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: June 26, 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.

[*] We are highly indebted to Pierre Alquier for his generous help. We thank Gilles Stolz for extremely useful comments and multiple discussions. We are also very grateful to our discussants Christian Julliard and Sebnem Kalemli Ozcan and to seminar participants at ENSAE, Nowcasting, the LBS AQR academic symposium, the BIS and the Women in Economics Chicago Conference for comments. Our thanks also go to Marco Lo Duca, to Katja Taipalus and to XXX for help with the datatabase. Jeremy Fouliard thanks the BIS for its hospitality. jfouliard@london.edu. 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

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 overfitting.

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. Online machine learning is specifically geared at quasi-real time prediction in situations where the true models driving outcomes are not known and are time varying. This approach can be described as ”meta-statistic” since the aim is to find the best sequential linear combination of experts. 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. 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 more general learning rates. 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) to illustrate the power of our methodology. We assess predictive ability using four model aggregation rules and we present AUROC results. In section 5 we build our own experts using millions of logit models to test whether an increase in the number of experts has significant effects on forecasting performance. 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 6 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. Our method can of course be applied to real time predictions but we lack the vintage series of our economic indicators to do the back testing in 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).

- **Macroeconomic indicators** : Current account, Consumer Price Index, GDP, M3, Unemployment rate, Cross-border capital flows, Total Liquidity Index.

- **Credit indicators** : Bank (or Total) credit, Household debt, Debt Service Ratios (household, non-financial corporations, non-financial sector), aggregate leverage, skewness of leverage.

- **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, Stock prices, Financial Conditions Index, Risk Appetite Index.

We draw from OECD's Main Econonomic indicators and National Accounts databases, the BIS and Cross Border Capital data (see more details in Appendix).

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 linear 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:

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

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 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

[1] The bound of the regret guarantees that forecasterfis 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 forecasterfis strategy includes time-varying weights. Forecasterfis 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.

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

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.

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-12th 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 $P_j = \arg\min_{p_j} d(p, y) = \arg\min_{p_j} || \sum_{j=1}^{N} p_{j,t} f_{j,t} - y_t||$

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

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 previous 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* :

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

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.

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 $ln(N)$ (or min $ln(\frac{1}{w_{j,0}})$ for non uniform initial weights), so that it is often better to use a large number of experts.

We will pick different sets of experts in section 4 and in section 5. In section 4 we pick "off-the shelf" experts used in the literature and in Central Banks to predict financial crises. The beauty of our approach is that we can include *any* type of experts and therefore be very oecumenical in terms of methodology. In section 5, we will build systematically a large number of experts using logit regressions on exhaustive combinations of variables.

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 add a series of models that we constructed ourselves (logits with elastic net penalties). 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 models we use in our forecasting exercise and refer the reader to the appendix for more details.

- X1: Dynamic probit model (panel estimates) - Bank of Portugal
- X2: Panel logit - fixed effect - PCA (panel estimates) -Bank of England

- X3: Panel logit - fixed effect -ECB (panel estimates)
- X4: Bayesian Random Coefficient Logit (country estimates) -Oesterreichiches Central Bank
- X5 : Binary Classification tree (panel estimates)
- X6 : Binary Classification tree (country estimates)
- X7 : Logit with elastic net penalty (housing) (country estimates)
- X8 : Logit with elastic net penalty (real economy)(country estimates)
- X9 : Logit with elastic net penalty (BIS variables) (country estimates)
- X10 : Logit with elastic net penalty (Bank) (country estimates)
- X11 : Logit with elastic net penalty (Monetary) (country estimates)
- X12 : Logit with elastic net penalty and bubble variable (country estimates) -Bank of Finland
- X13 : Logit with elastic net penalty (cross border variables) (country estimates)
- X14 : Logit Bashful: logit with 20 selected variables (based on AUROC) (country estimates)
- X15 : Logit Lazy: logit with every variables (country estimates)
- X16 : Logit with elastic net penalty (housing + real economy) (country estimates)
- X17 : Logit with elastic net penalty (housing + BIS) (country estimates)
- X18 : Logit with elastic net penalty (housing + BIS+ real economy) (country estimates)

We now have a total of 18 experts of all stripes and shapes including some models with common components and classification trees. 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, Spain,Italy, Spain, Sweden, the UK and the US. 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 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 (in yellow on Figure 2) from 2011 q1 till 2013 q4. We estimate the expert models on sample 1985Q1-2000Q2 a period during which France experienced the first systemic crisis, which was linked to real estate in par-

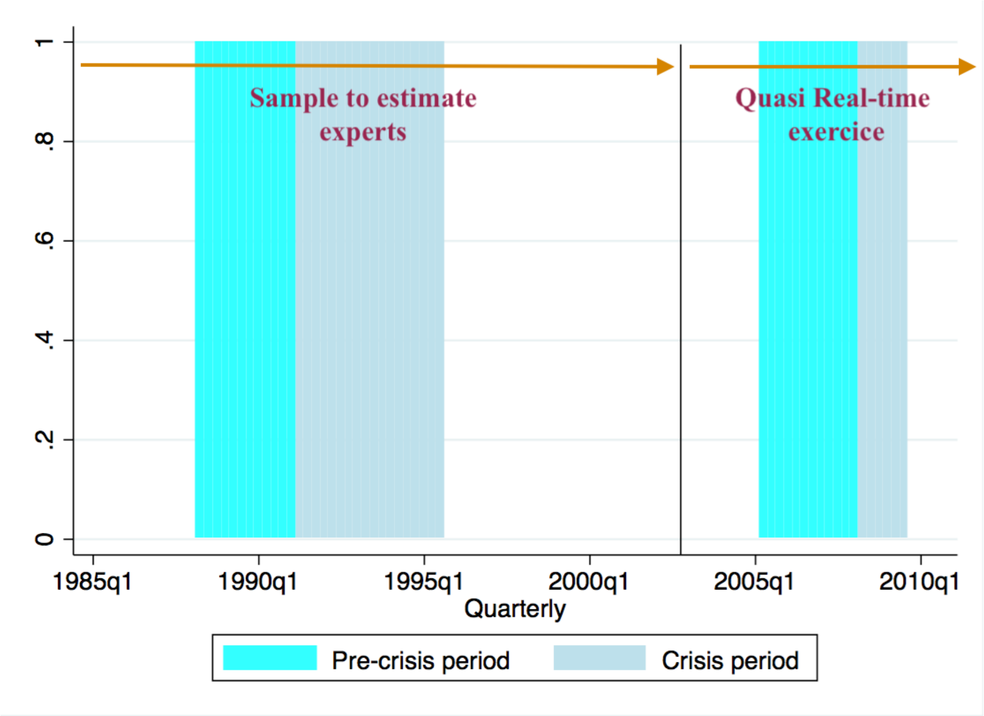

Figure 1: France

ticular. We present results for out-of-sample pre-crisis prediction for 2000Q3 to 2017Q2. This includes the period with the second systemic crisis following the collapse of Lehman Brothers as well as the euro area sovereign debt crisis, which is not classified as a systemic crisis in France.

Out-of-sample prediction of systemic crises.

Figure 2 presents the results for the EWA aggregation rule. It shows that the probability of being in a pre-crisis in 2003 and 2004 was low with a sharp increase starting in 2005q1. Since the probability increases over time, the model provides a very good early warning system. The 12 quarter ahead crisis probability keeps rising till 2006q4 where it reaches 1 and remains very elevated (between 0.8 and 1 till 2008q1. The model also performs very well as the crisis starts in 2008 Q1: the probability drops like a stone. There is then 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.

The other 3 aggregation rules (ML, OGD and Ridge) are presented in the appendix D. The results are remarkably consistent. One of the main difference across the different aggregation rules in terms of methodology is the way the learning rate is picked. For both the EWA and the

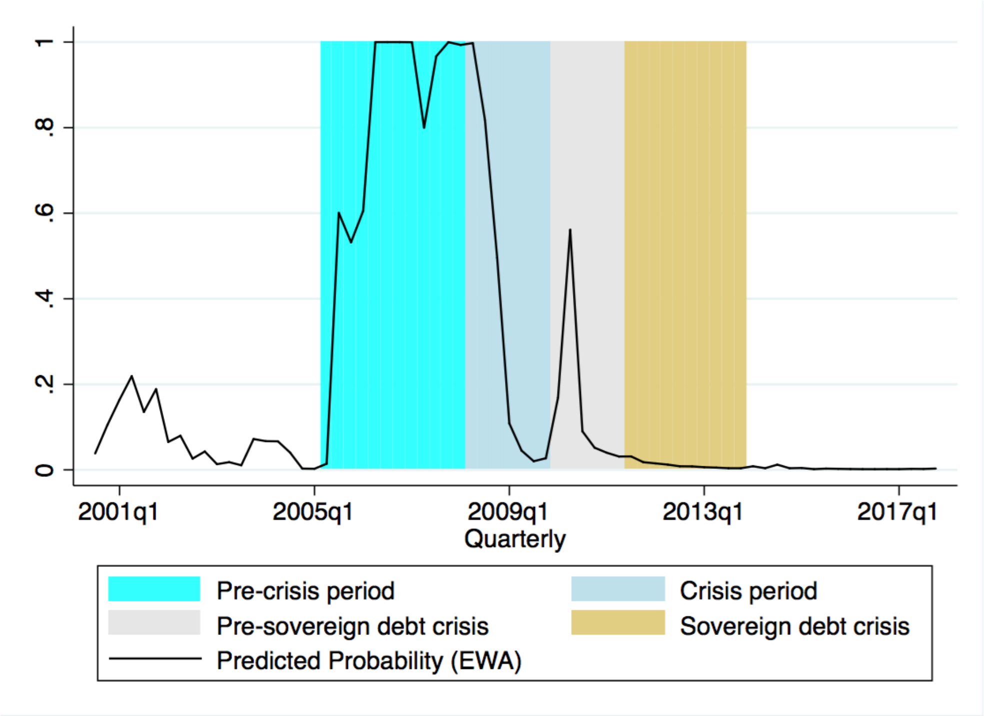

Figure 2: France: Predicted probability - EWA (quasi-real time from 2000Q3 to 2017Q2)

ML it is optimised upon whereas for the OGD and the Ridge the theoretically calibrated value of the learning rate is used. As a result we usually find very similar results for EWA and ML and also very similar results for OGD and Ridge (different speeds of learning). 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 very close to its theoretical asymptotic value of the best convex combination of experts (0.233 versus 0.212 for the best convex combination known ex post). EWA, ML, OGD, Ridge all do a lot better than uniform weights. The Ridge minimises RMSE and does better than the best convex combination but not as well of course as the best linear combination of experts (ex post).

Figure 3 shows the time varying weights associated to each of our 18 experts for the EWA aggregation rule. We see that there is some updating of weights when information is revealed and that the optimal forecast for each of our rules puts some positive weights on most of our models while towards the end of the sample a small number of models tend to dominate. Towards the

Online Aggregation Rule	RMSE
EWA	0.233
Uniform	0.351
ML	0.236
OGD	0.282
Ridge	0.208
Best convex combination	0.212
Best linear combination	0.117

Table 1: RMSE of different aggregation rules and expert. France: quasi-real time from 2003Q3 to 2009Q3

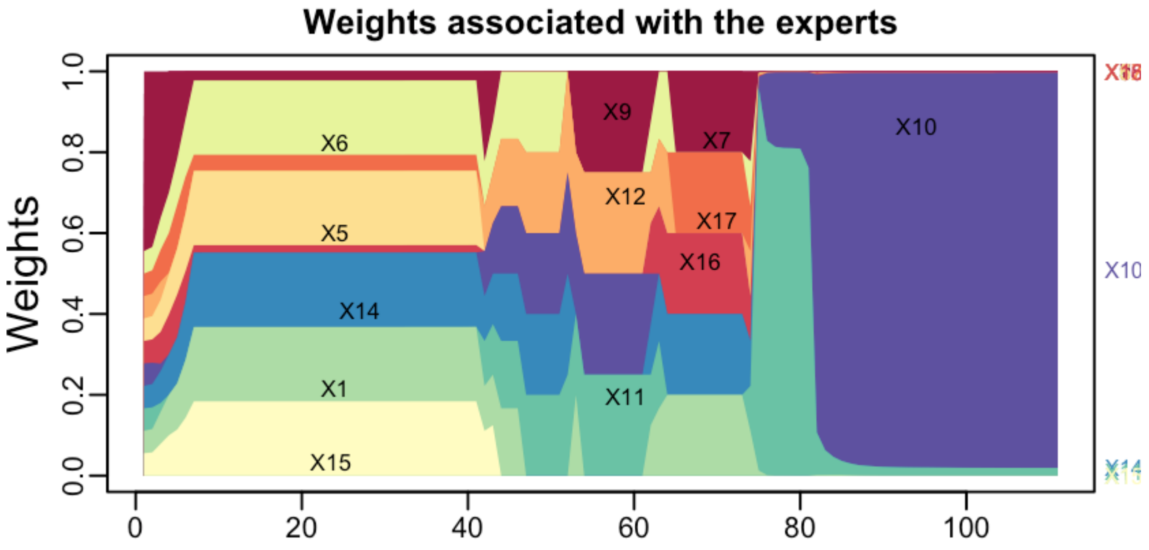

Figure 3: France: Weights. EWA

beginning of the sample X1, X5, X6, X14 and X15 dominates[3]. One common point to all these models is that they contain housing variables (price-to-rent, and price-to-income in particular). Hence it seems that development in the housing sector in France are closely related to financial instability towards the beginning of the sample. As time passes and information is revealed however weights are changing and X9, X10 , X11 and X12 become the relevant models[4]. Hence we seem to be moving to a build up of risk reflected by credit, monetary and banking variables. But this evolves further as X7, X14, X16 , X17 [5] and X1 become dominant. Those models reflect both credit and real estate. Finally X 11 (monetary becomes almost entirely dominant and at the end of the sample one model has almost a weight of 1: X10 (banks).

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

Figures 4 shows for the EWA aggregator the distribution of weights associated to each experts: the black bar is the median weight over the prediction period and the rectangle spans the first quantile to the third quantile. Although there are some variations across aggregation rules (see Appendix D), the same subset of models tends to attract higher weights. These are for France X1, X14, X6, X16. [6]. Figure 5 shows the average loss of each expert (which is different from the cumulated loss) over the forecasting period. Interestingly, the expert X16 is systematically the one with the lowest average loss (see Appendix D) while experts X3 [7] and X5 (panel classification tree) come systemically last. Both of these experts are estimated on a panel (and not country by country). We note that the average loss of the EWA rule is below the loss of each expert as well as below the loss of an uniform aggregation rule.

[3] They correspond to: X1: Dynamic probit model (panel estimates) with 4 variables selected with AUROC (Total Credit to private non-financial sector (2y change); Price-to-rent; Price-to-income; Consumer prices (2y change). X5: binary classification tree (panel). X6 : Classification tree (country estimates). X14: Logit bashful; X15: Logit lazy

[4] X9: logit with BIS variables; X10: logit with bank and risk appetite variables; X11 logit with monetary variables and X12 ; logit with asset bubble variable)

[5] X17: housing and BIS; X16: housing and real economy; X7: housing

[6] X16 : Logit with elastic net penalty (housing + real economy) (country estimates)(Price-to-rent ; Price to income; Price to rent 1y change; Price to income 1y change; Real estate price ; Real estate price 1y and 2y change; GDP (nominal 1y change 2y change); GDP (Per capita per person per hour); Multifactor productivity ; Oil price ; Oil price 1y change 2y change).

[7] X3: Panel logit - fixed effect: Banking credit gap-to-trend; Banking credit 1-y change; Consumer prices; Share price index fi?? 1y change; Real estate price 1-y change.

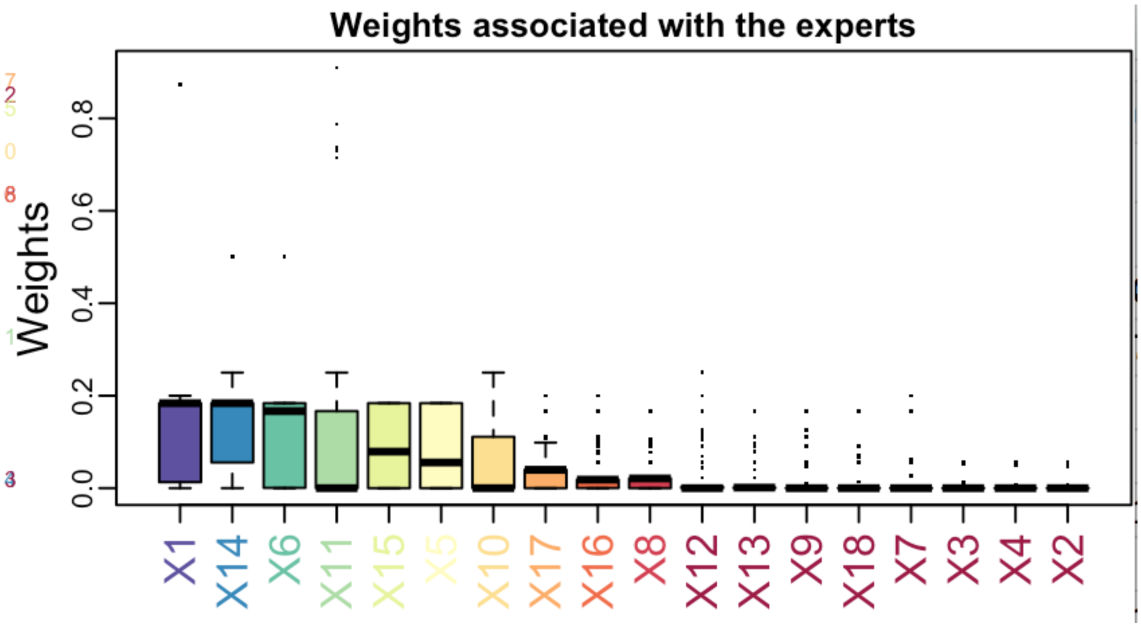

Figure 4: France: Distribution of Weights. quasi-real time from 2003Q3 to 2009Q3. EWA

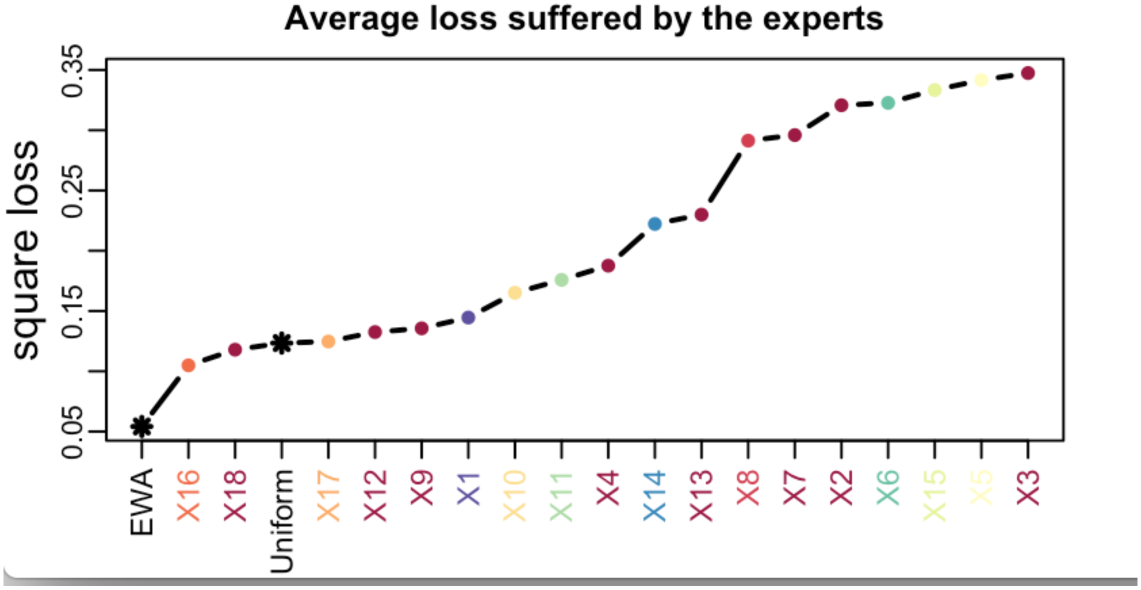

Figure 5: France: Average loss. quasi-real time from 2003Q3 to 2009Q3. EWA

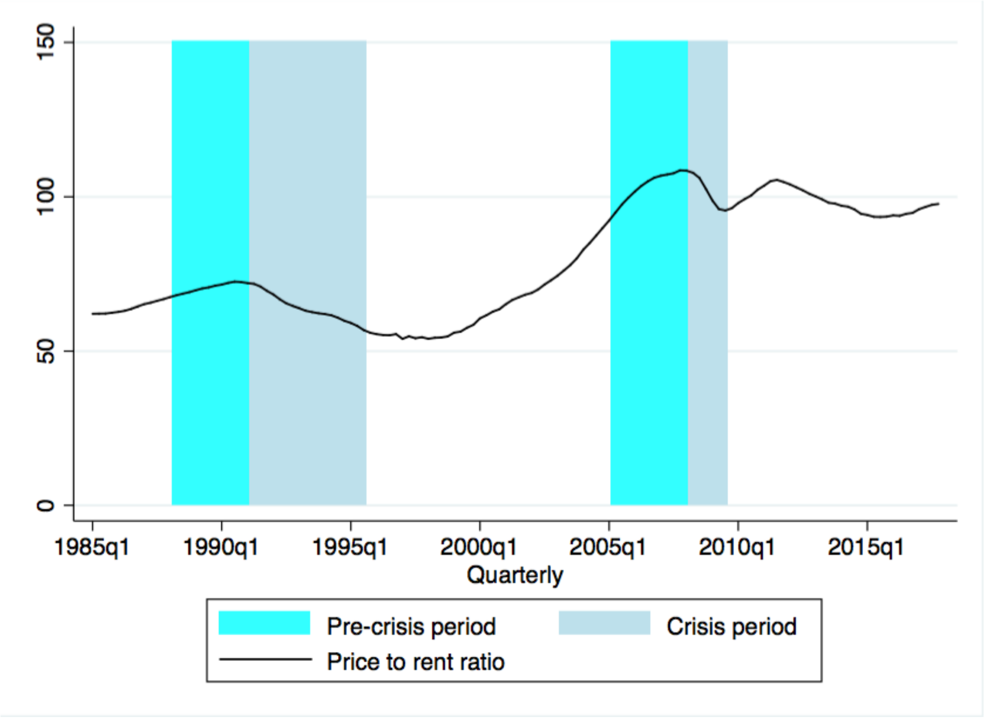

Figure 6: Price to rent - France

As an illustration, we plot in Figure 6 one of the key variables coming up in many of the most frequently selected models for France: the price to rent ratio. Indeed and even though it needs not be the case its time variation follows closely the pre- crisis dates.

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. Figures 7 plot the ROC curves for France for the EWA aggregation rule (see Appendix D for the other 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 ranges from 0.938 for the ML aggregation rule to 1 for the ridge (and 0.988 for the EWA). Furthermore the time profile of the signal is such that it would have been very valuable for policy makers.

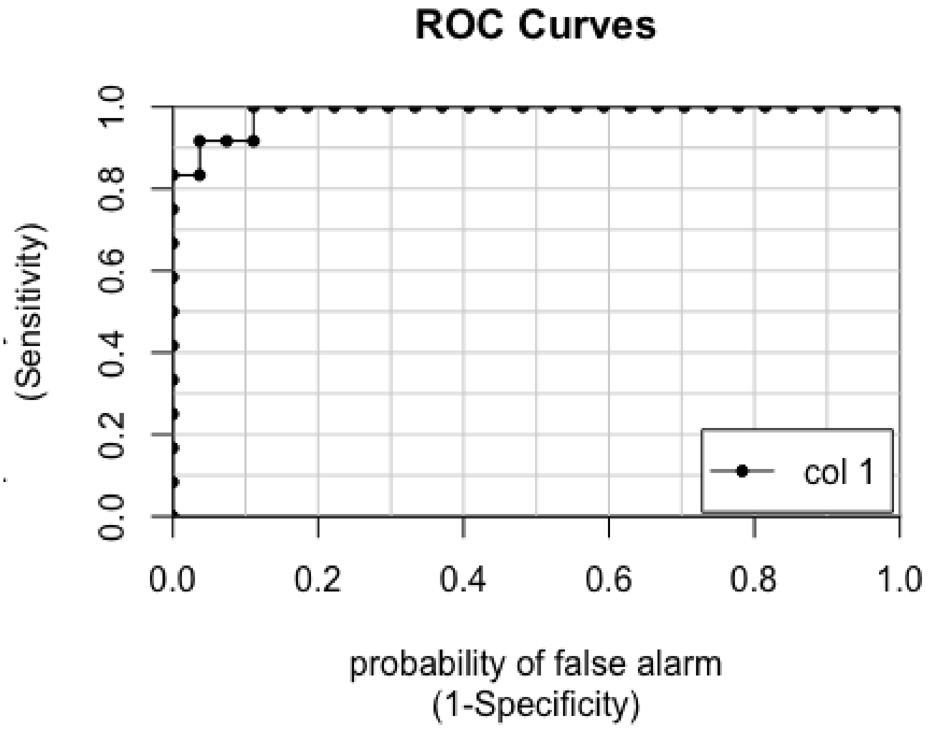

Figure 7: France: AUROC=0.988. EWA

5.2 Germany

We now turn to Germany. We estimate the expert models on sample 1985Q1-2001Q4. 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 8 presents the timing of pre-crises in Germany in light blue. The systemic crises are is in grey. We present results for out-of-sample prediction for 2002Q1 to 2010Q1.

Out-of-sample prediction of systemic crises.

Figure 9 (see also Appendix D) presents the predicted probability for the EWA aggregation rule. The consistency across aggregation rules is, once more, remarkable. The probability of being in a pre-crisis in 2003 was already non zero (10 to 20%) but there was a sharp increase starting in 2004q1. The probability kept increasing till about 60% in 2005 and stabilised at that high level until the crisis started. Just like for France, the model shows a remarkable ability to pick up the turning

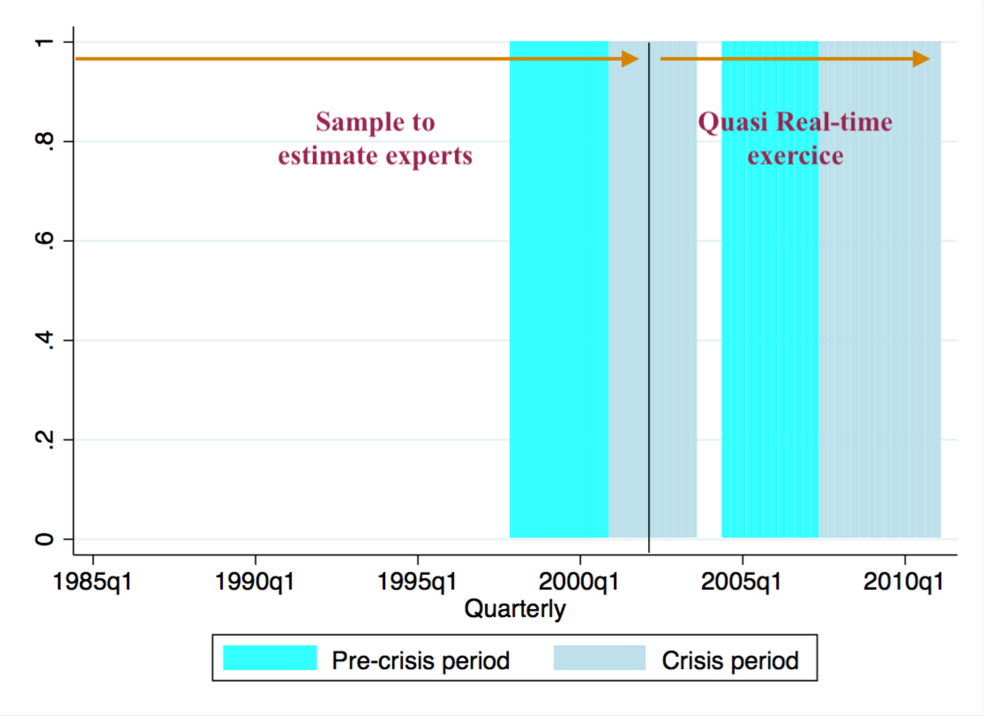

Figure 8: Germany

points. The model also performs very well as the crisis starts: the probability drops quickly.

Table 2 presents the RMSE of our different aggregation rules. We note that the EWA and the ML RMSE are very close to zero (0.0738 and 0.00658) and so is the best convex combination of experts (0.00168 known ex post). EWA, ML, OGD, Ridge all do a lot better than uniform weights. We are therefore in an interesting case where some experts seem to have very small forecasting errors. As we study the weights of our aggregation rule (Figure 10 and Appendix D) we note that a small subset of experts carry a lot of weight at the end of the sample. The distribution of weights reflect the importance of 3 experts for Germany and the average loss tells the same story. There is therefore strong dominance of 3 experts in German data. These are X16 (housing and the real economy), X17 (housing and BIS) and X18 (housing, BIS and the real economy). Just like for France, housing and credit variables seem therefore to contain key information about the likelihood of a crisis.

Figure 13 plots the ROC curves for Germany. We have the remarkable result of a ROC curve of 1 for EWA and OGD and of 0.929 for the Ridge. This shows the power of our approach: when the forecasting performance of some experts is high, our aggregation rules minimizing regret has

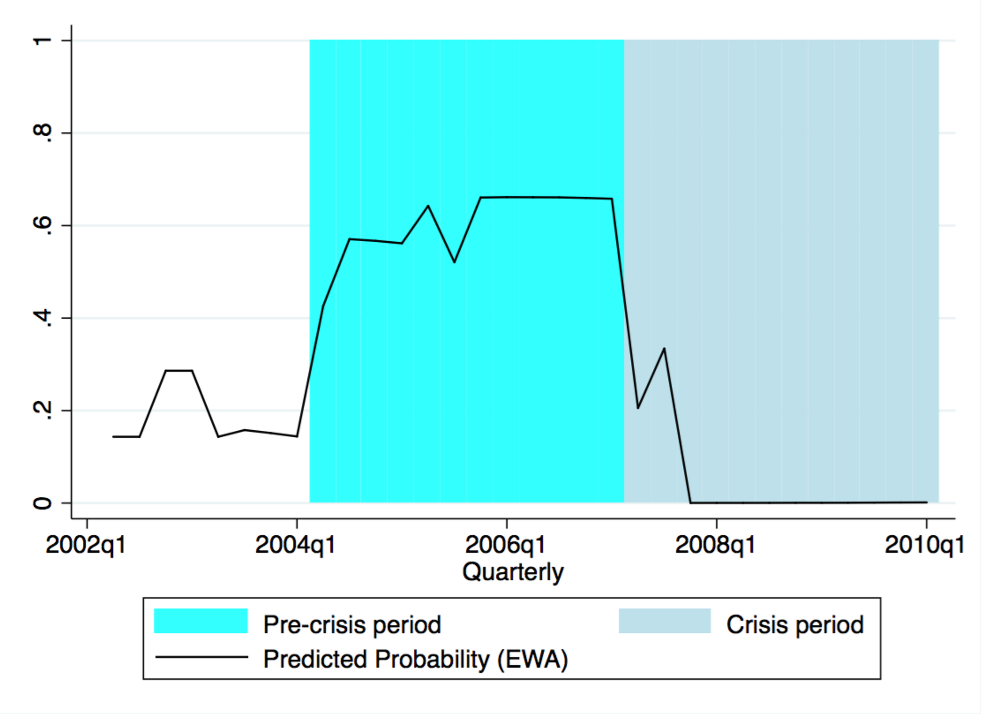

Figure 9: Germany: Predicted probability - EWA (quasi-real time from 2002Q1 to 2010Q1)

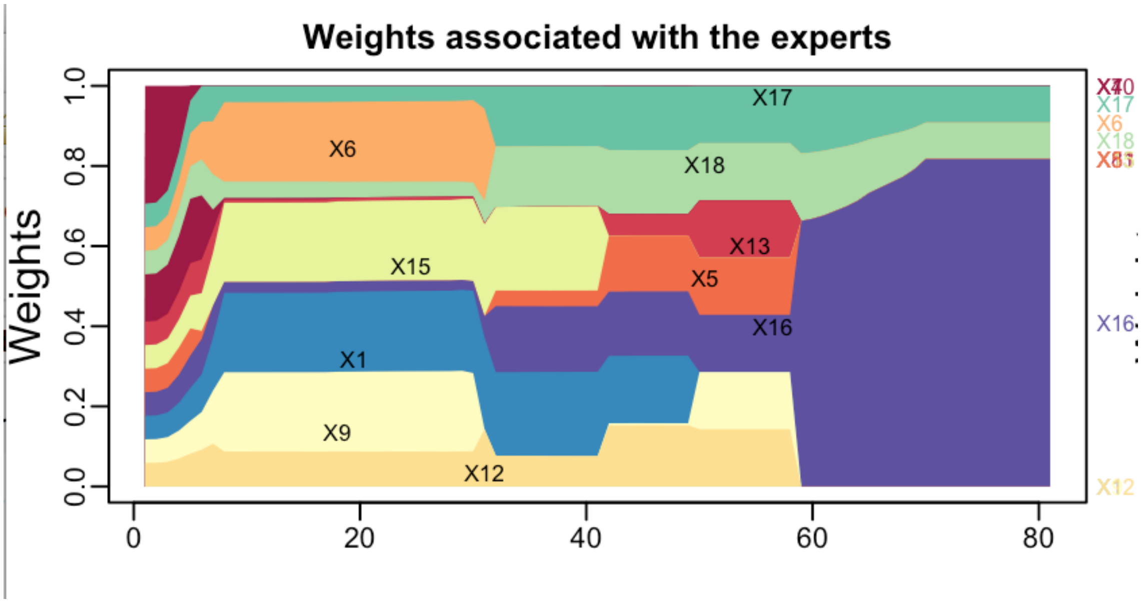

Figure 10: Germany: Weights. quasi-real time from 2002Q1 to 2010Q1. EWA

Online Aggregation Rule	RMSE
EWA	0.0738
Uniform	0.288
ML	0.0658
OGD	0.106
Ridge	0.101
Best convex combination	0.00168
Best linear combination	0.0003

Table 2: RMSE of different aggregation rules and expert. Germany: quasi-real time from 2002Q1 to 2010Q1

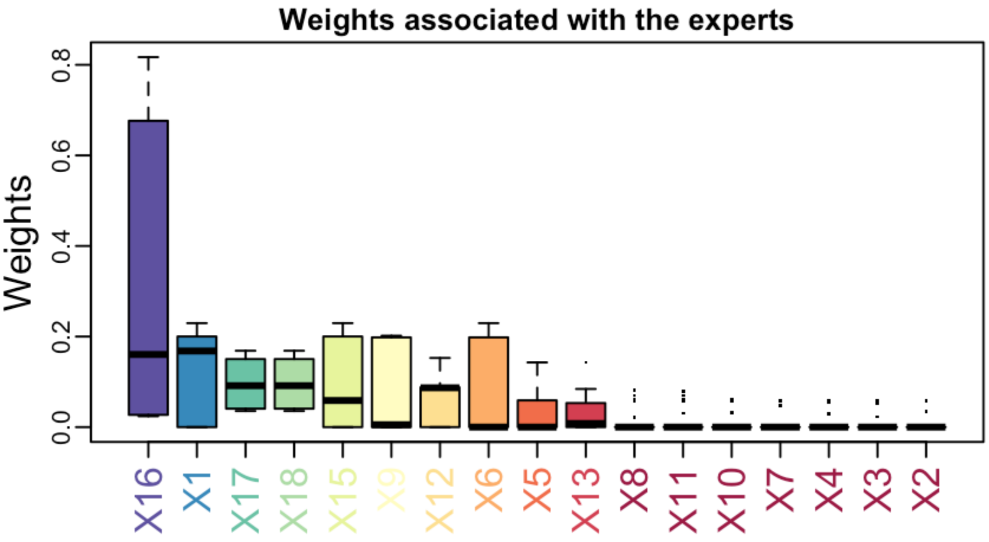

Figure 11: Germany: Distribution of Weights. quasi-real time from 2002Q1 to 2010Q1. EWA

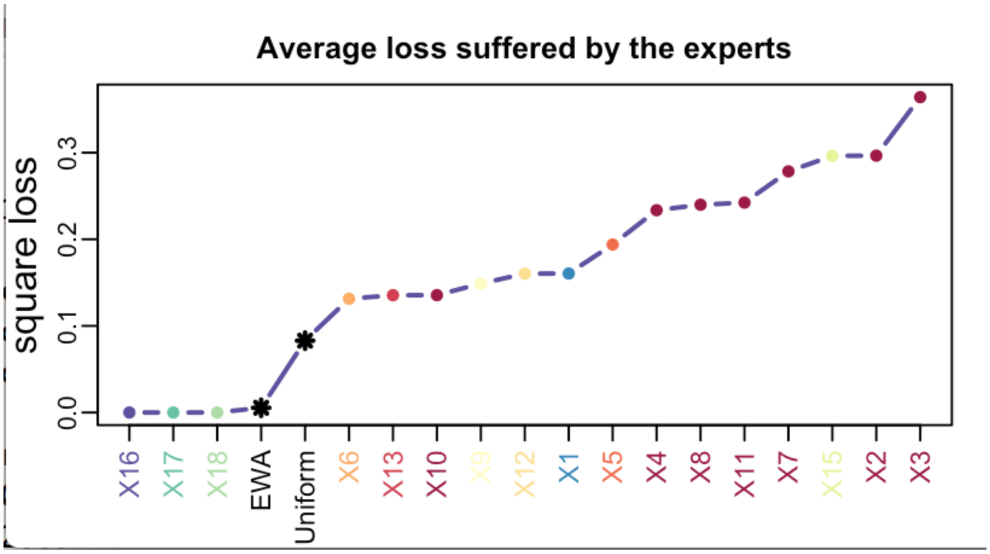

Figure 12: Germany: Average loss. quasi-real time from 2002Q1 to 2010Q1. EWA

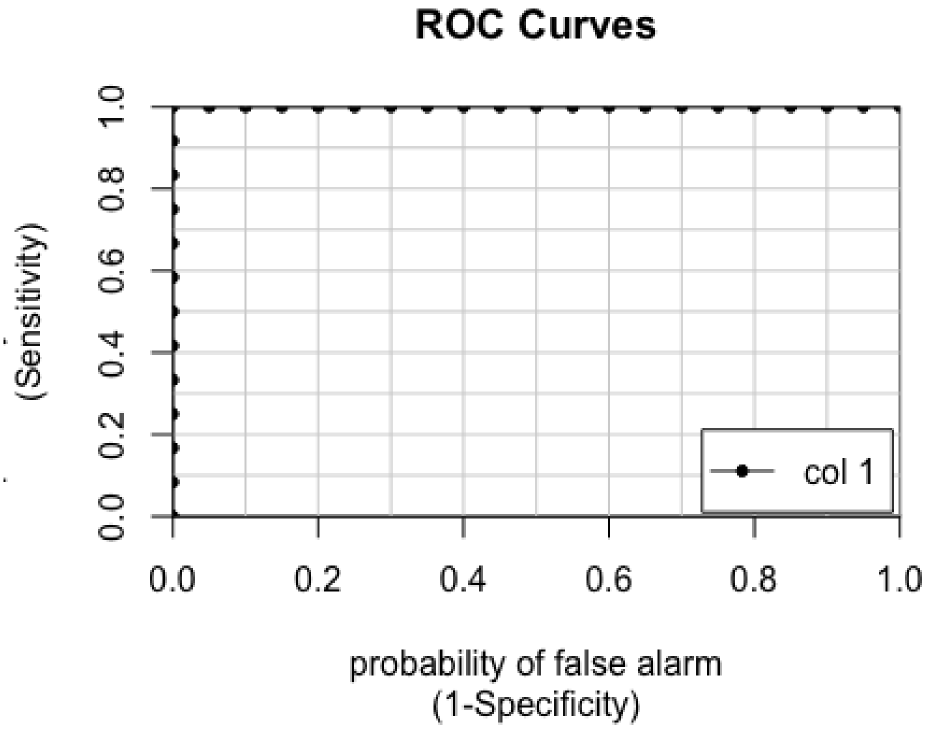

Figure 13: Germany: AUROC=1. quasi-real time from 2002Q1 to 2010Q1. EWA

a very high degree of reliability.

5.3 Italy

For Germany and France, the timing of the systemic crisis is closely related to the collapse of Lehman Brothers (2008q1 for France, 2007q2 for Germany). For Italy however the first systemic crisis occurred in 1991 Q3 and finished in 1997 Q4 but the most recent crisis started in 2011Q2 (and ended in 2013 Q4). This is the euro area crisis. We note that all aggregation rules (including the best convex combination ex post) pick up the increase in pre-crisis probability in 2008 q2 but also all show a decline in the probability of being in pre crisis in 2009 q2 before rebounding. This shows that none of our experts is able to give a stable signal that particular quarter[8] . Interestingly for Italy X1 is very dominant towards the end of the sample while X6 and X14 are important at the beginning. This combination of model reflects risk both in credit and in the housing market.

[8] The Aquila earthquake happened in 2009 q2

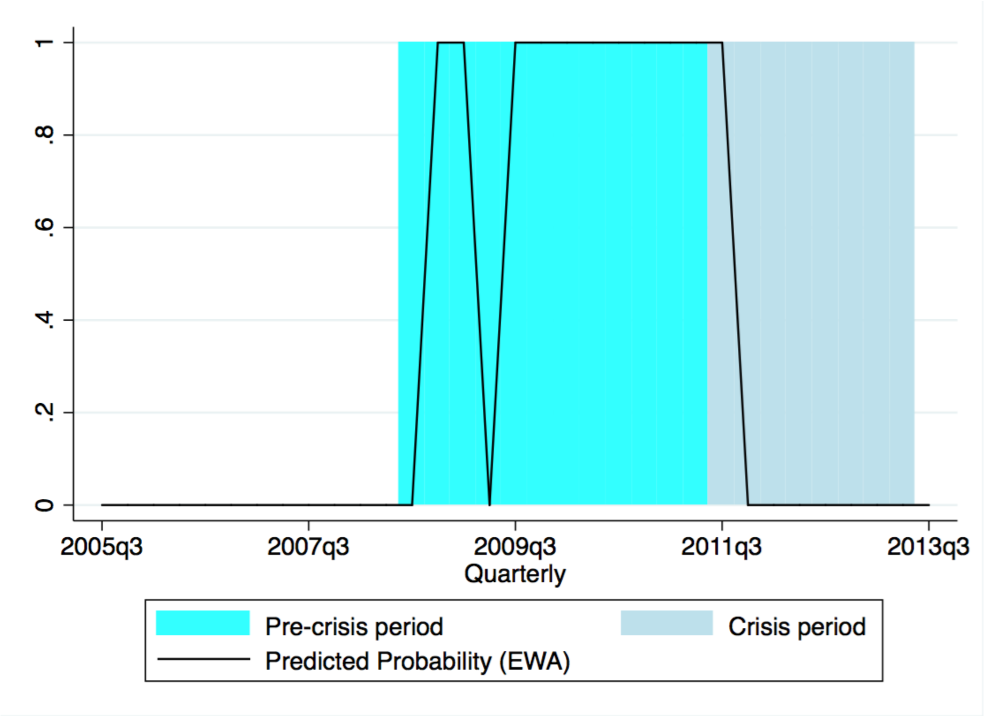

Figure 14: Italy. Predicted probability - EWA (quasi-real time from 2005Q3 to 2013Q3)

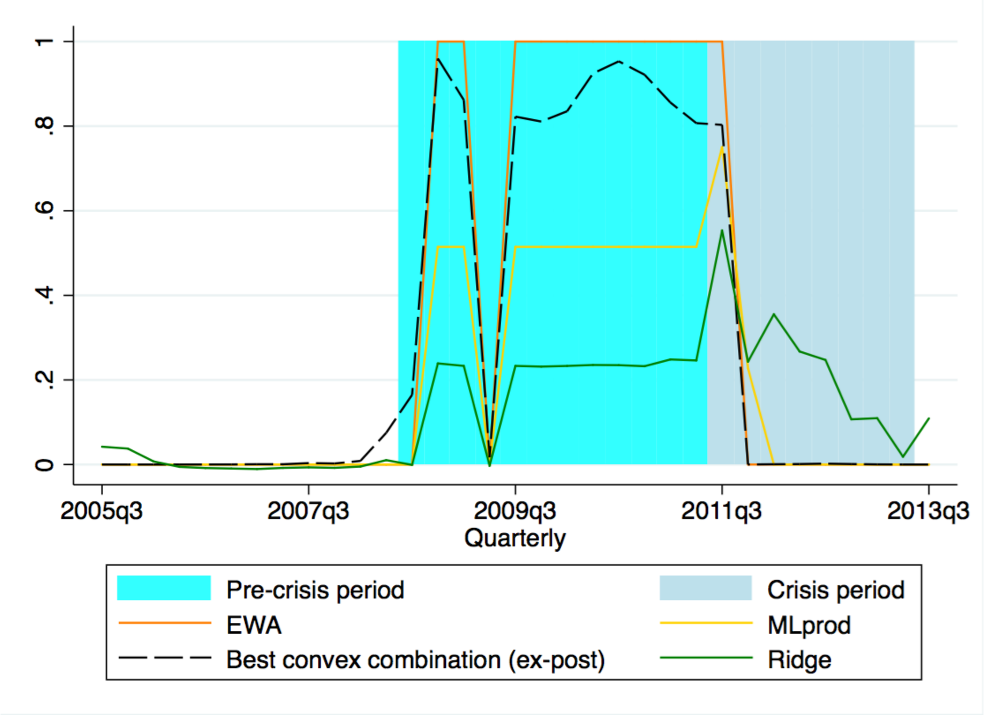

Figure 15: Italy. Different probability estimates.

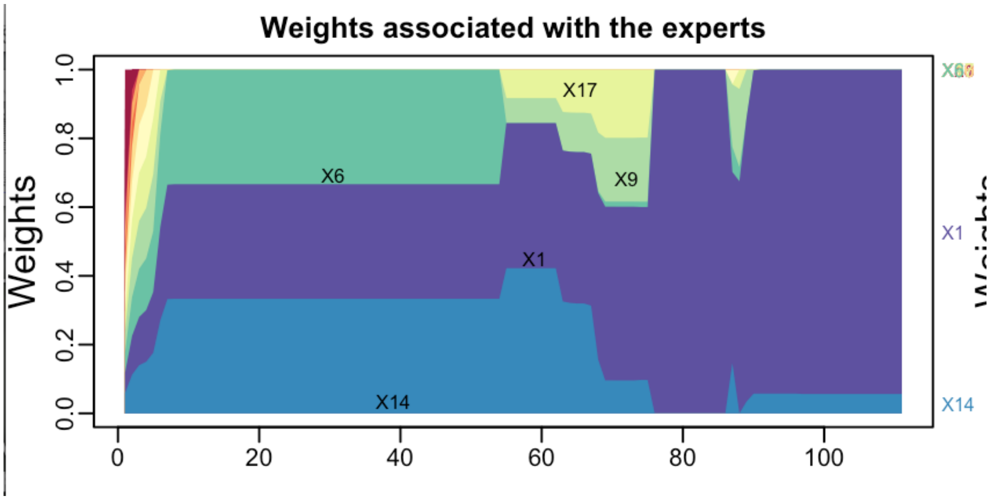

Figure 16: Italy. Weights- EWA

5.4 Spain

For Spain, the crisis started in 2009 Q1 and ended in 2013q4. Interestingly even before the collapse of Lehman Brothers we can pick up a sudden increase in pre crisis probability (in 2006 Q1). This probability reaches one 2007 q4 and then drops with the beginning of the crisis in 2009 q1. One model is very dominant for Spain at the end of the sample: it is the logit (elastic net penalty) with housing, BIS and real economy variables. In appendix D we present the other aggregation rules. The ML is very similar to the EWA. The OGD and the Ridge have slightly different properties. Interestingly the OGD puts some weights on the banks variables but X18 remains dominant.

5.5 Spain

5.6 Sweden

In Sweden, the crisis starts in 2008 q3 and ends in 2010 q3. Sweden went through a systemic real estate crisis previosuly (1991 q1 to 1994 q2). Again our aggregation rule picks up the turning point in the increased probability of pre crisis in 2005 q3 and goes to a probability of 1 in 2007q3.

In terms of the dominant models, the pattern is now familiar. Towards the beginning of the sample, the more ”general models” with an important real estate component tends to dominate. In an intermediate period the model with banks and risk appetite (X10) appears and towards the

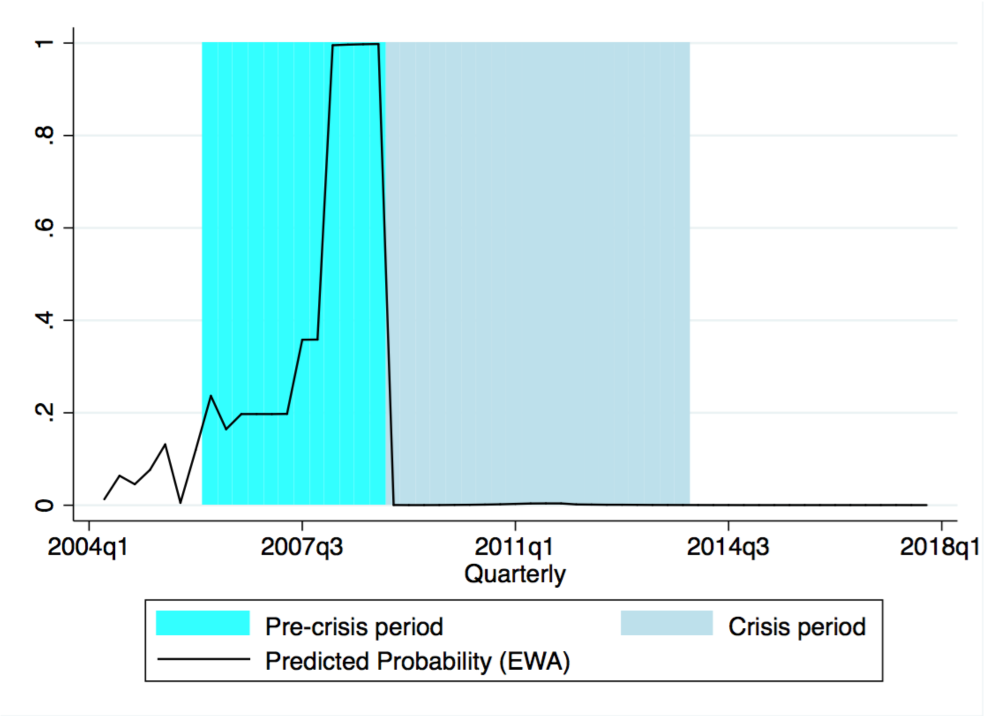

Figure 17: Spain

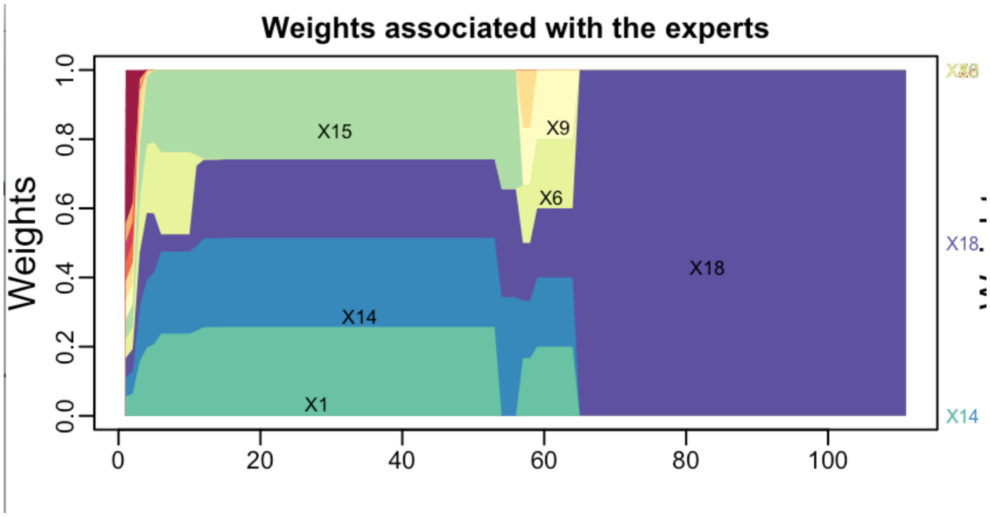

Figure 18: Spain

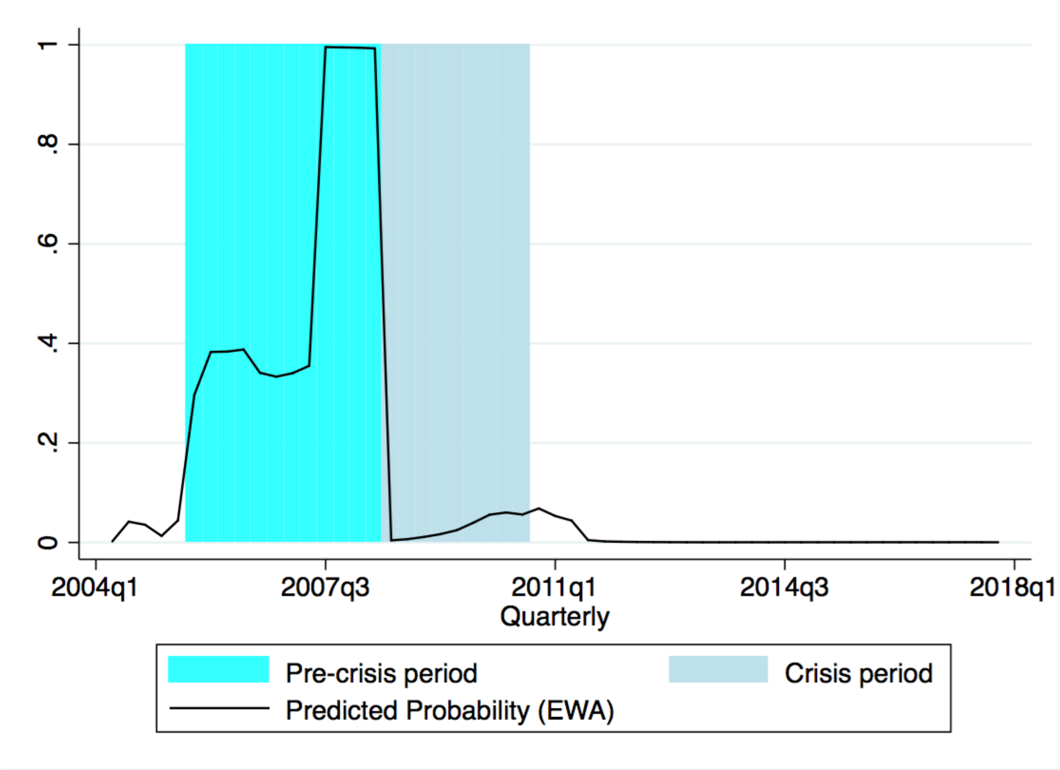

Figure 19: Sweden

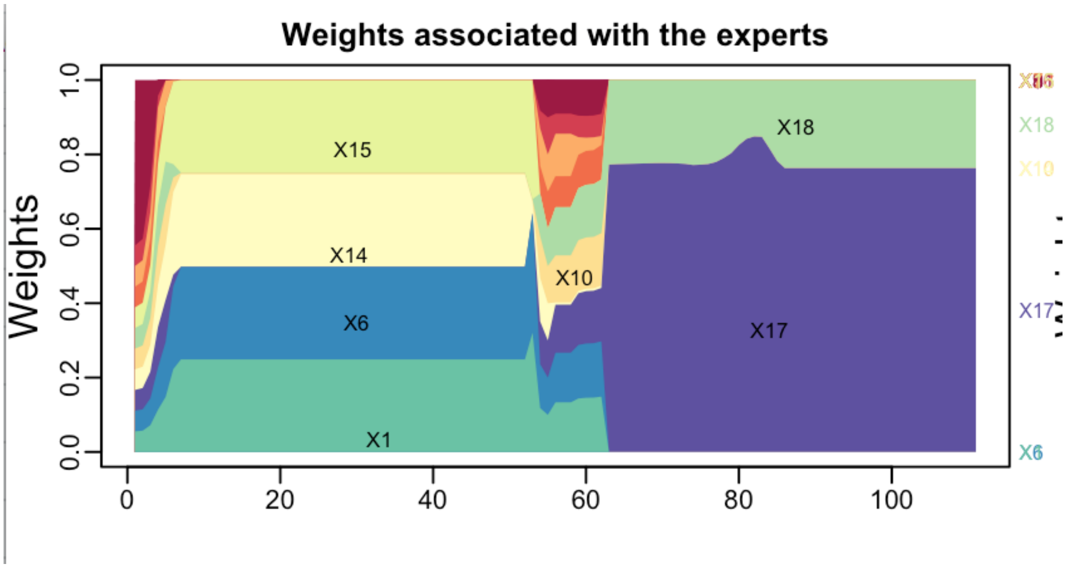

Figure 20: Sweden: weights

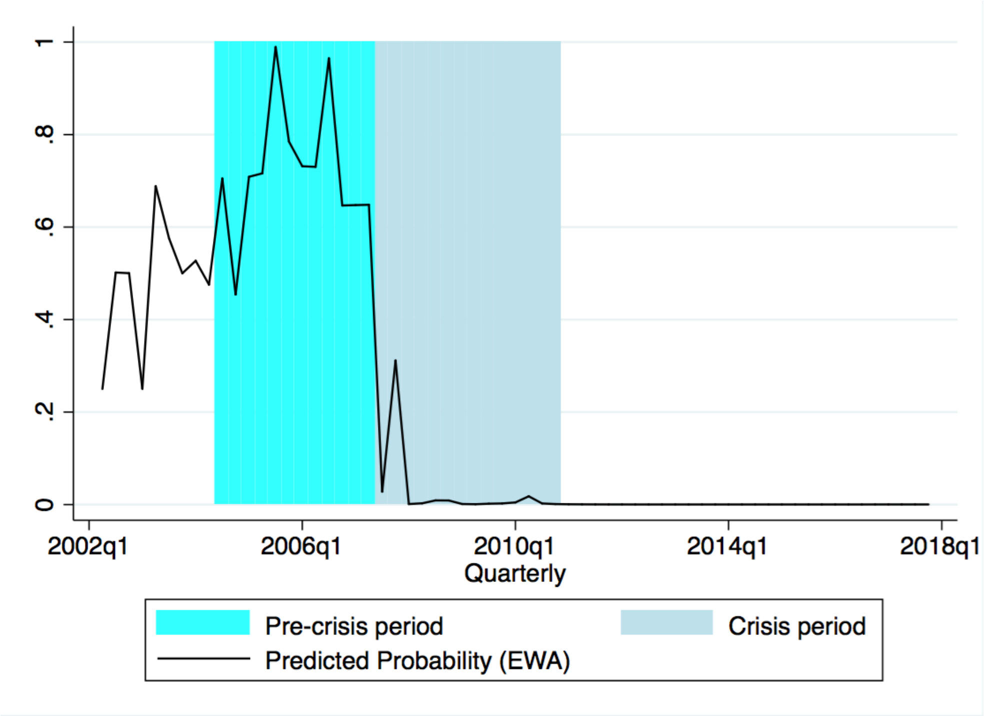

Figure 21: UK

end the dominant models are the elastic net models with housing the real economy and the BIS variables.

5.7 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 probability profile for the UK is much more jagged. It is possible that we are missing some important experts. Nevertheless, the probability increases very significantly in 2006 q1 and remains elevated until the beginning of the crisis. The experts most responsible for the prediction towards the end of the sample is the logit with the housing and the BIS variables.

6 Millions of logit experts

Increasing the number of experts is likely to give us even better results. We do this in a systematic way by computing millions of logit experts. Let's define Ω as the set of Early Warning Indicators such that $\Omega = (X_1, X_2, ..., X_n)$ with $card(\Omega) = n$. $P(\Omega) = \{\emptyset, \{X_1\}, \{X_1, X_2\} ...\}$ is the whole

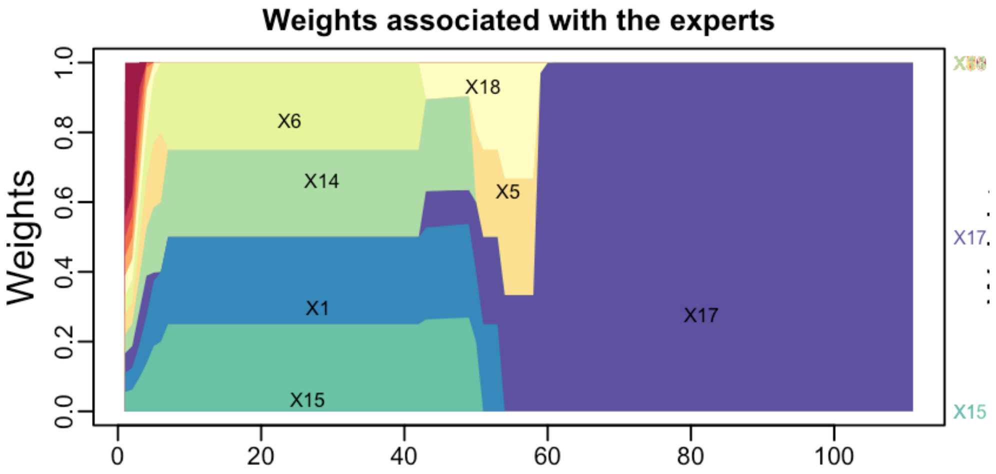

Figure 22: UK

group of subsets of Ω. Since we do not know the best variables to predict pre-crisis periods, we estimate the basic logit equation for each non-empty element of $P(\Omega)$:

$$I_t = F(\alpha + \sum_{k=1}^{K_0} \beta_k X_{k,n,t-1}) \tag{4}$$

where F is a logistic function such as $F(Z) = \frac{\beta_k e^Z}{1+e^Z}$, K_0 is the number of variables to be included in the regression, I_t the pre-crisis indicator α and β_k parameters to estimate.

At each time instance $t-1$, each expert predicts the probability of pre-crisis period for the period t. Indeed the fitted value of the logit estimation can be interpreted as the estimated conditional probability of pre-crisis period :

$$\widehat{f}_{j,t} = Pr[I_t = 1|X_k] = F(\widehat{\alpha} + \sum_{k=1}^{K_0} \widehat{\beta}_k X_{k,t-1}) \tag{5}$$

Note that the lag on $X_{k,t-1}$ accounts for the delay for the delay in the avaibility of data for the forecaster.

We use all the possible combination of variables. Due to computational power we have to restrain the number of variables. We select indicators which have an AUROC greater than 0.74 on a restricted sample (reflecting out-of-sample time analysis). Limit similar to medical science limit. Gives us around 20-25 variables which corresponds to 1.5 to 3 millions of experts for each country. Everything is estimated country by country.

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 18 experts, some of them being central bank financial crises models, we find that, for France, the best performing experts are several and tend to be related to housing variables and the real economy, but also feature some credit variables. For Germany a trio of experts dominates and generates very low forecast errors. These experts reflect also housing and real economy variables. We get similar results for Spain, Italy, Sweden and the UK. Clearly it is very important to allow for time varying weights. Real estate variables and risk appetite and banking variables are important at different times. Models estimated country by country (and not on a panel) tend to generate lower average losses. 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. The out-of-sample forecasting ability of our EWA expert is outstanding with AUROCs close to 1 in several instances. Our method is very flexible: we could incorporate many more experts (deep learning, subjective judgement) and potentially increase further the performance of our model. In a companion paper we use our methodology of online learning on historical data (Jorda Schularick and Taylor dataset) and are able to predict the Great Recession in quasi real time. The set of variables important for the great depression seem to differ from the set of variables needed to predict 2008 in the US. An obvious other application of our methodology is to predict recessions, something we will tackle in future work.

Appendix

A Data

B 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 $\ell_t \in [0, 1]^K$,

$$\sum_{t=1}^{T} \hat{\ell}_t \leq \min_{1 \leq j \leq N} \left\{ \sum_{t=1}^{T} \ell_{j,t} + \sqrt{N(1 + ln(1 + T))(1 + \sum_{t=1}^{T} (\hat{\ell}_t - \ell_{j,t})^2} \right\} \tag{6}$$

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 $\eta_t > 0$ *Initialization : an uniform vector* p_1*. 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* $\hat{y}_t \in [0, 1]$:

$$R(\mathcal{R}_\eta) \leq \inf_{v \in \mathbb{R}^N} \left\{ \lambda ||v_2^2|| \right\} + N \times ln(1 + \frac{T}{\lambda N}) \tag{7}$$

C Experts

- X1 (France): Bank of Portugal : Dynamic probit model. (panel) 4 variables selected with AUROC : - Total Credit to private non-financial sector (2y change) - Price fi??to-rent - Price-to-income - Consumer prices (2-y change)
- X1 (Germany) Bank of Portugal : Dynamic probit model. (panel) 4 variables selected with AUROC : - Total credit to non-financial corporations (2y change) - Total Credit to private non-financial sector (1y change) - GDP$_G 2 - Real Estate Prices - 2y change$

- X2 (France): Bank of England : Panel logit - fixed effect - PCA: 4 variables selected with PCA: - GDP Per Hour Index - GDP Per Person Index - M3 - Price-to-rent
- X2 (Germany): Bank of England : Panel logit - fixed effect - PCA: 4 variables selected with PCA : - Price to rent - Total Credit to private non-financial sector - gap to trend - GDP Per hour index - M3

- X3: ECB : Panel logit - fixed effect: - Banking credit fi?? gap-to-trend - Banking credit 1-y change - Consumer prices - Share price index ?? 1y change - Real estate price 1-y change

- X4: Austrian central bank : Bayesian Random Coefficient Logit : (panel) - Share price Index fi?? 1y change - GDP fi?? 1y change - Banking credit 1y change - Real estate price fi?? 1y change - Total Liquidity Index fi?? 1y change - Financial Condition Index 1-y change

- X5 : Binary Classification tree (panel)

- X6 : Binary Classification tree (country)

- X7 : Logit with elastic net penalty (housing): (country) - Price-to-rent - Ptice to income - Price to rent 1y change - Price to income 1y change - Real estate price - Real estate price 1y and 2y change

- X8 : Logit with elastic net penalty (real economy): (country) - GDP (millions dollar 1y change 2 y change) - GDP (Per capita per person per hour) - Multifactor productivity - Oil price - Oil price 1y change 2 y change

- X9 : Logit with elastic net penalty (BIS variables): (country) Every credit variable % GDP 1y change 2y change gap to trend

- X10 : Logit with elastic net penalty (Bank): - Risk Appetite - Financial Condition Index - Share price Index - Equity holding - Liquid Assets

- X11 : Logit with elastic net penalty (Monetary): (country) - M3 - Short term interest rate (nominal) - Short term interest rate (real) - Consumer prices

- X12 : Bank of Finland: Logit with elastic net penalty and bubble variable (country)

- X13 : Logit with elastic net penalty (cross border variables) (country)

- X14 : Logit Bashful : logit with 20 variables selected by AUROC (country)

- X15 : Logit Lazy: Logit with every variables (country)

- X16 : Logit with elastic net (housing + real economy) (country)

- X17 : Logit with elastic net penalty (housing + BIS) (country)

- X18 : Logit with elastic net penalty (housing + BIS+ real economy) (country)

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 $\alpha = 0.7$.

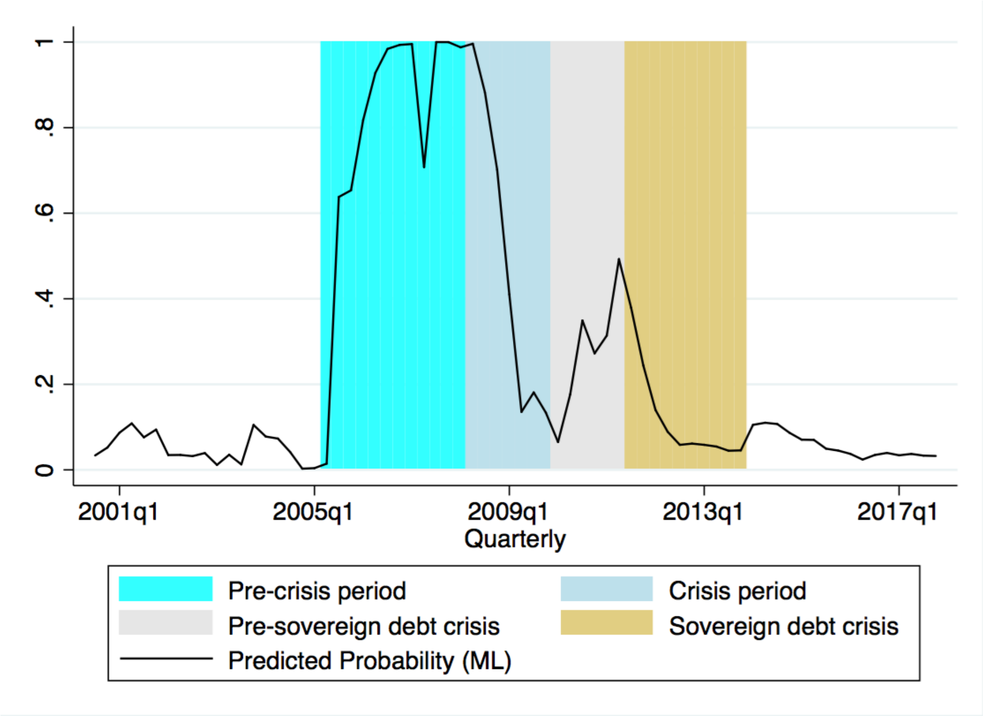

Figure 23: France: Predicted probability - PML (quasi-real time)

D Results France

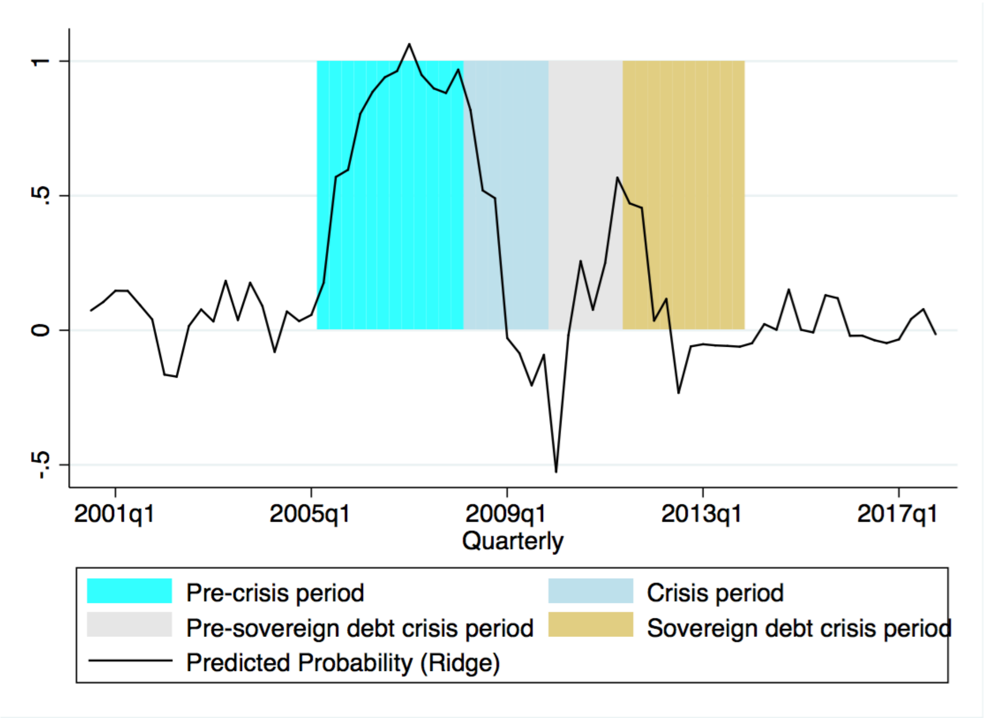

Figure 24: France: Predicted probability - Pridge (quasi-real time)

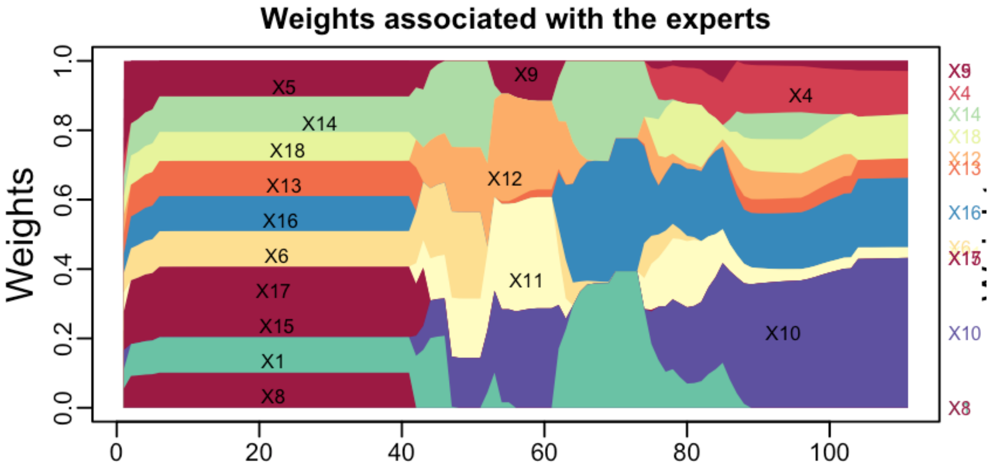

Figure 25: France: Weights. quasi-real time. ML

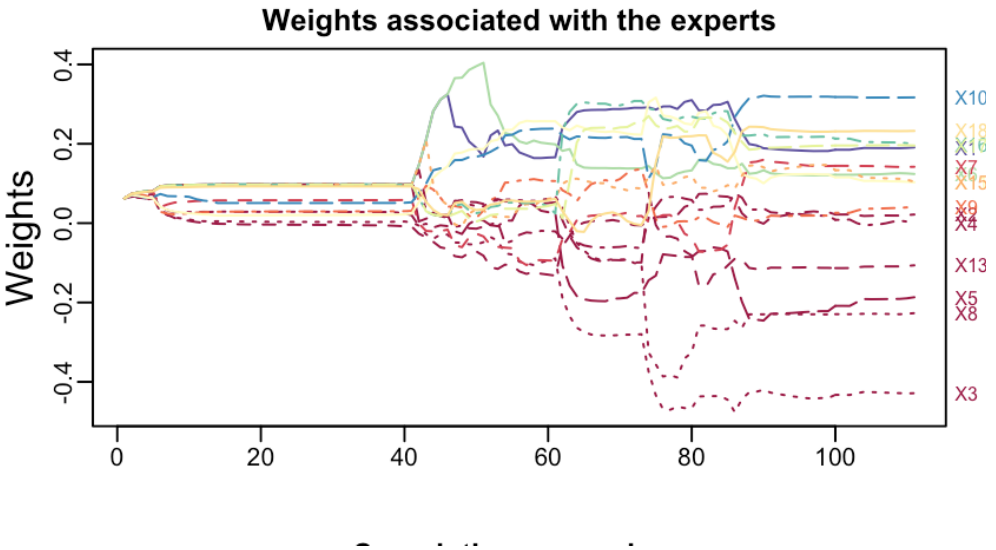

Figure 26: France: Weights. quasi-real time. Ridge

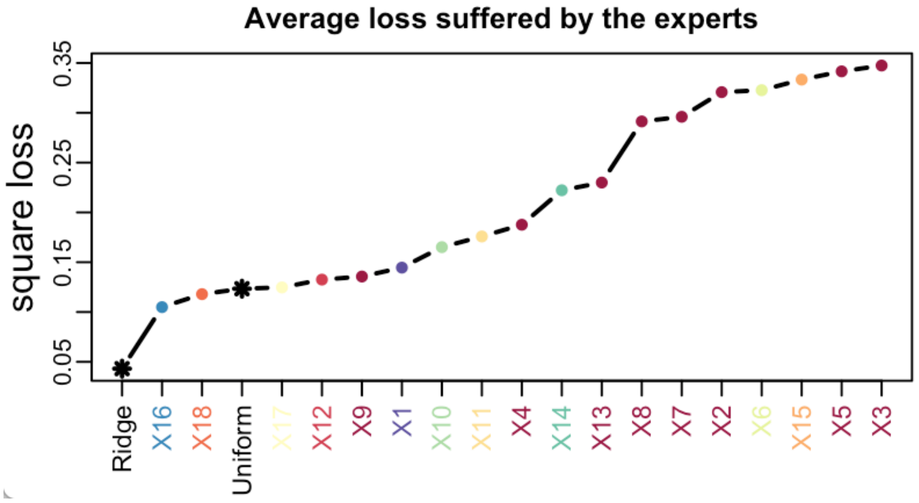

Figure 27: France: Average Loss quasi-real time. Ridge

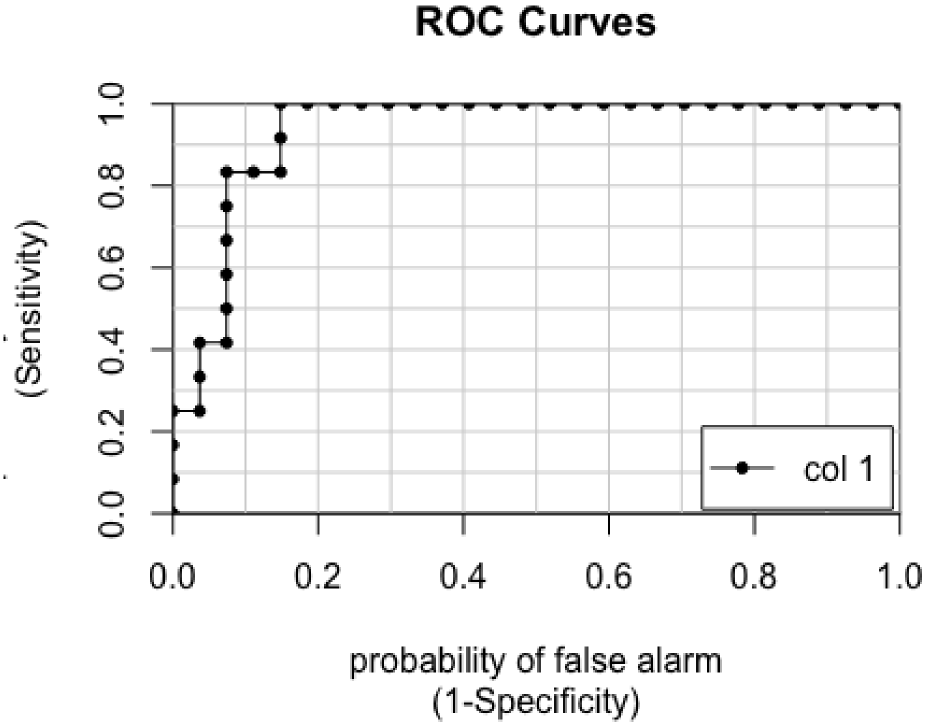

Figure 28: France: AUROC=0.938. quasi-real time from 2003Q3 to 2009Q3. OGD

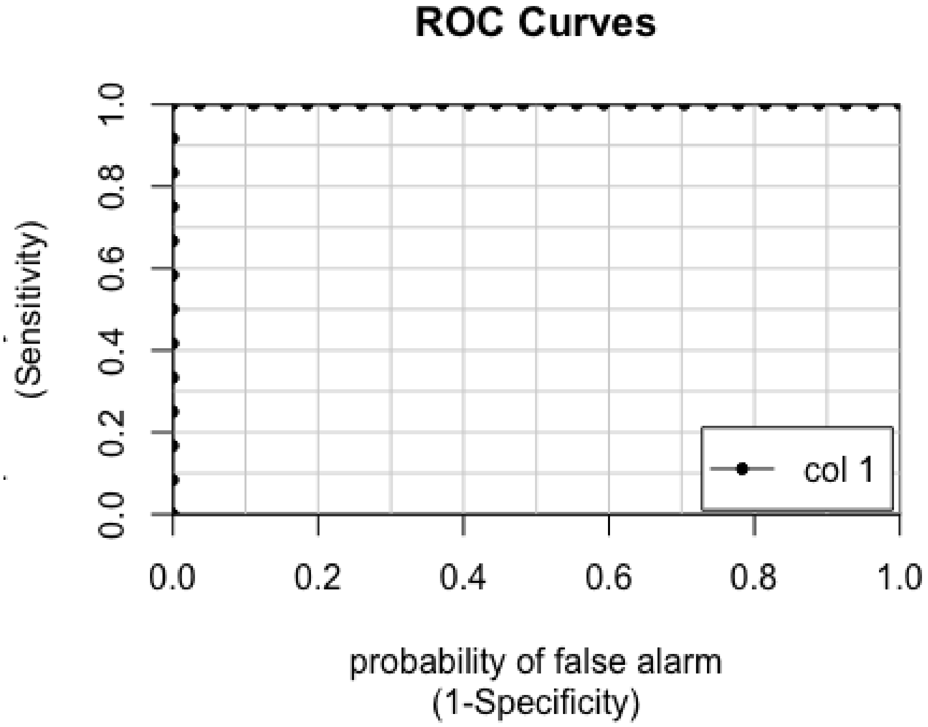

Figure 29: France: AUROC=1. quasi-real time from 2003Q3 to 2009Q3. Ridge

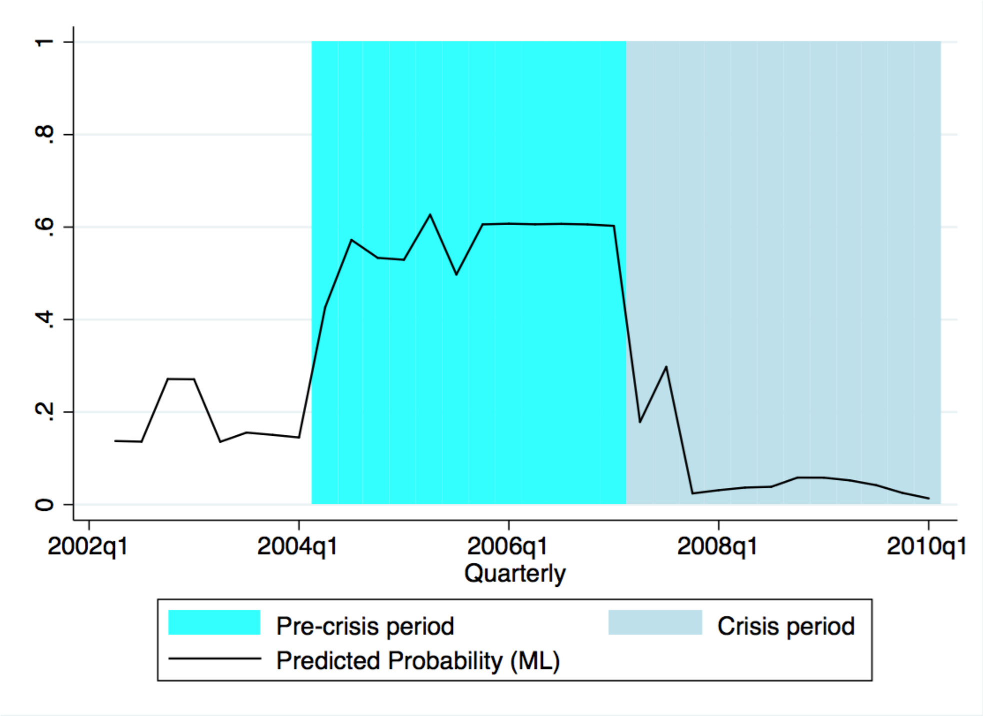

Figure 30: Germany: Predicted probability - ML (quasi-real time from 2002Q1 to 2010Q1)

E Results Germany

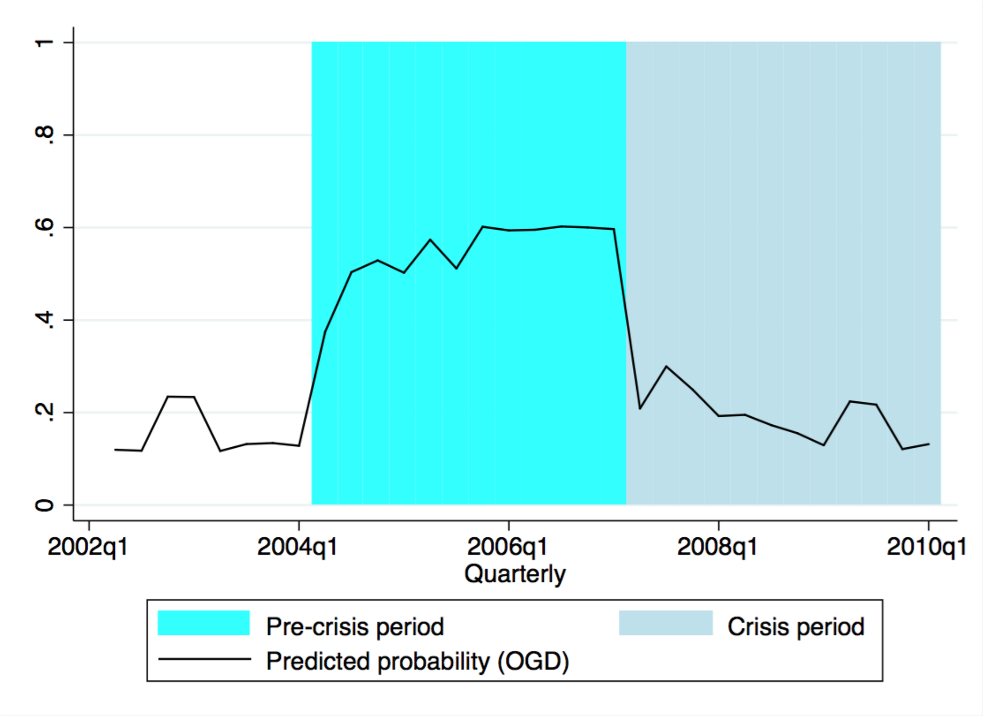

Figure 31: Germany: Predicted probability - OGD (quasi-real time from 2002Q1 to 2010Q1)

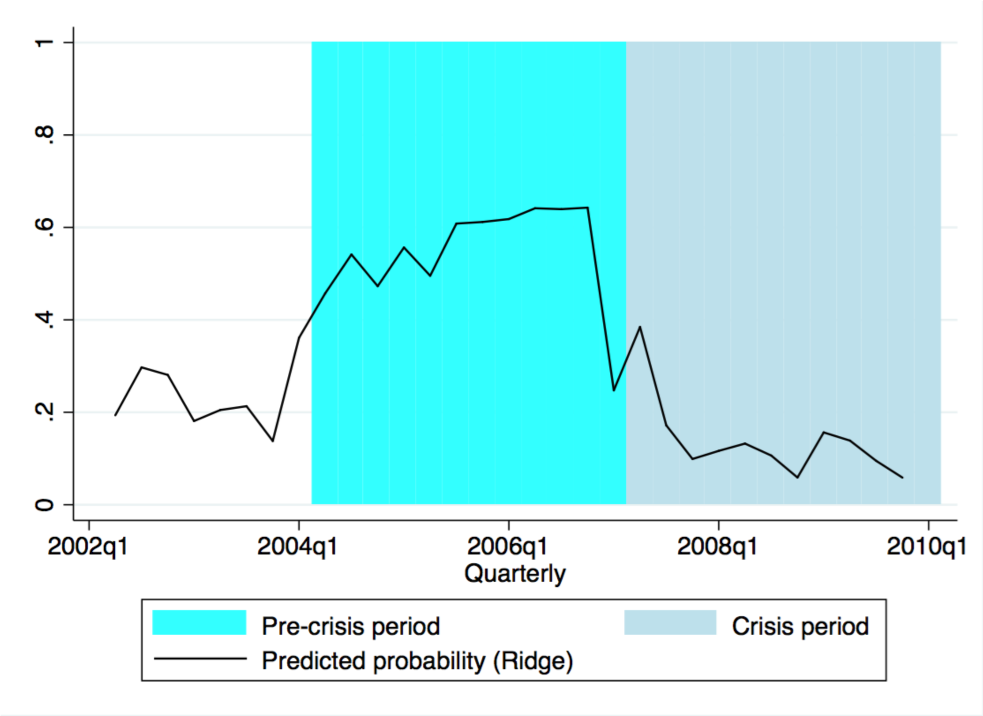

Figure 32: Germany: Predicted probability - Ridge (quasi-real time from 2002Q1 to 2010Q1

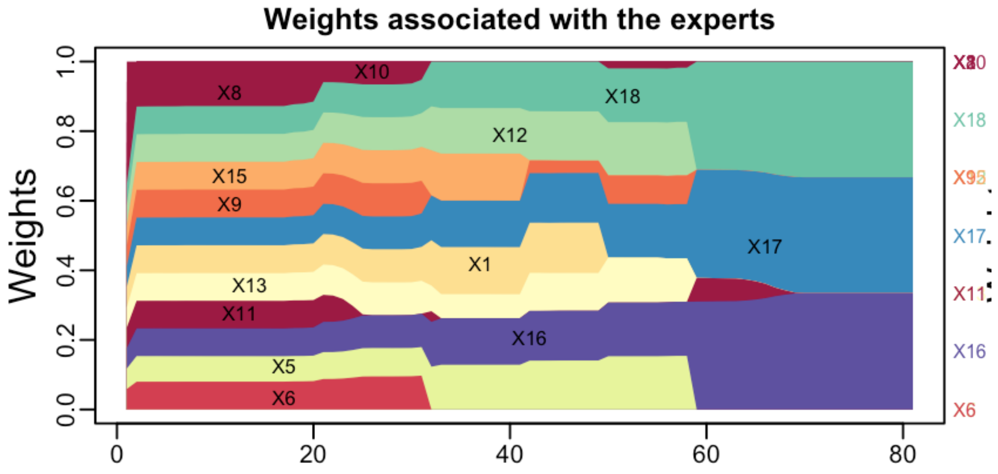

Figure 33: Germany: Weights. quasi-real time from 2002Q1 to 2010Q1. ML

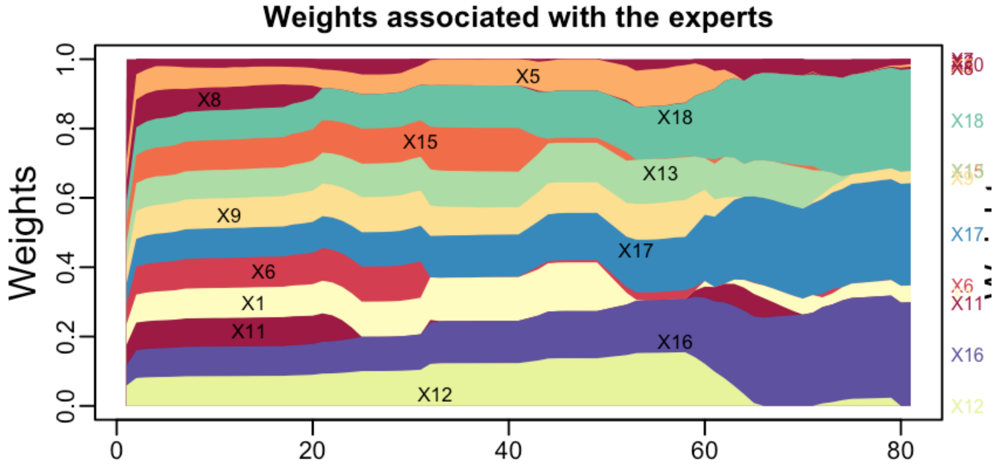

Figure 34: Germany: Weights. quasi-real time from 2002Q1 to 2010Q1. OGD

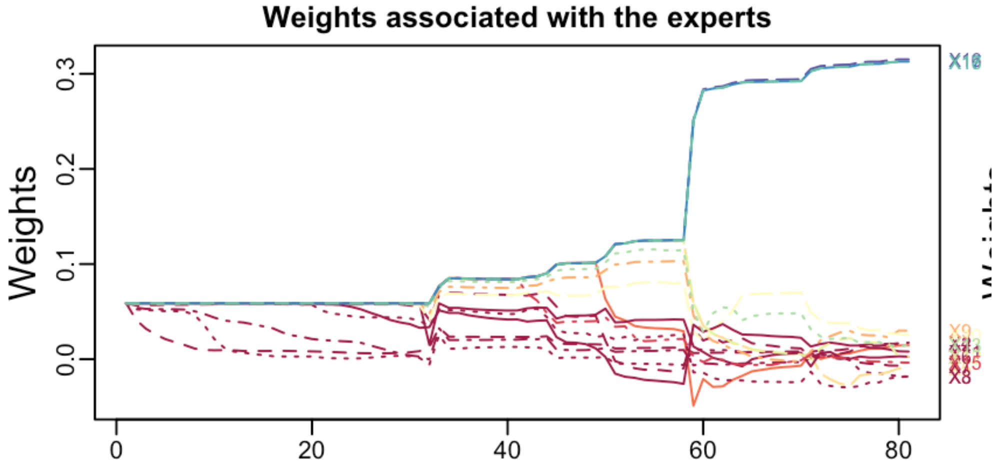

Figure 35: Germany: Weights. quasi-real time from 2002Q1 to 2010Q1. Ridge

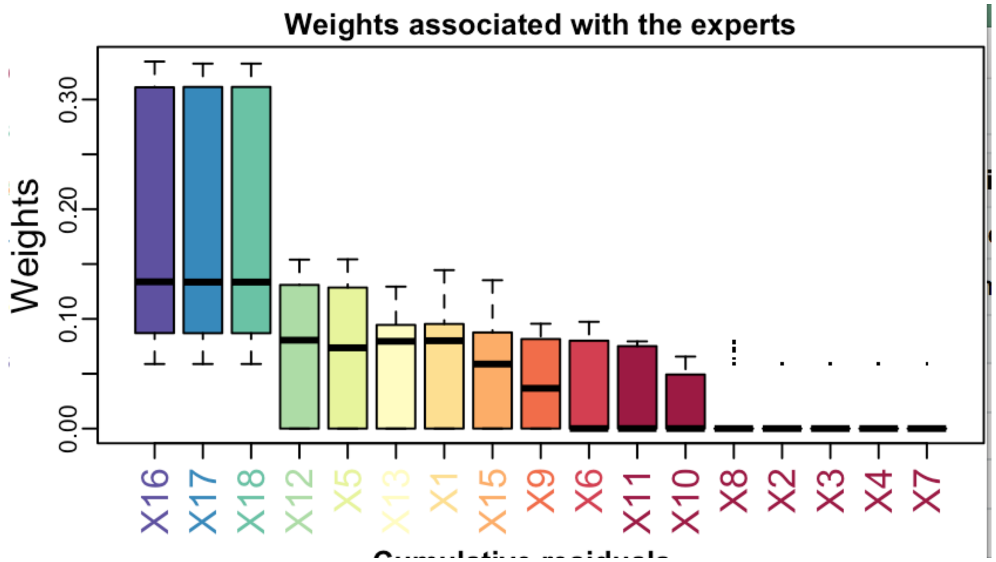

Figure 36: Germany: Distribution of Weights. quasi-real time from 2002Q1 to 2010Q1. ML

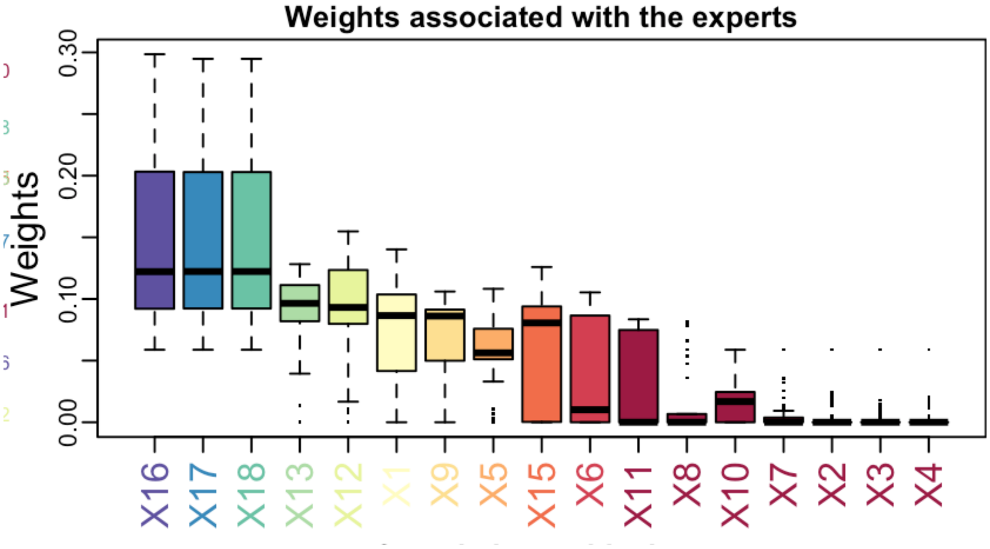

Figure 37: Germany: Distribution of Weights. quasi-real time from 2002Q1 to 2010Q1. OGD

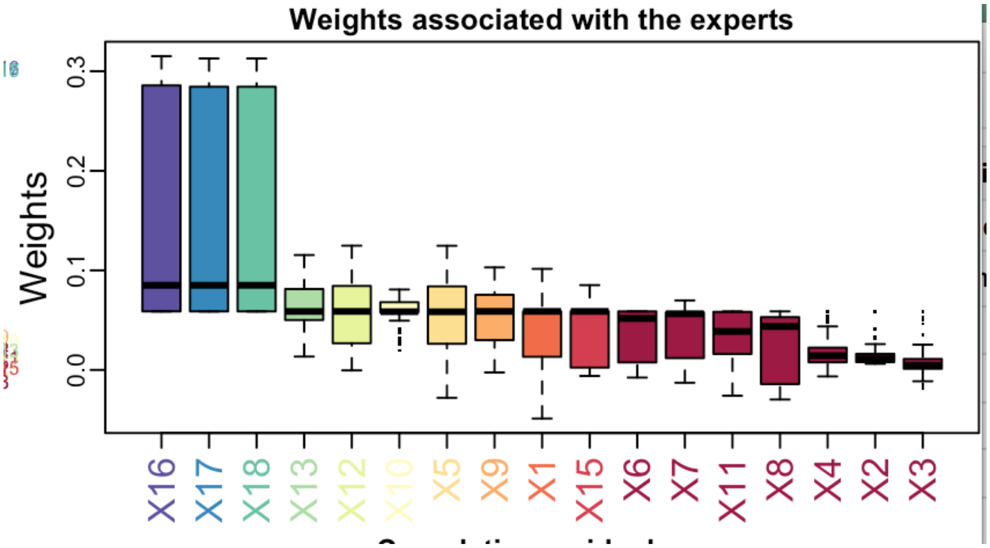

Figure 38: Germany: Distribution of Weights. quasi-real time from 2002Q1 to 2010Q1. Ridge

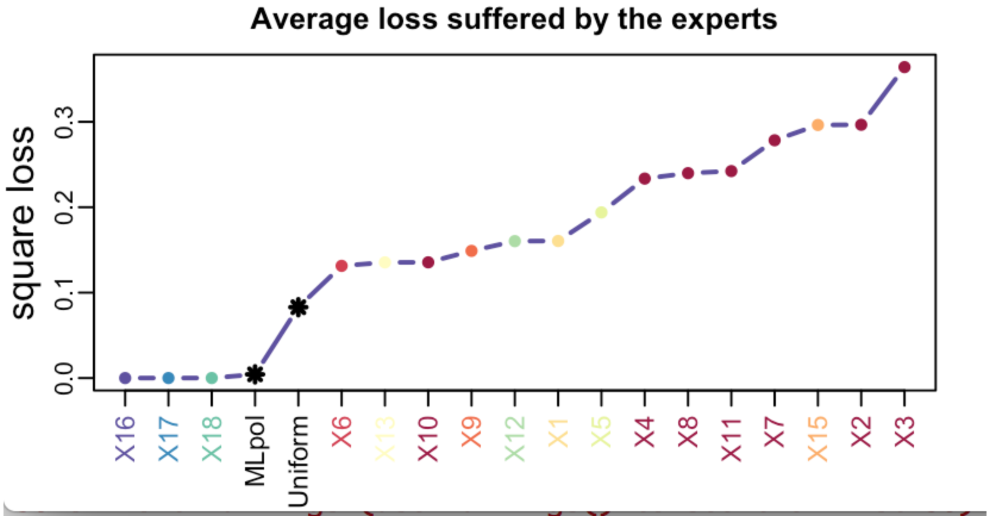

Figure 39: Germany: Average loss. quasi-real time from 2002Q1 to 2010Q1. ML

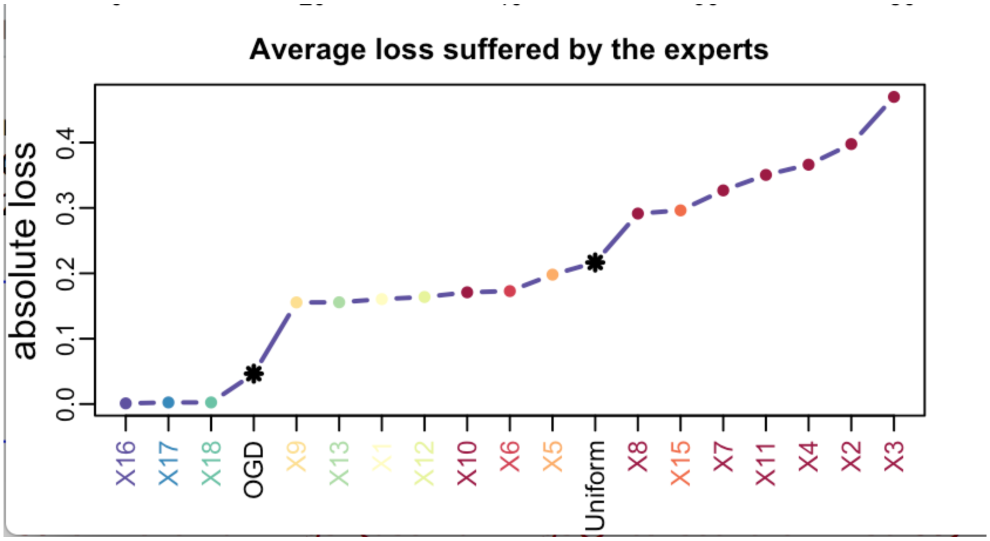

Figure 40: Germany: Average loss. quasi-real time from 2002Q1 to 2010Q1. OGD

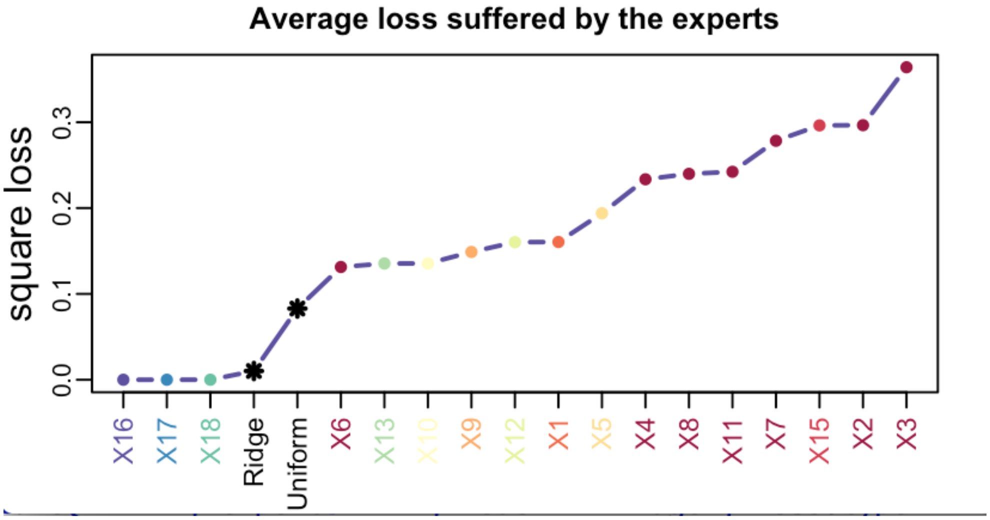

Figure 41: Germany: Average loss. quasi-real time from 2002Q1 to 2010Q1. Ridge

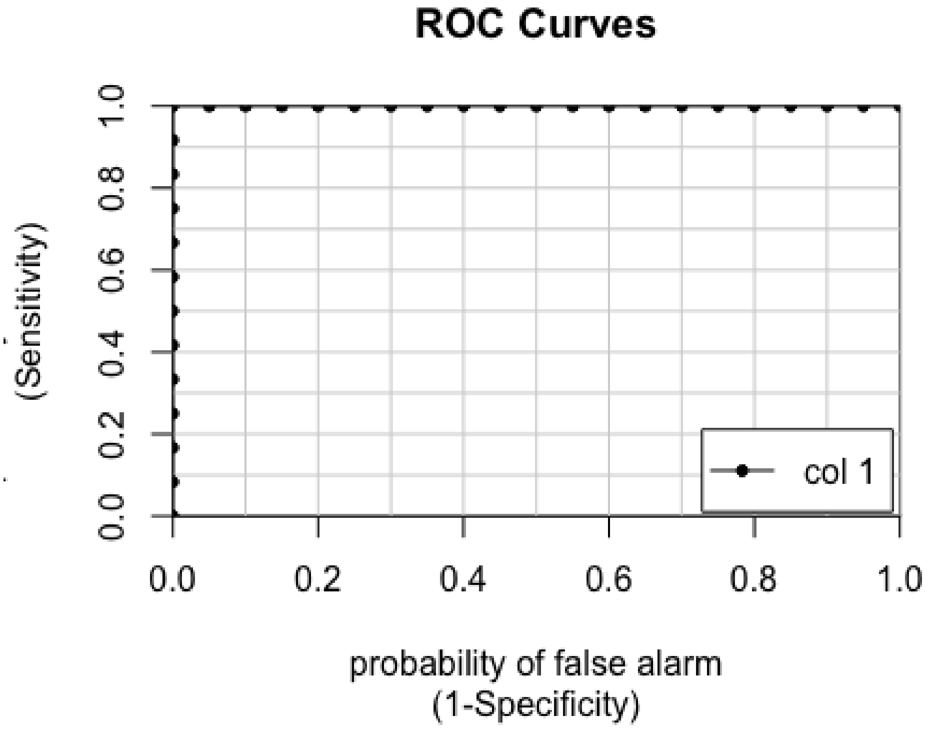

Figure 42: Germany: AUROC=1. quasi-real time from 2002Q1 to 2010Q1. OGD

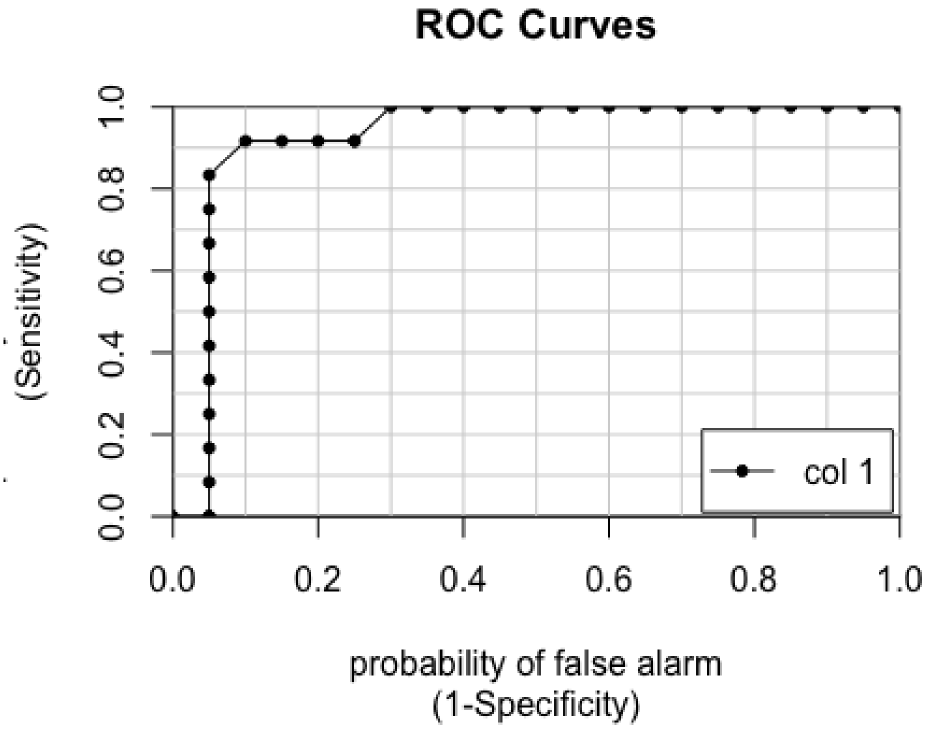

Figure 43: Germany: AUROC=0.929. quasi-real time from 2002Q1 to 2010Q1. Ridge

F Italy

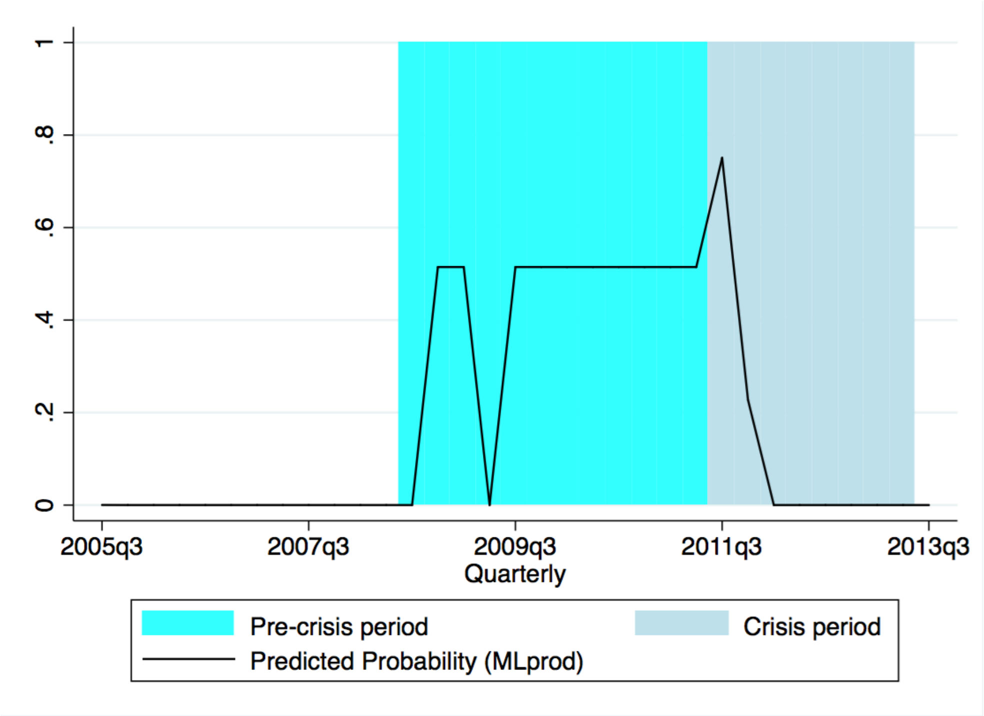

Figure 44: Italy. Predicted probability - ML (quasi-real time from 2005Q3 to 2013Q3)

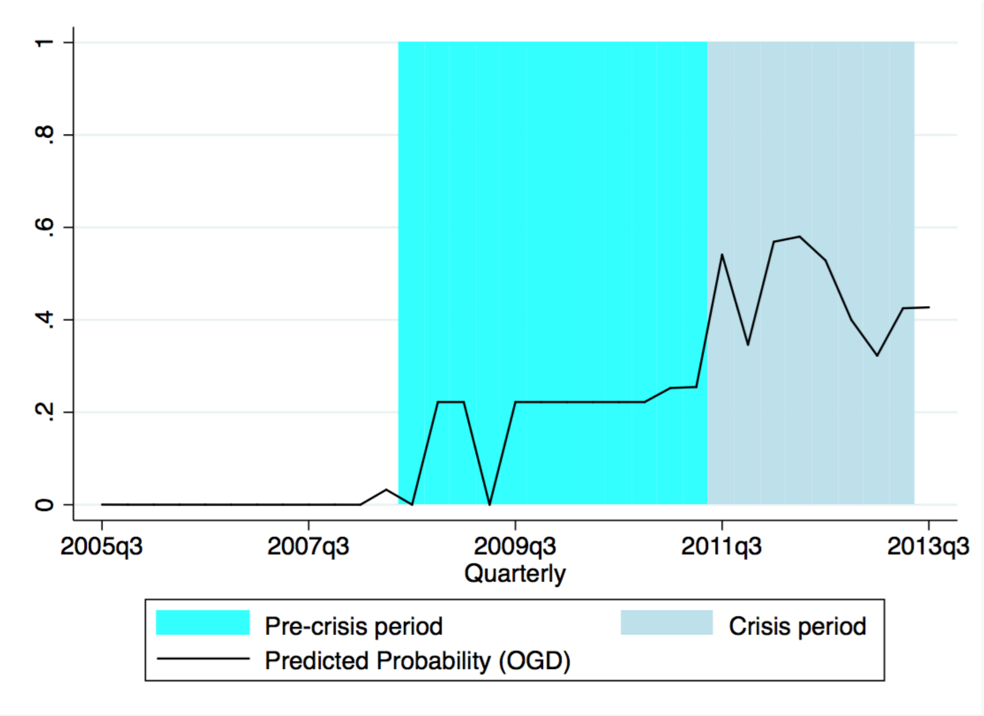

Figure 45: Italy. Predicted probability - OGD (quasi-real time from 2005Q3 to 2013Q3)

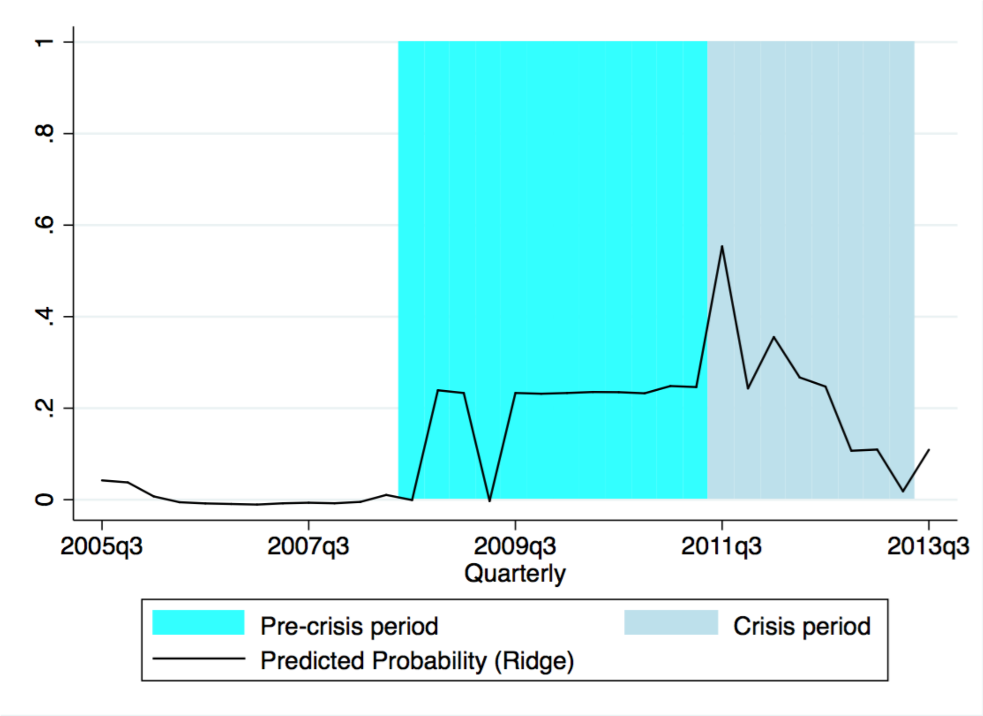

Figure 46: Italy. Predicted probability - Ridge (quasi-real time from 2005Q3 to 2013Q3)

G Spain

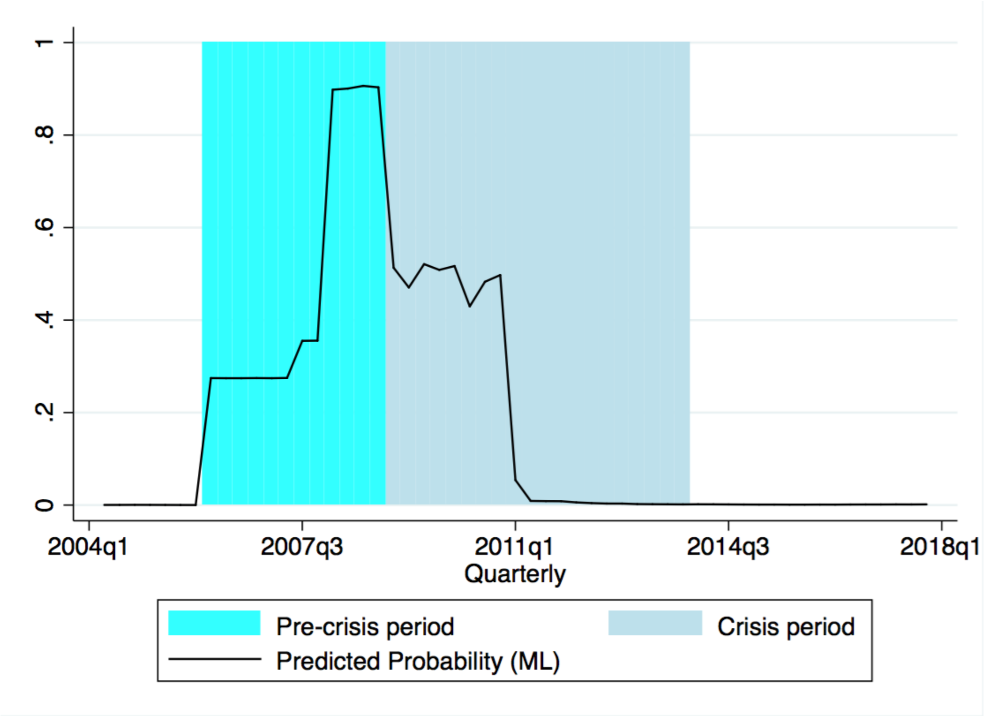

Figure 47: Spain. Predicted probability - ML (quasi-real time from 2004Q2 to 2018Q1)

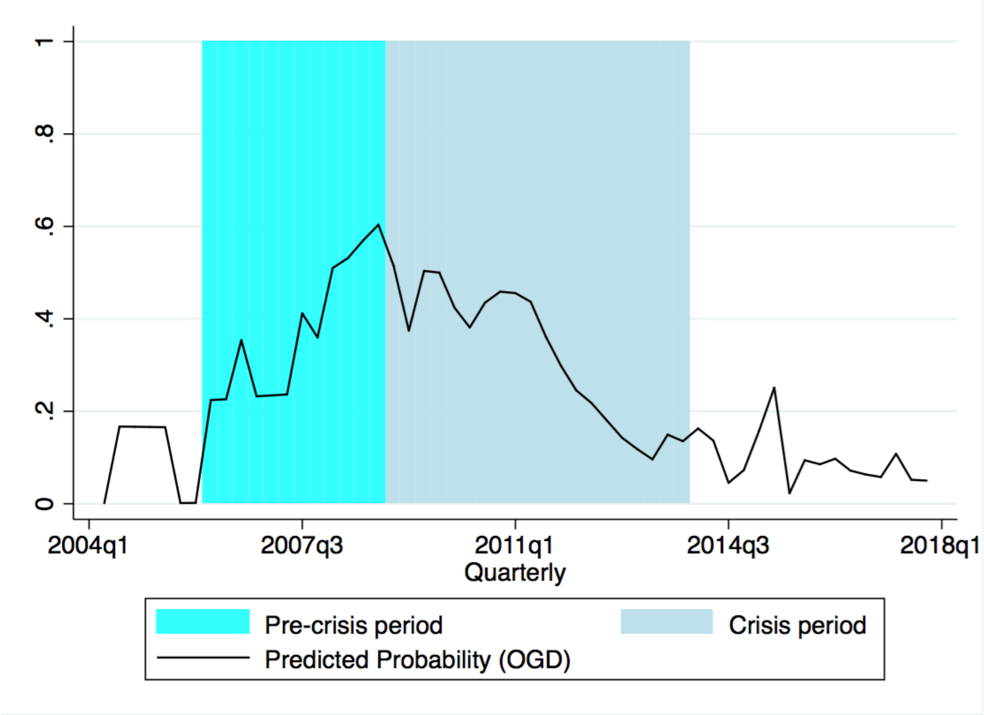

Figure 48: Spain. Predicted probability - OGD (quasi-real time from 2004Q2 to 2018Q1)

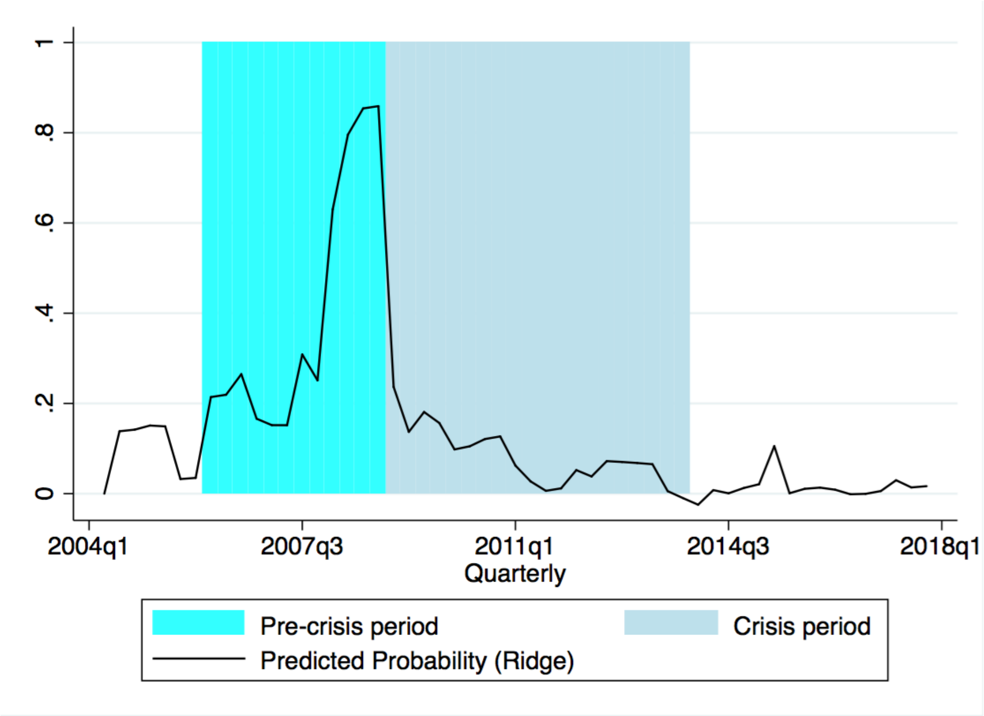

Figure 49: Spain. Predicted probability - Ridge (quasi-real time from2004Q2 to 2018Q1)

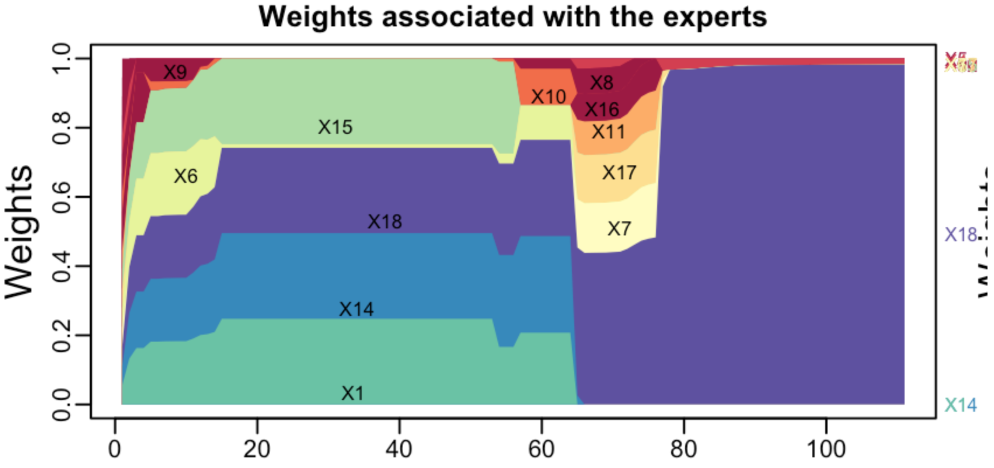

Figure 50: Spain: ML

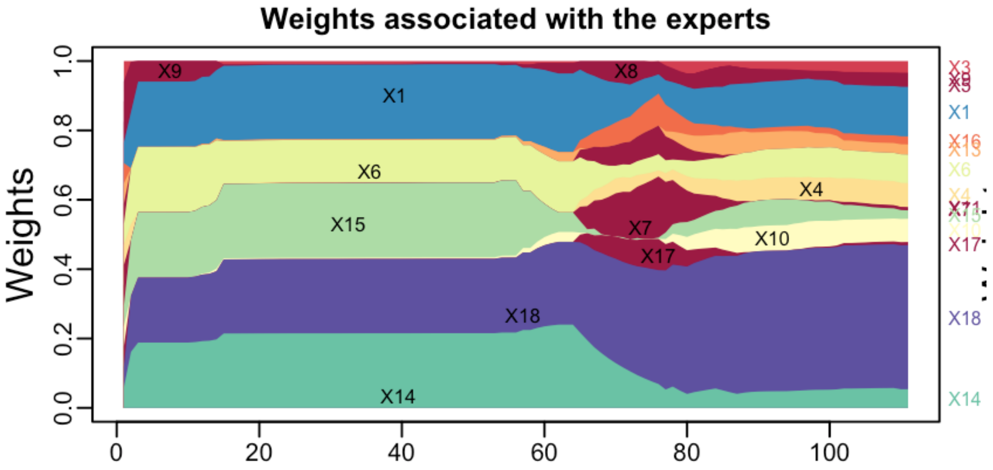

Figure 51: Spain: OGD

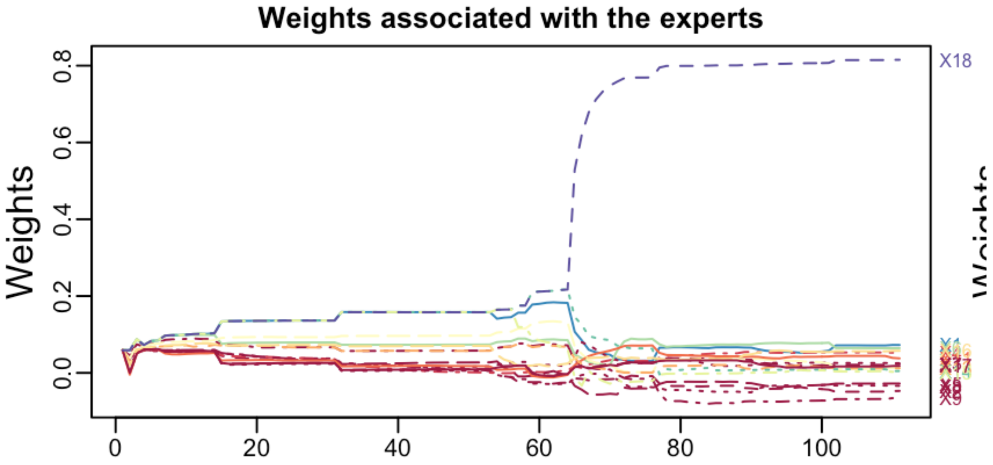

Figure 52: Spain: Ridge

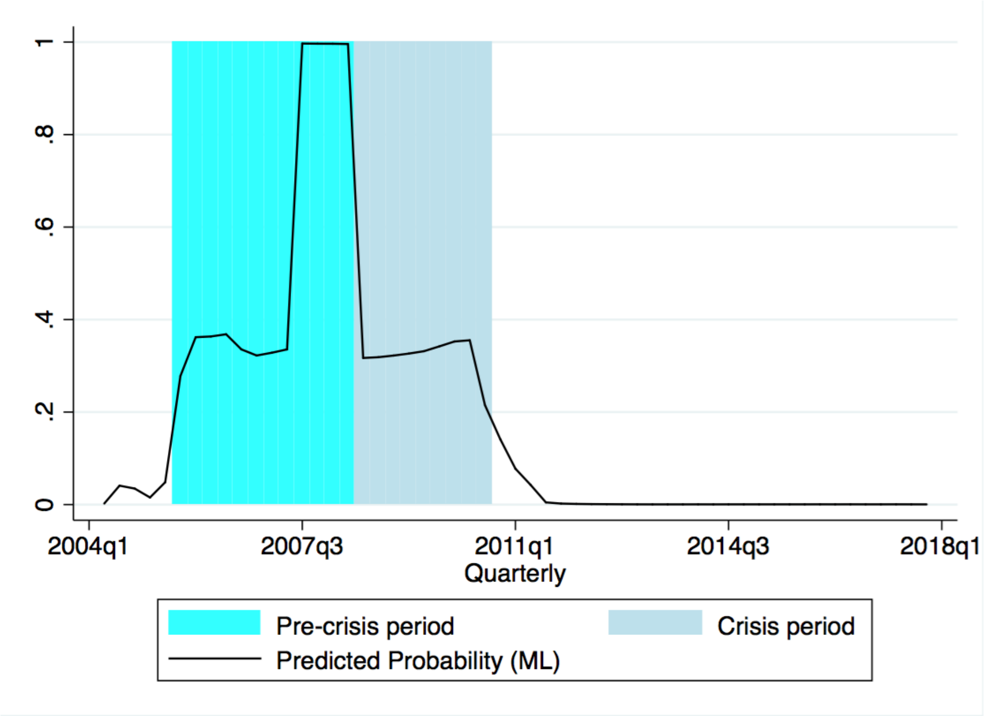

Figure 53: Sweden. Predicted probability - ML (quasi-real time from 2004Q2 to 2018Q1)

H Sweden

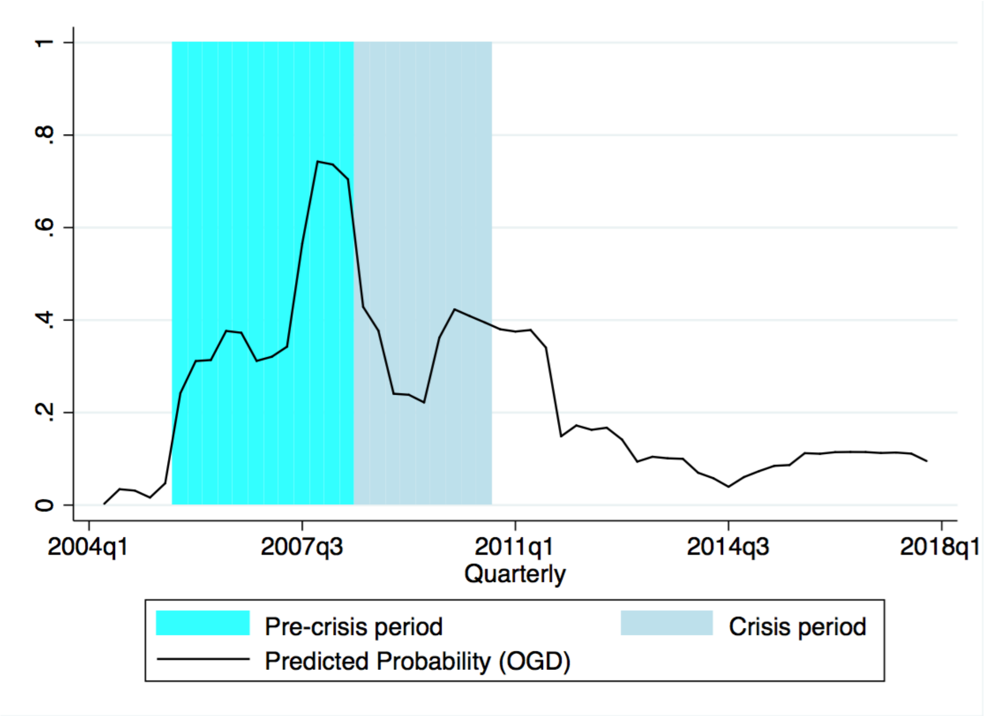

Figure 54: Sweden. Predicted probability - OGD (quasi-real time from 2004Q2 to 2018Q1)

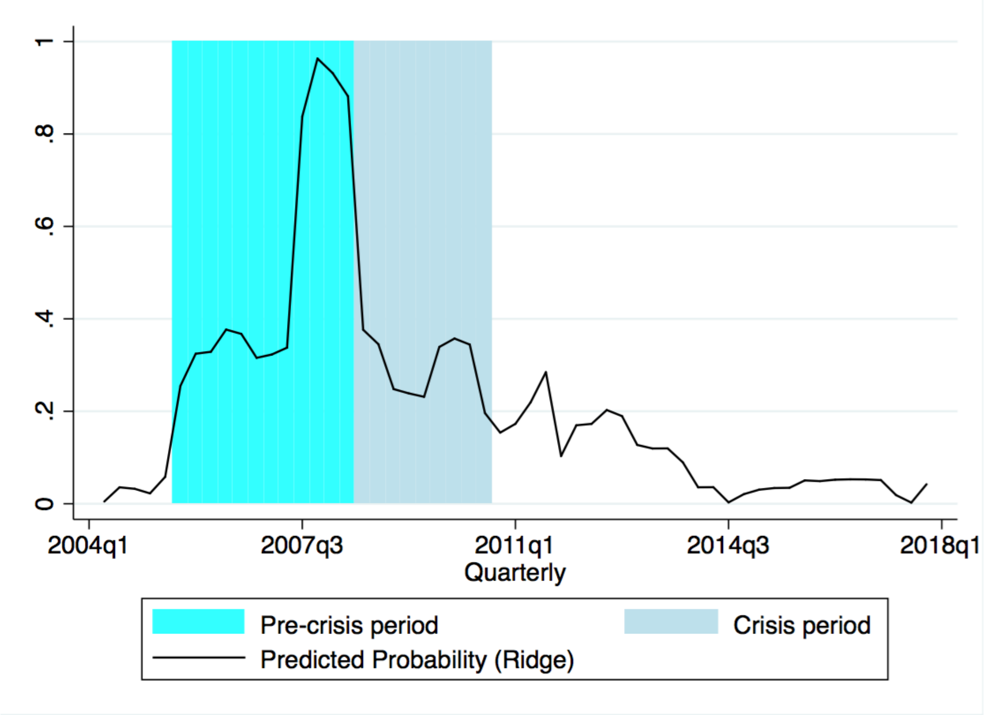

Figure 55: Sweden. Predicted probability - Ridge (quasi-real time from2004Q2 to 2018Q1)

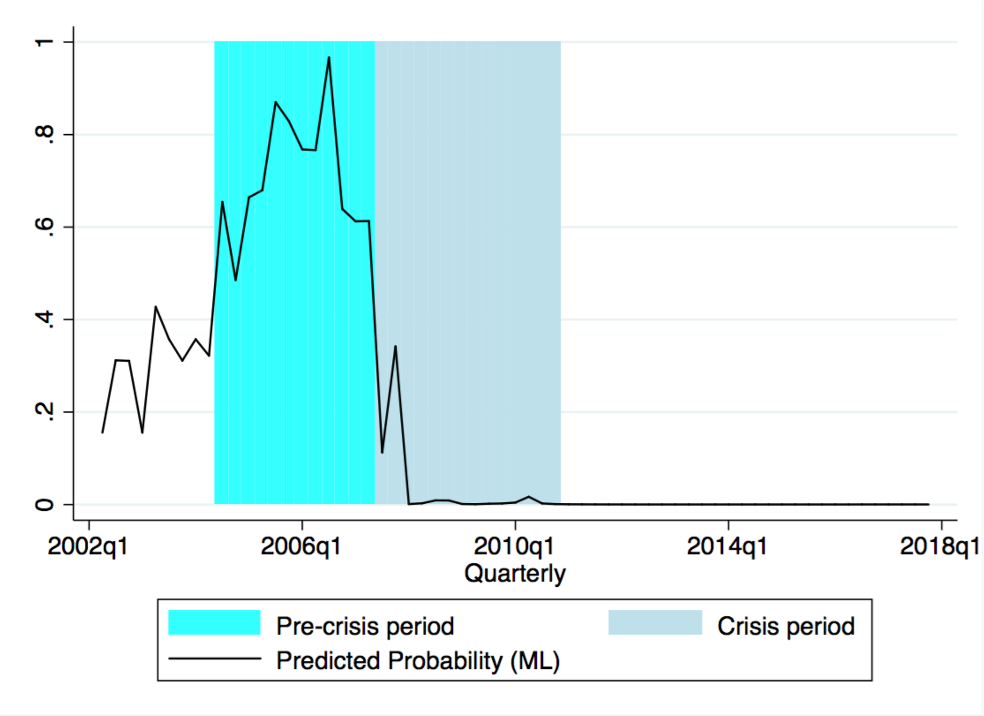

Figure 56: UK. Predicted probability - ML (quasi-real time from 2004Q2 to 2018Q1)

I UK