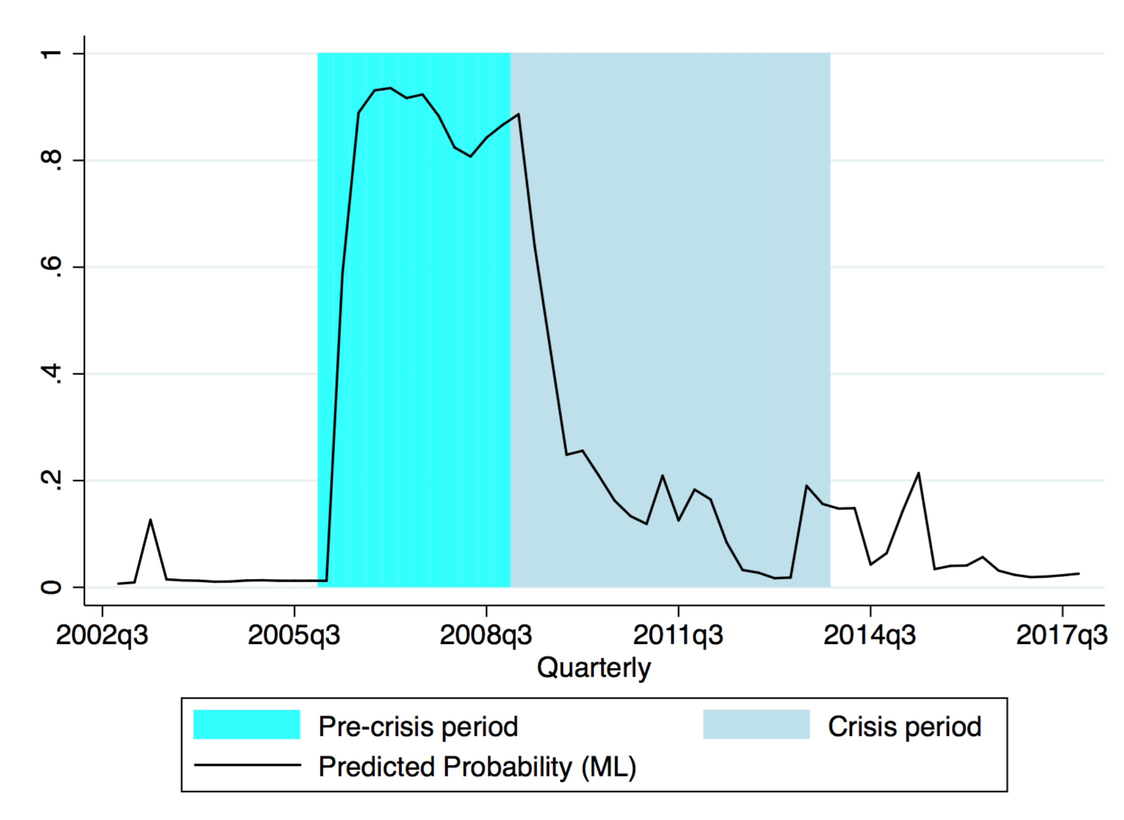

Figure 60: Spain. Predicted probability - ML (quasi-real time from 2002Q3 to 2017Q4)

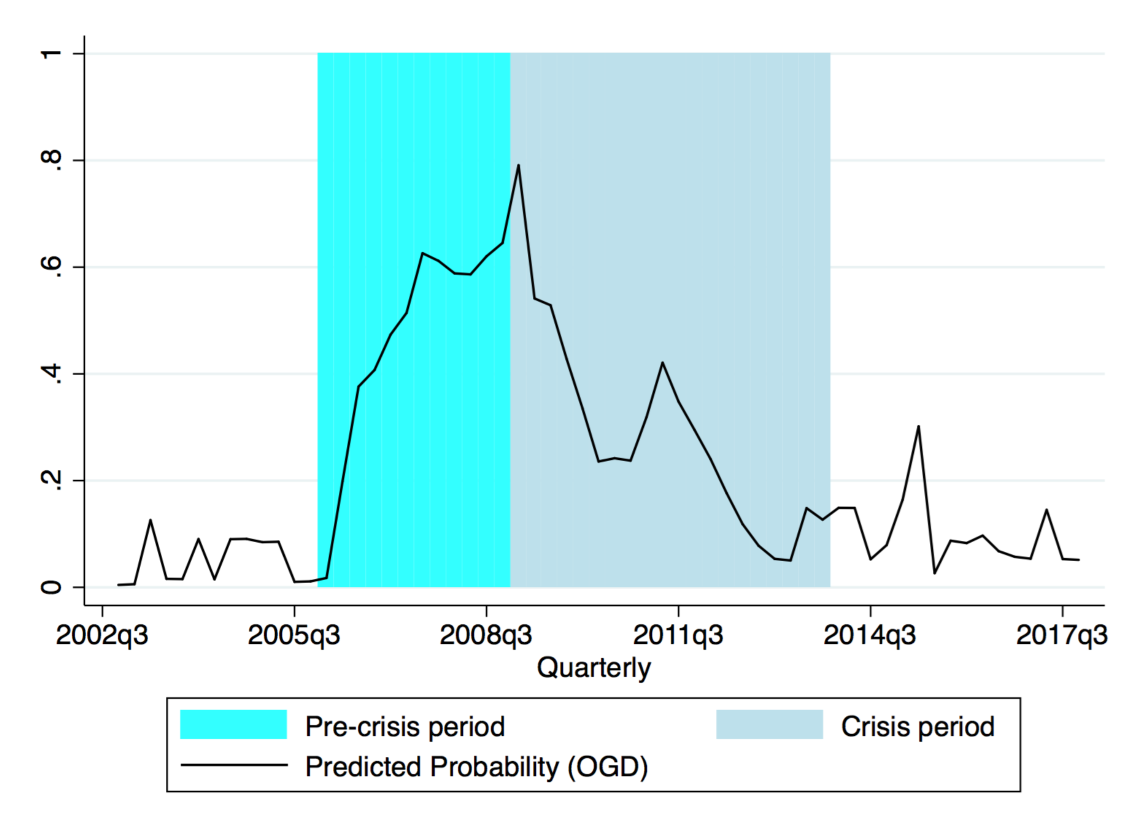

Figure 61: Spain. Predicted probability - OGD (quasi-real time from 2002Q3 to 2017Q4)

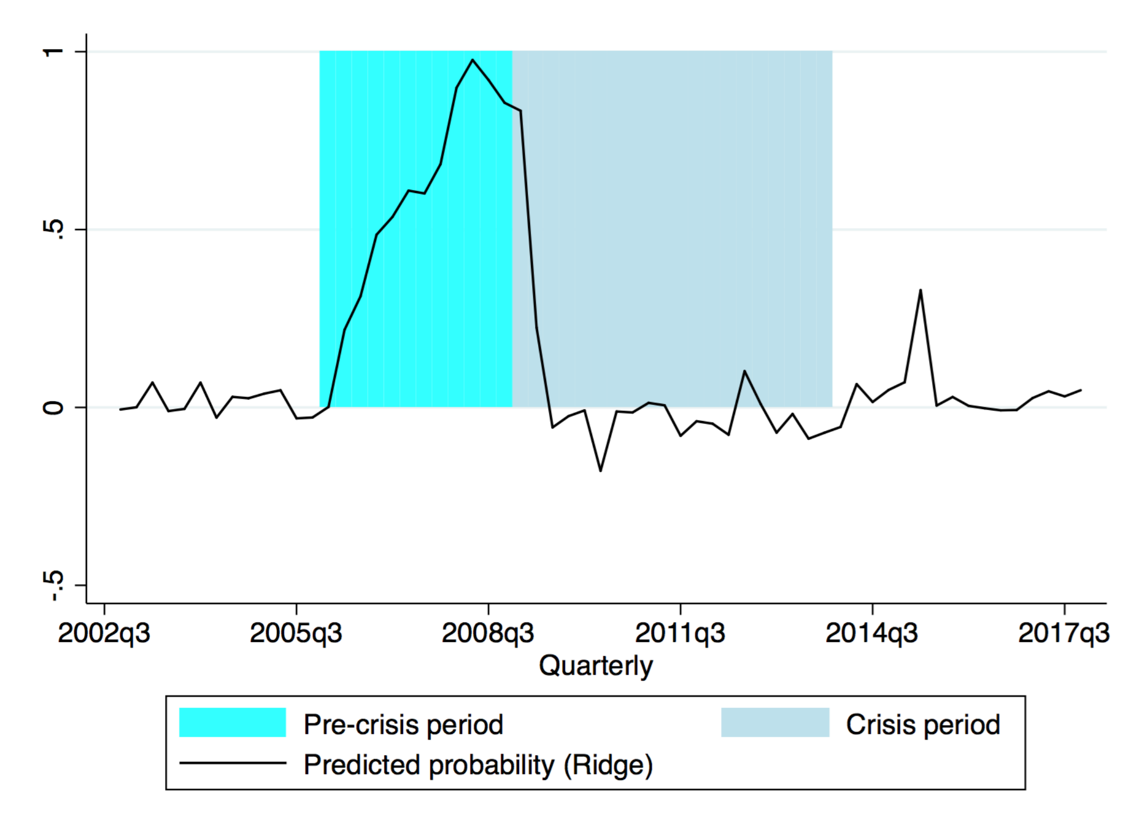

Figure 62: Spain. Predicted probability - Ridge (quasi-real time from 2002Q3 to 2017Q4)

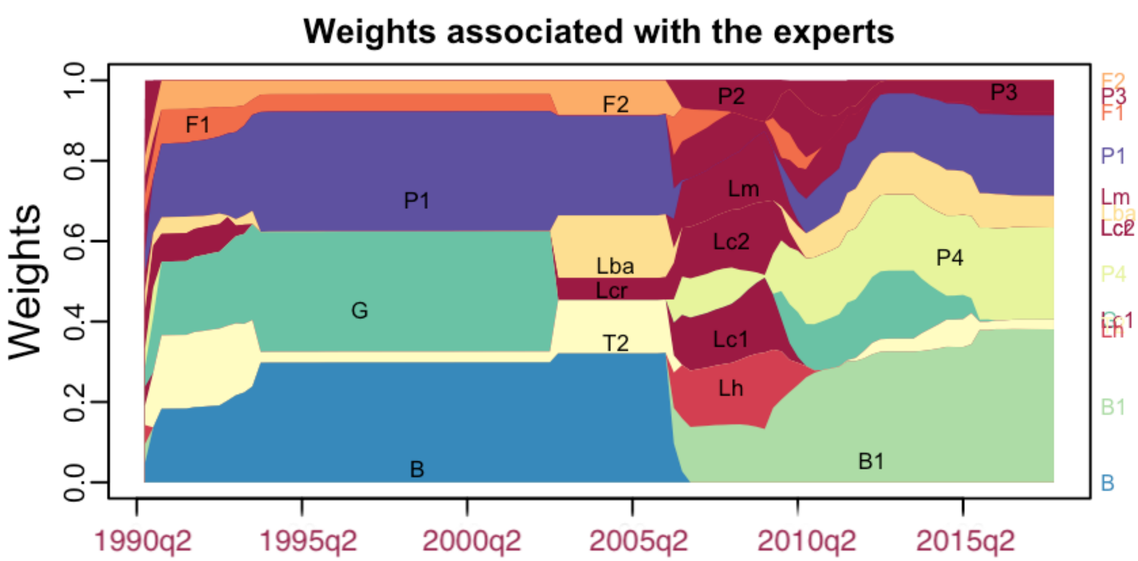

Figure 63: Spain: Weights. quasi-real time. ML

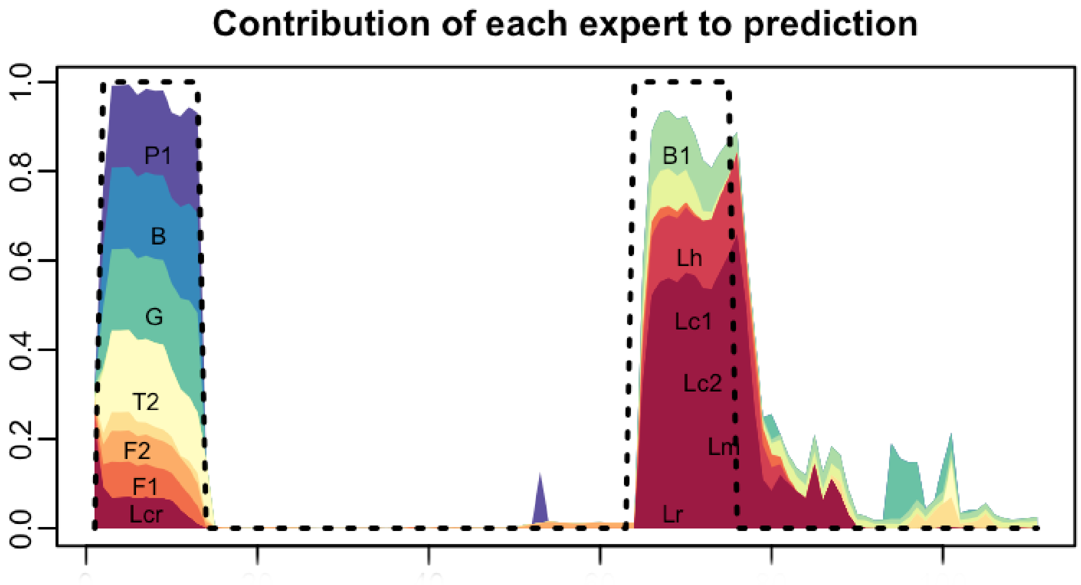

Figure 64: Spain: Experts contribution to forecast. quasi-real time. ML

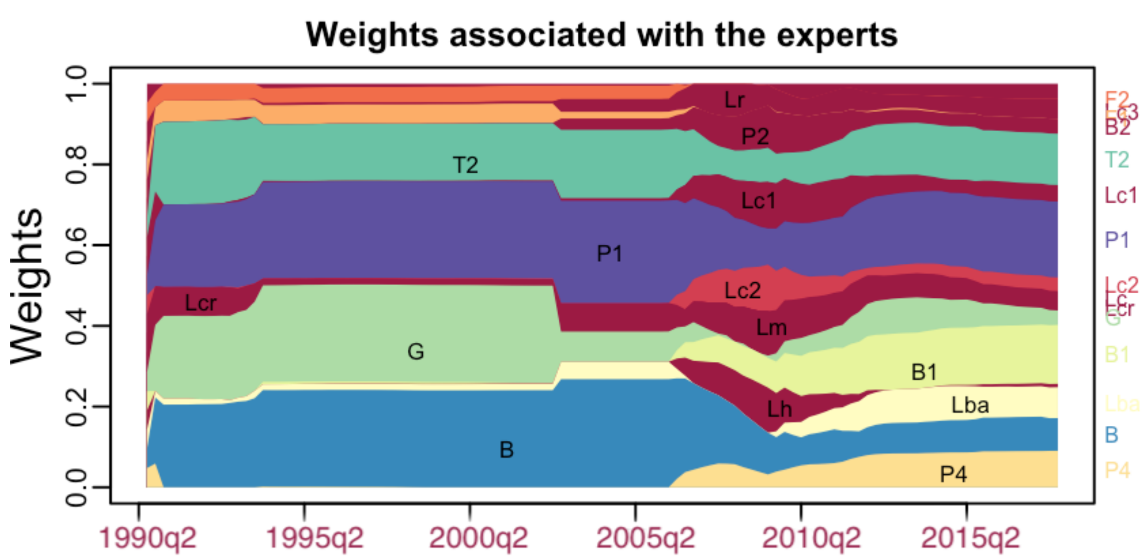

Figure 65: Spain: Weights. quasi-real time. OGD

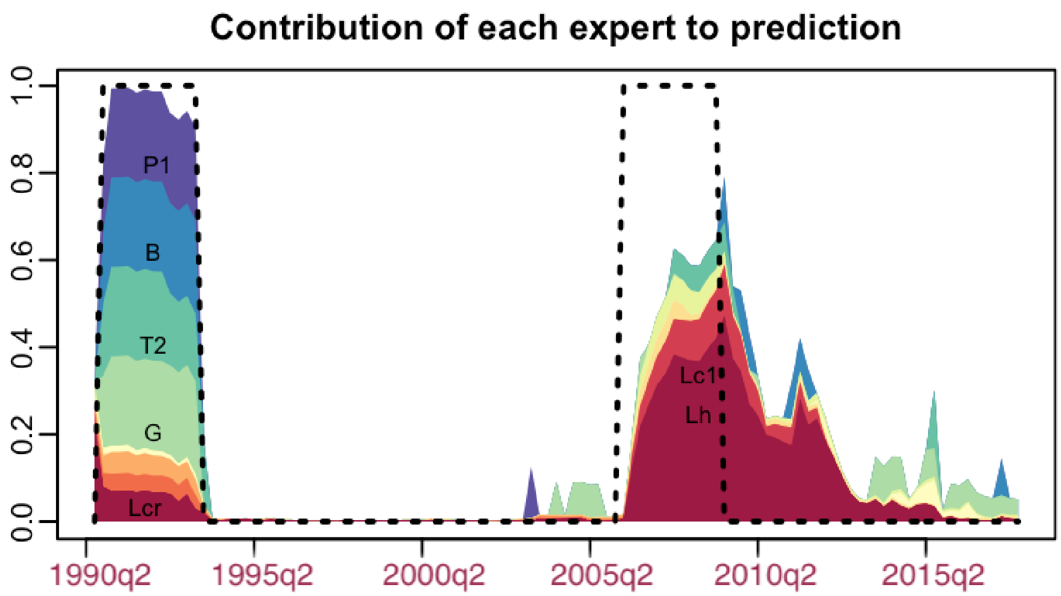

Figure 66: Spain: Experts contribution to forecast. quasi-real time. OGD

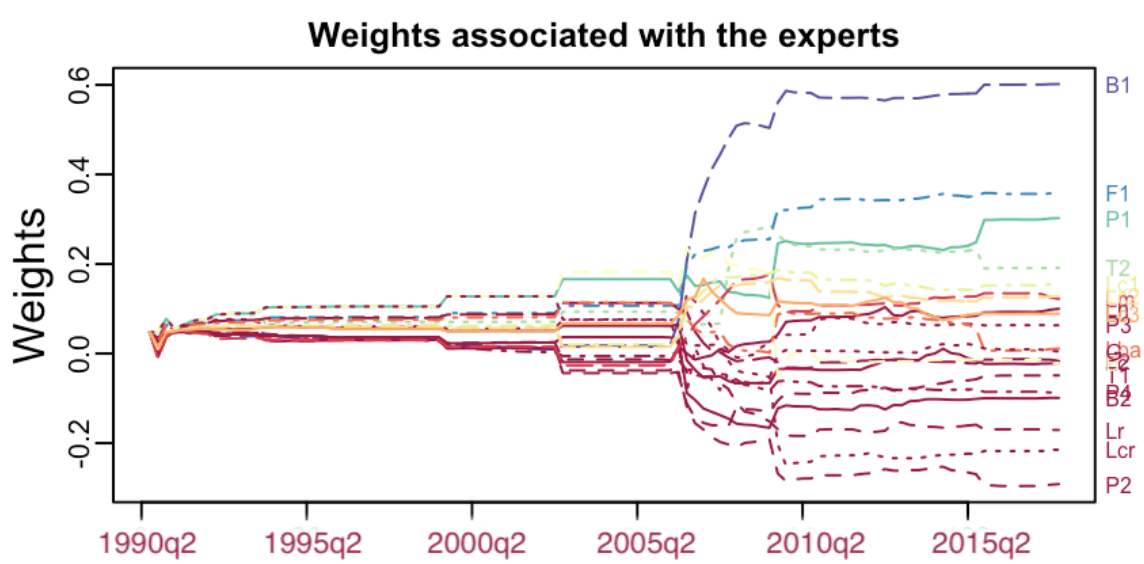

Figure 67: Spain: Weights. quasi-real time. Ridge

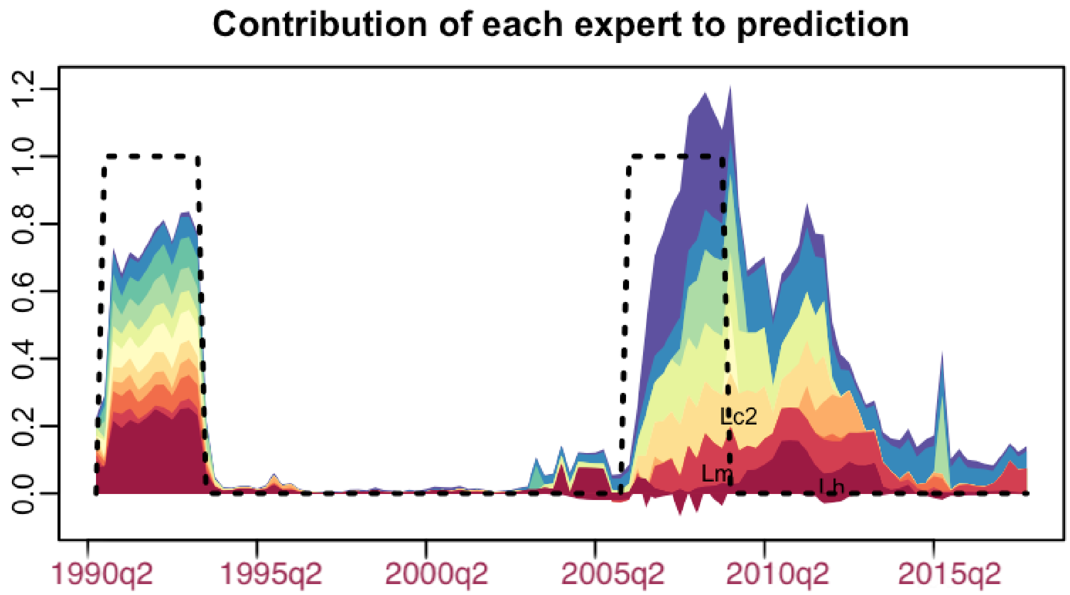

Figure 68: Spain: Experts contribution to forecast. quasi-real time. Ridge

I Results: US

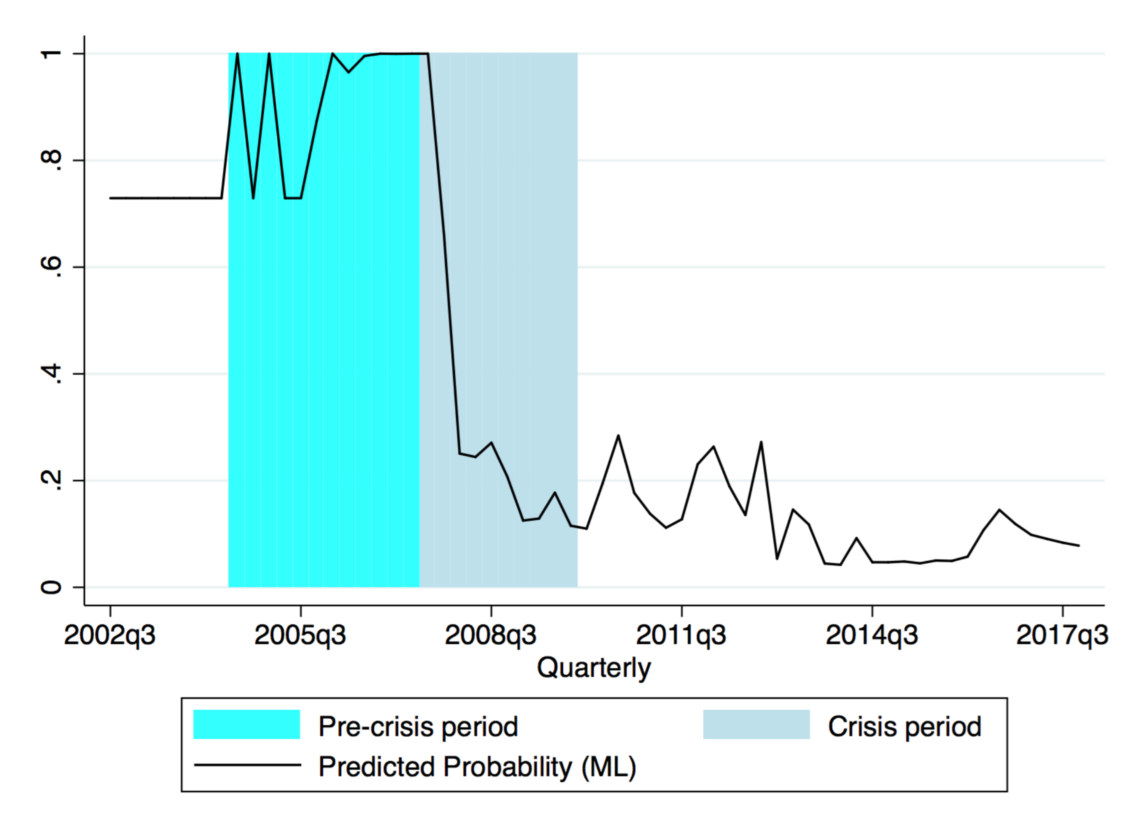

Figure 69: US. Predicted probability - ML (quasi-real time from 2002Q3 to 2017Q4)

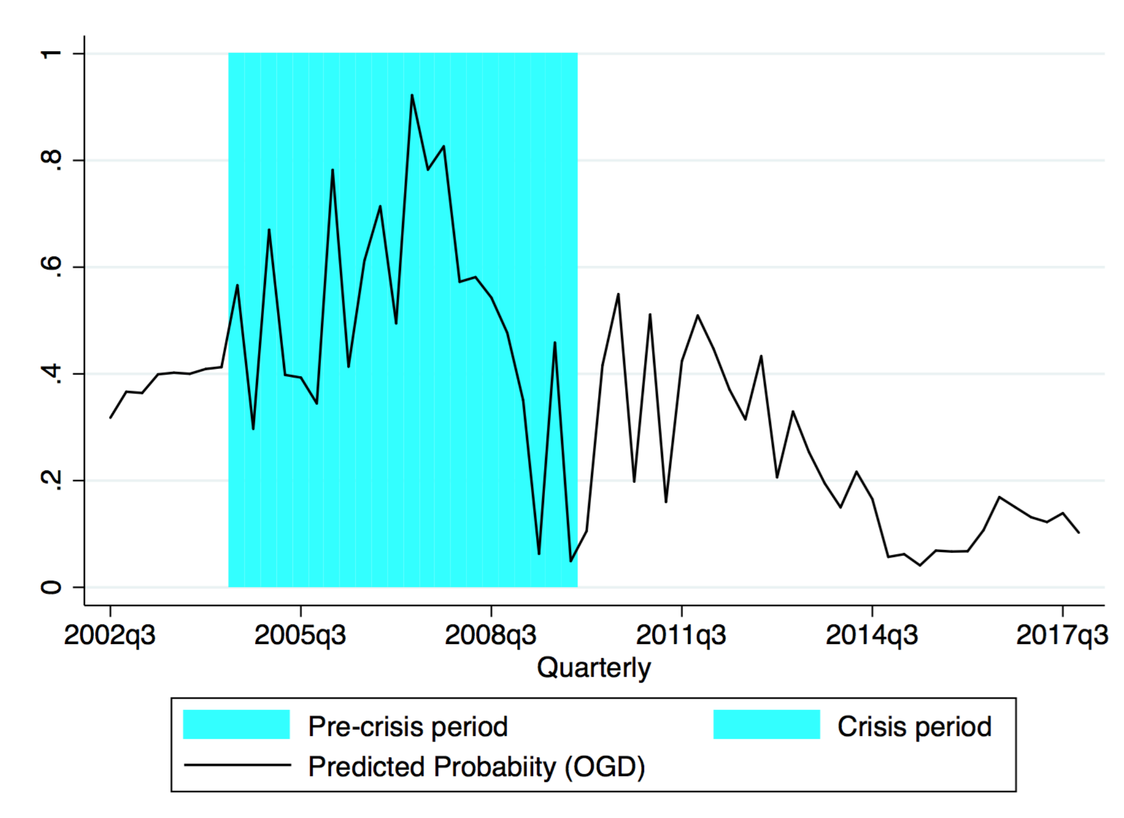

Figure 70: US. Predicted probability - OGD (quasi-real time from 2002Q3 to 2017Q4)

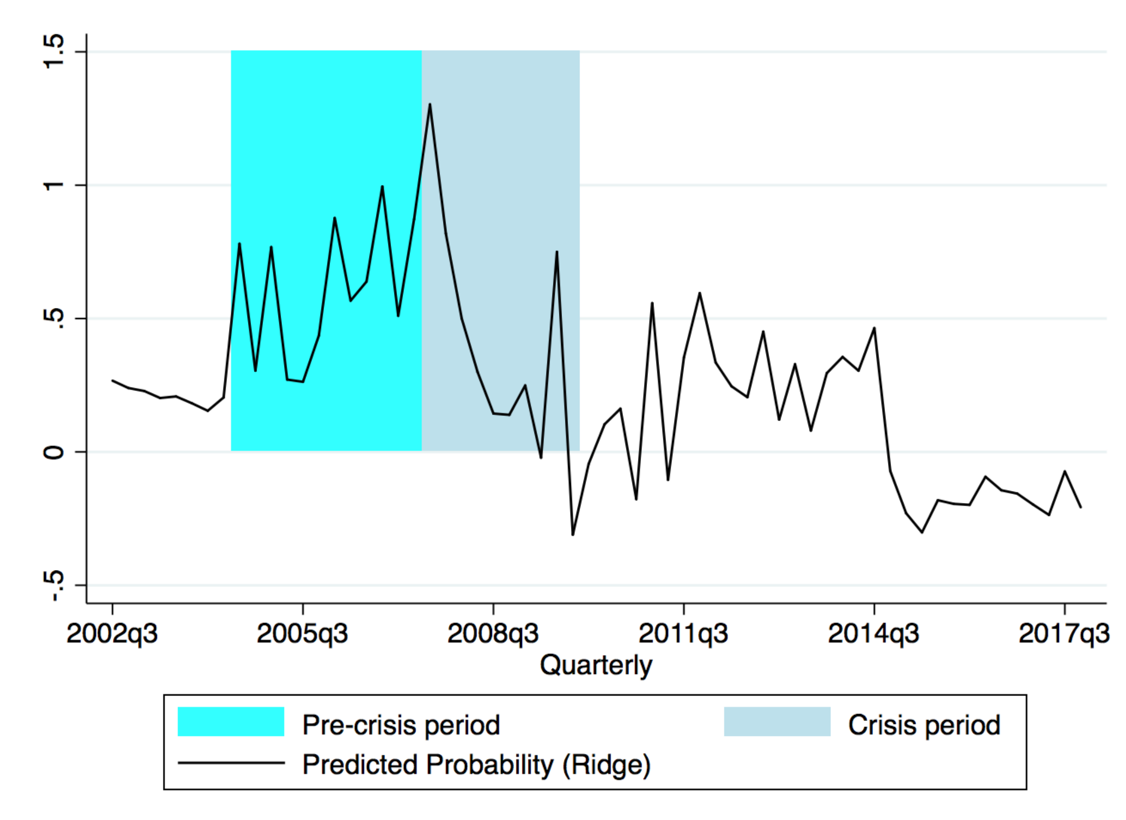

Figure 71: US. Predicted probability - Ridge (quasi-real time from 2002Q3 to 2017Q4)

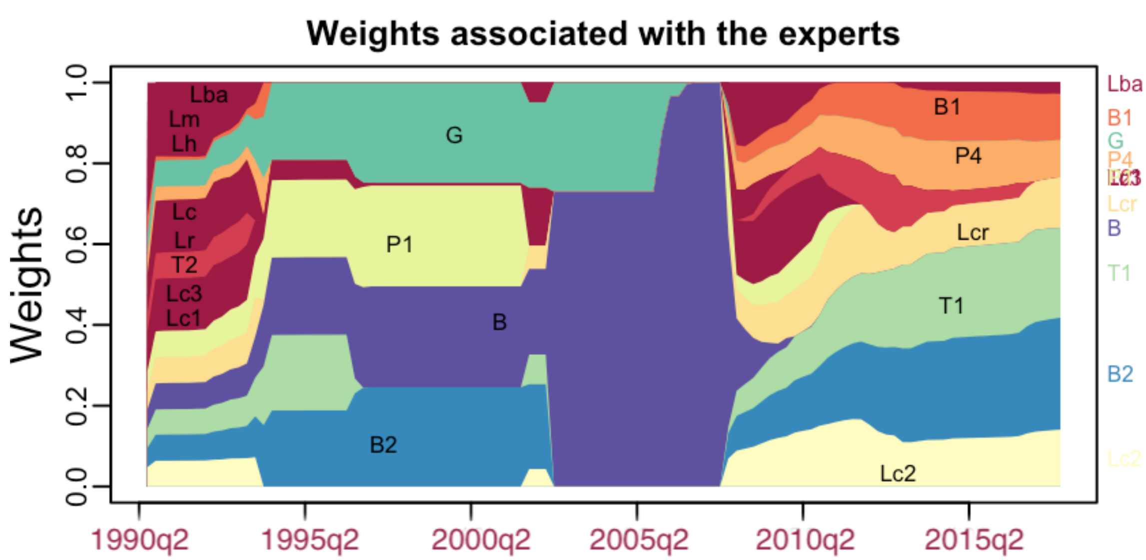

Figure 72: US: Weights. quasi-real time. ML

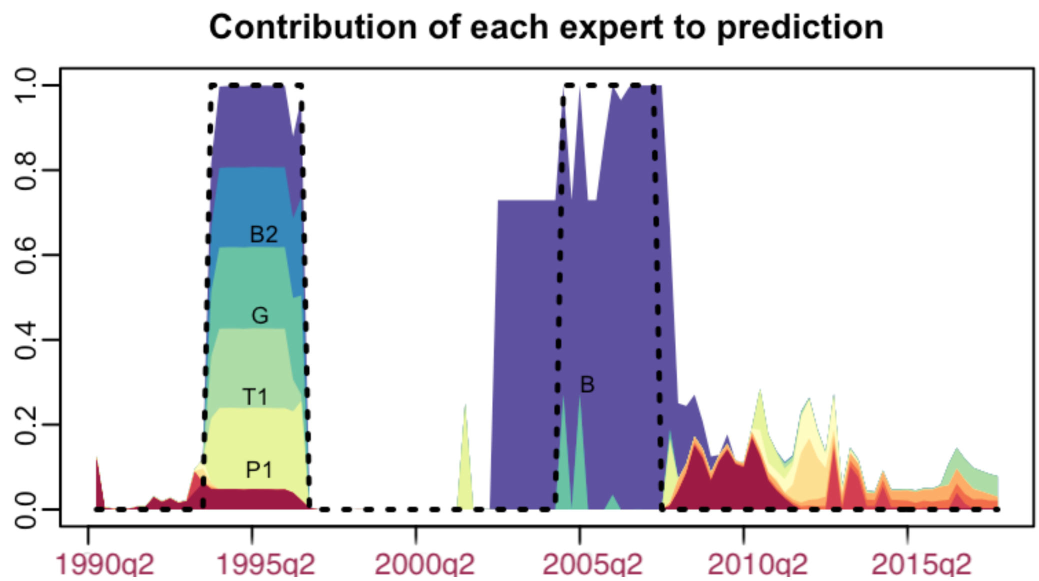

Figure 73: US: Experts contribution to forecast. quasi-real time. ML

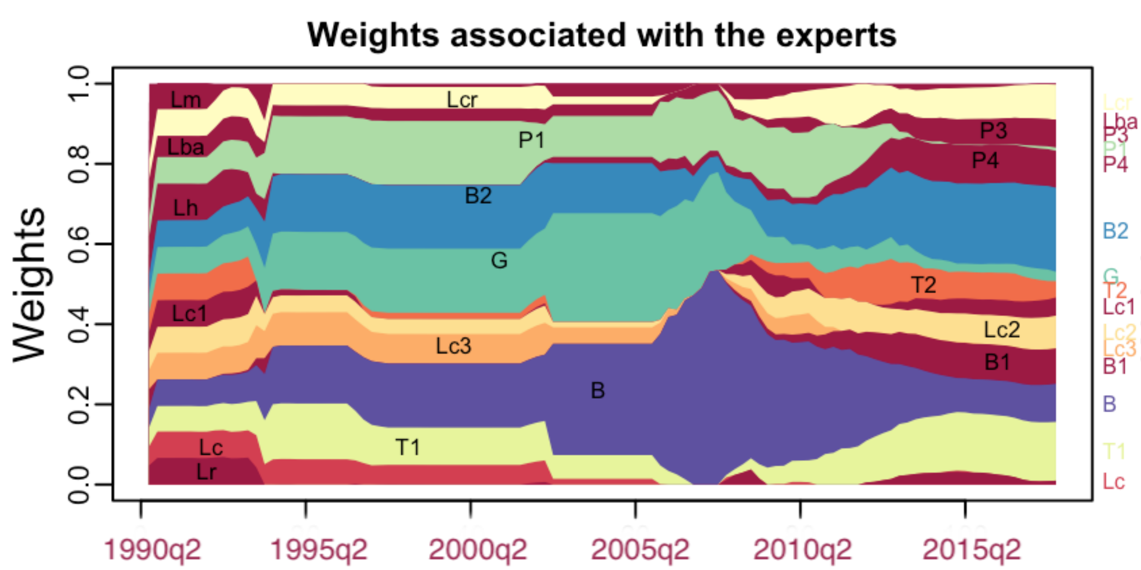

Figure 74: US: Weights. quasi-real time. OGD

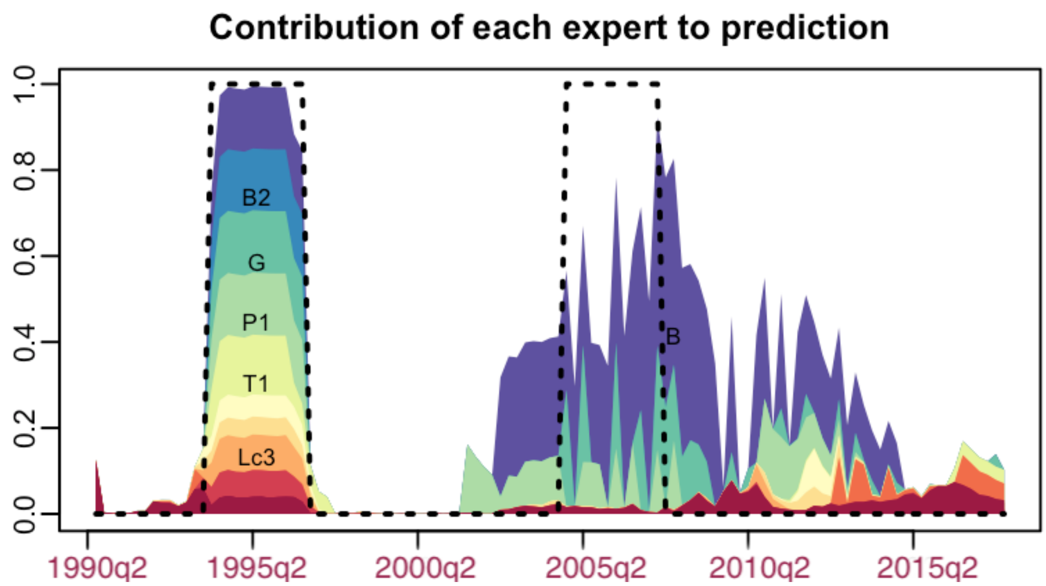

Figure 75: US: Experts contribution to forecast. quasi-real time. OGD

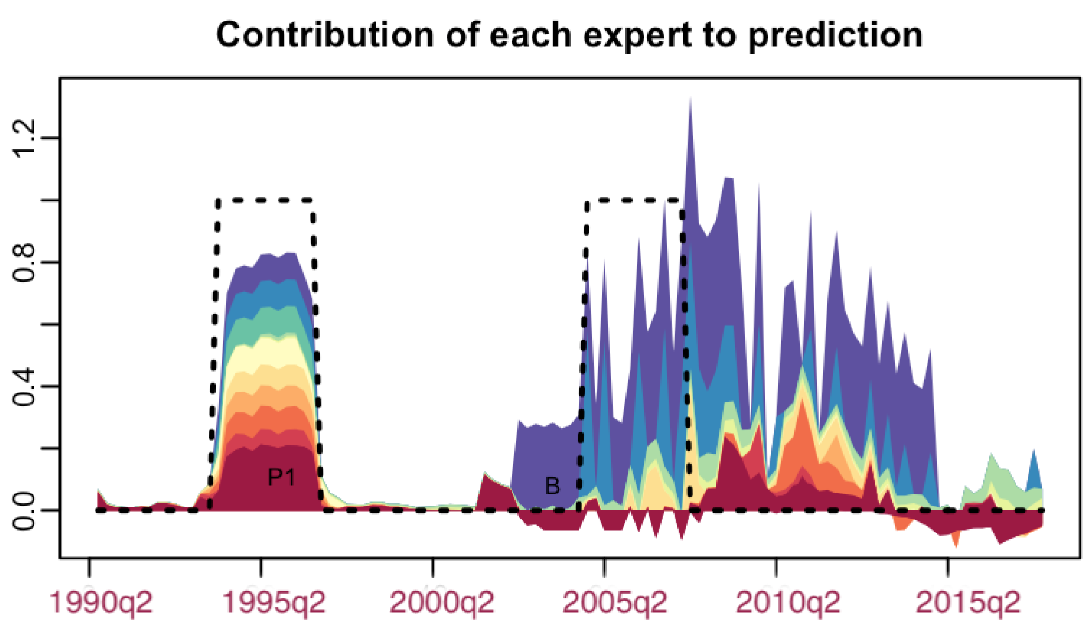

Figure 76: US: Weights. quasi-real time. Ridge

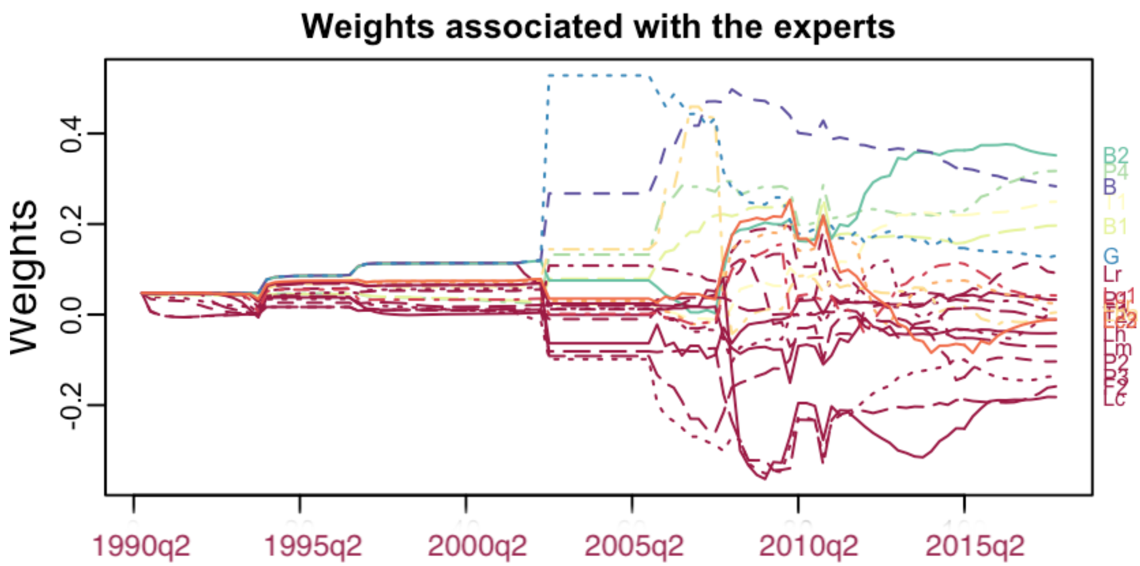

Figure 77: US: Experts contribution to forecast. quasi-real time. Ridge

J Results: France real time (vintage data)

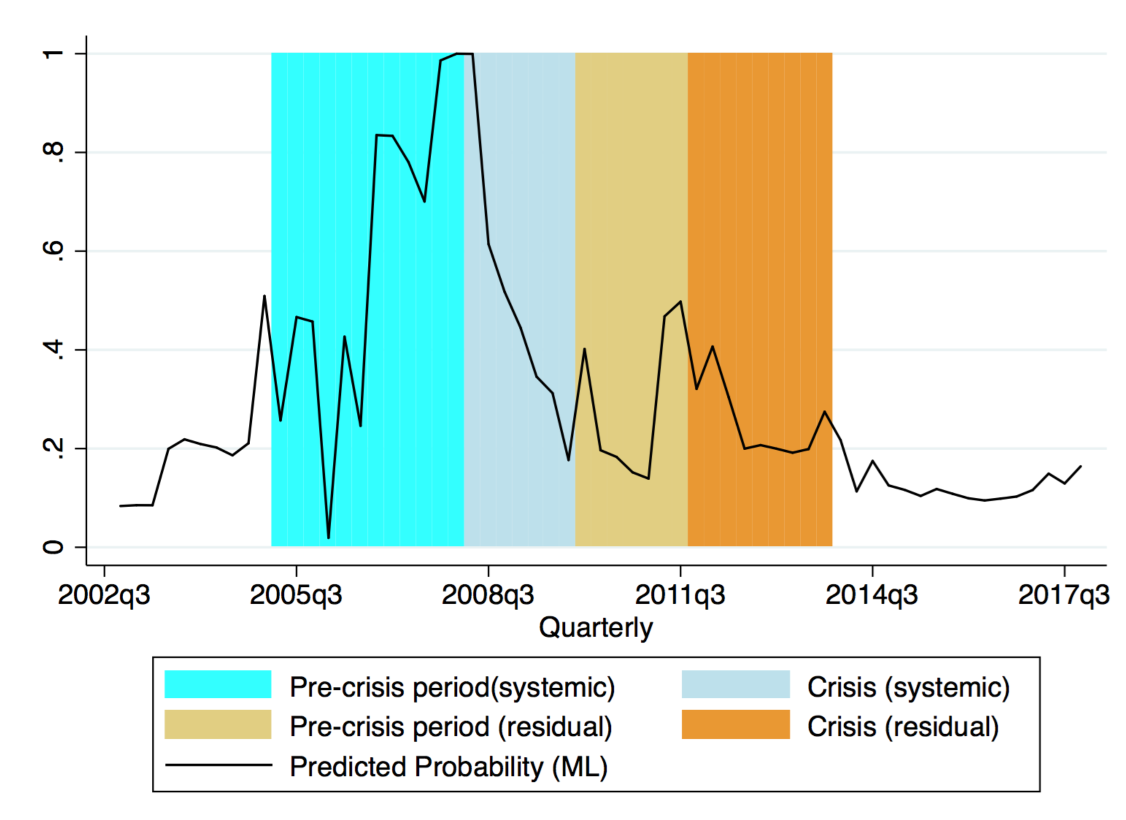

Figure 78: France. Predicted probability - ML (real time from 2002Q3 to 2017Q4)

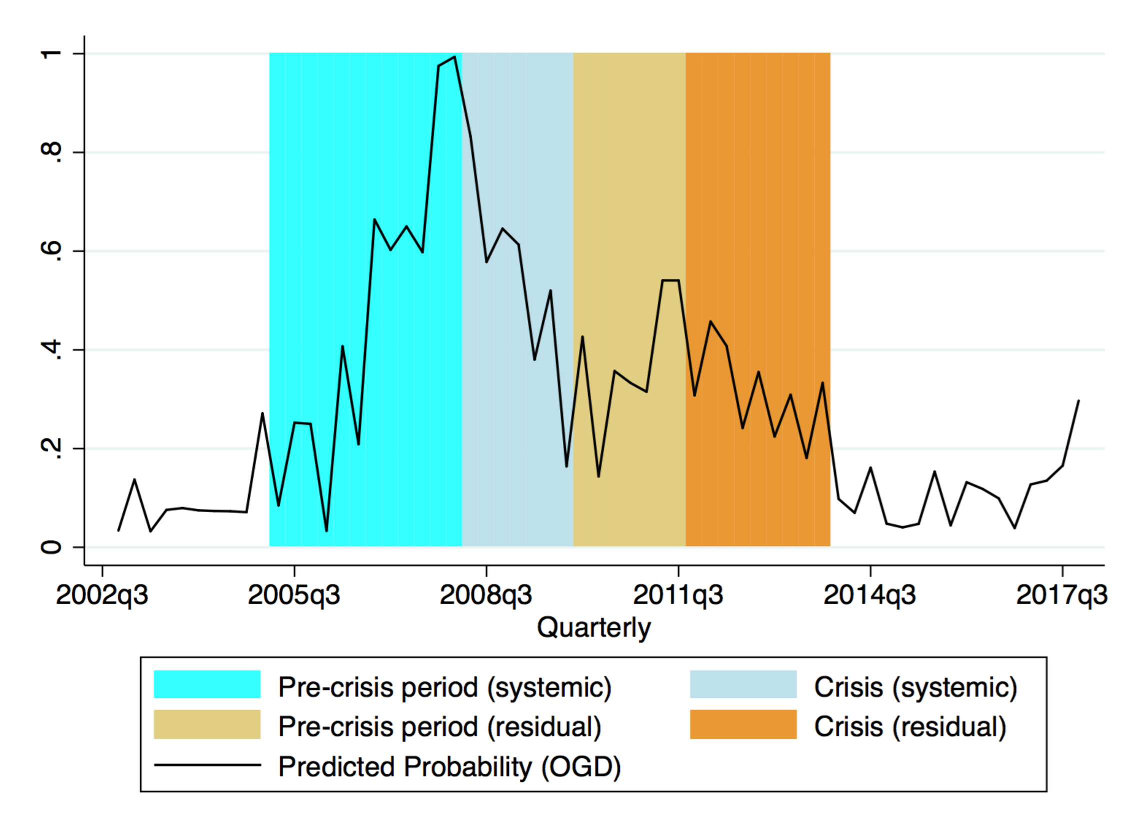

Figure 79: France. Predicted probability - OGD (real time from 2002Q3 to 2017Q4)

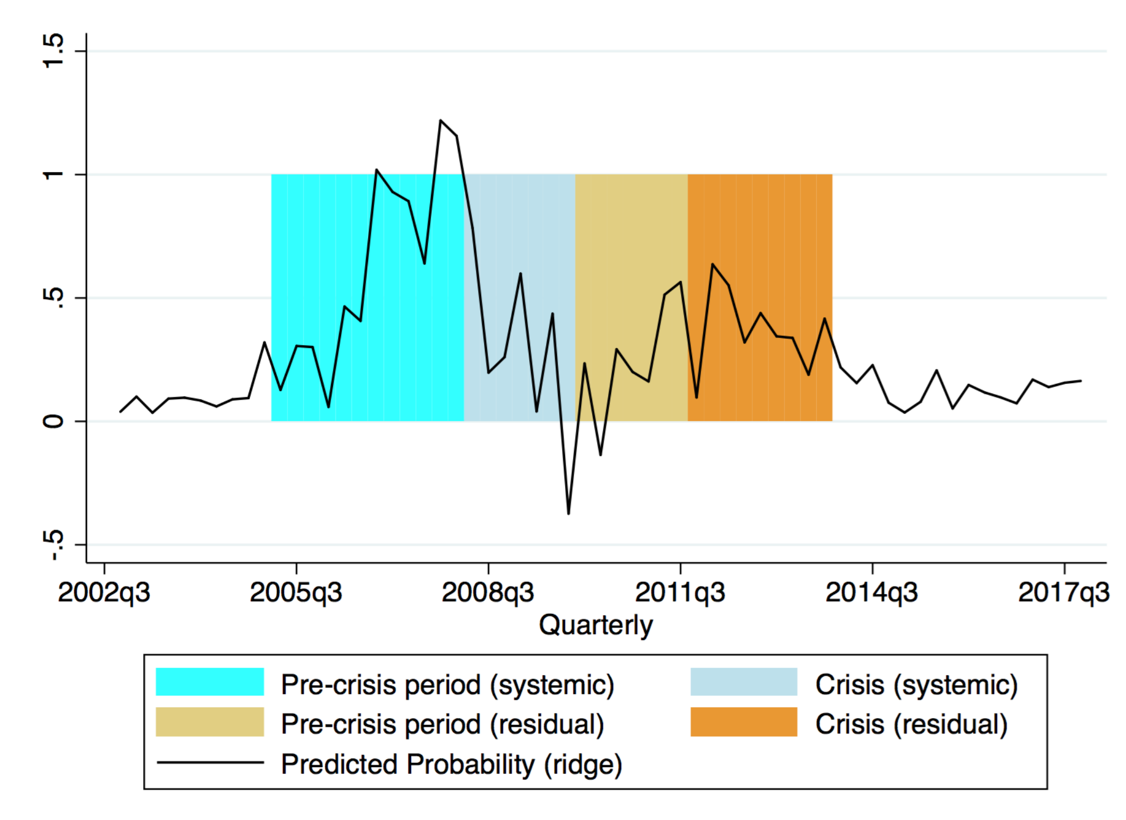

Figure 80: France. Predicted probability - Ridge (real time from 2002Q3 to 2017Q4)

Figure 81: France: Weights. real time. ML

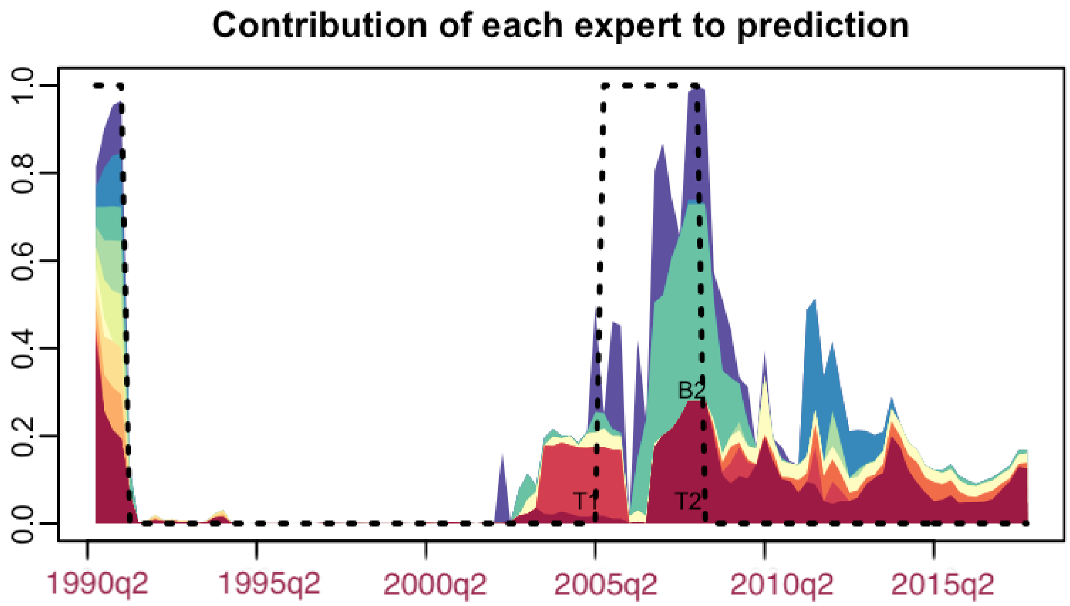

Figure 82: France: Experts contribution to forecast. real time. ML

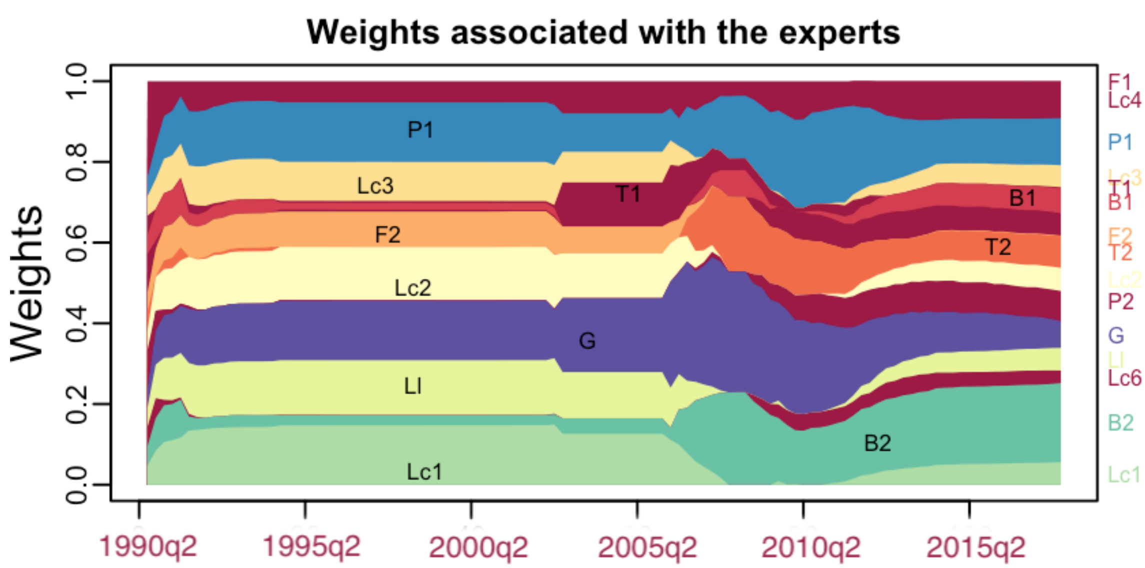

Figure 83: France: Weights. real time. OGD

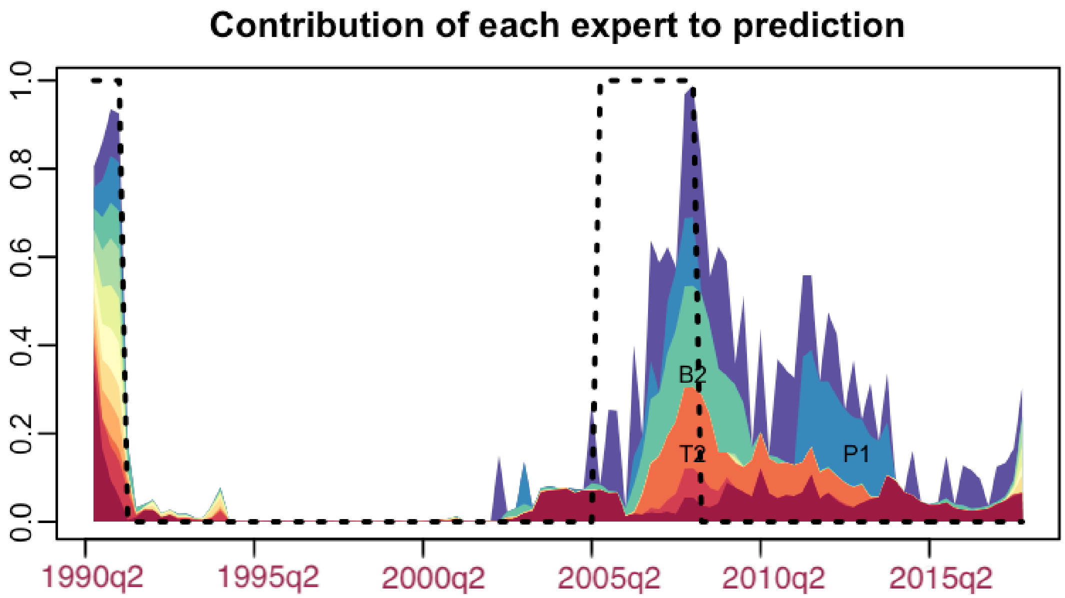

Figure 84: France: Experts contribution to forecast. real time. OGD

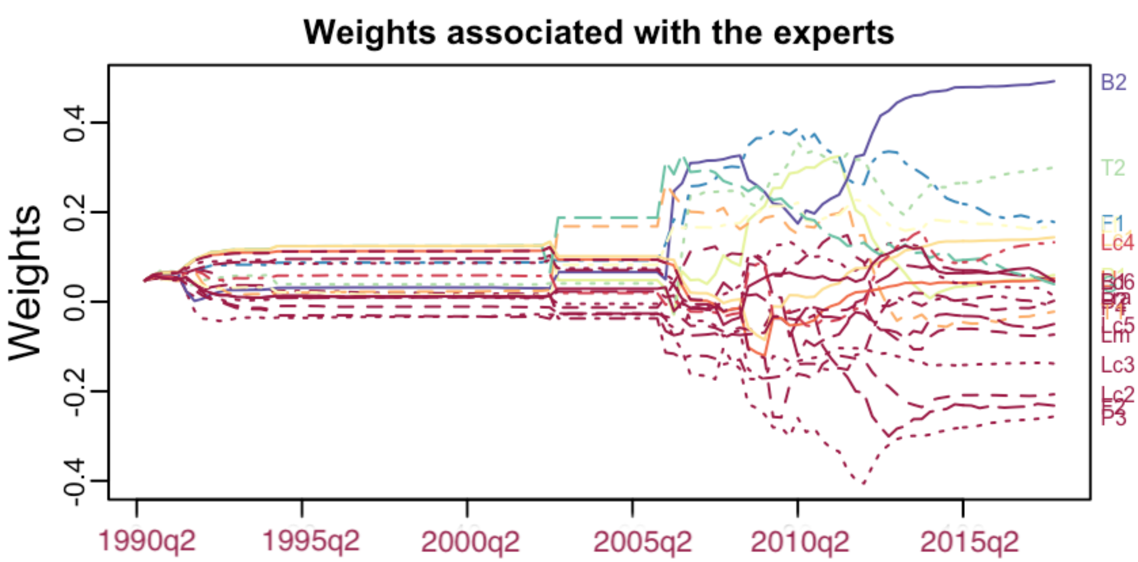

Figure 85: France: Weights. real time. Ridge

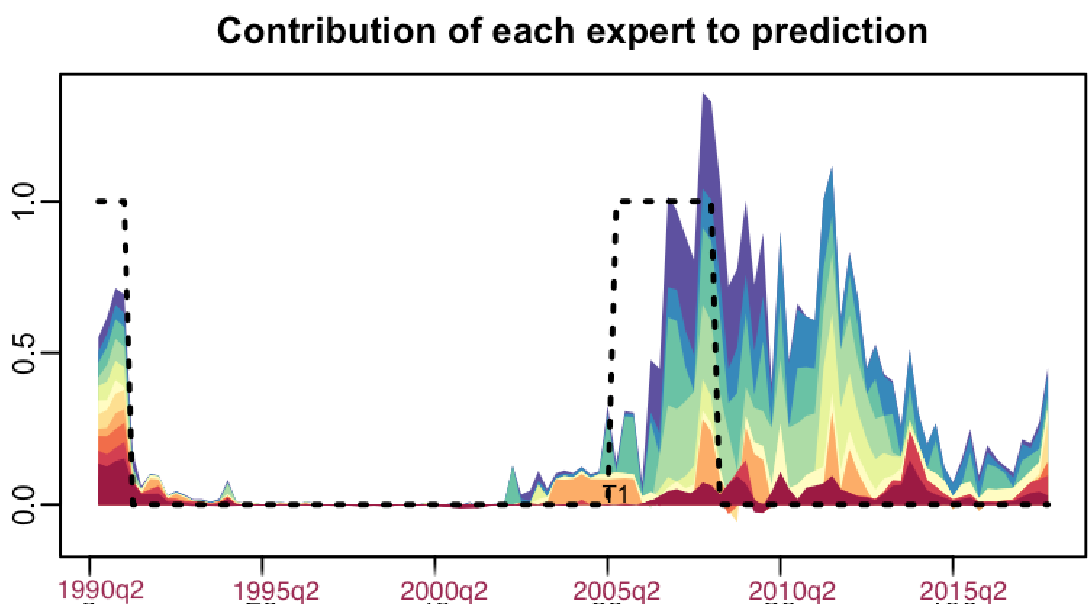

Figure 86: France: Experts contribution to forecast. real time. Ridge