BANK ĊENTRALI TA' MALTA
EUROSISTEMA
CENTRAL BANK OF MALTA

Forecasting Deflation Probability in the EA: A Combinatoric Approach

Luca Brugnolini[1]

WP/01/2018

[1] Luca Brugnolini is a Senior Economist at the Central Bank of Malta's Research Department, and a Ph.D. candidate at The University of Rome "Tor Vergata". The author would like to thank Roberto Motto, Carlo Altavilla, Giacomo Carboni, Alex Tagliabracci and Roberto Di Mari for their valuable suggestions on an earlier version of the paper. He is also grateful to Antonello D'Agostino and Giuseppe Ragusa, to colleagues at the Central Bank of Malta's Research Department, and to participants of the I Vienna Workshop on Forecasting. All errors are the sole responsibility of the author.

Corresponding author's email address: brugnolinil@centralbankmalta.org

Abstract

I assess and forecast the probability of deflation in the EA at different horizons using a binomial probit model. I select the best predictors among more than one-hundred variables adopting a two-step combinatoric approach and exploiting parallel computation in Julia language. I show that the best-selected variables coincide to those standardly included in a small *New Keynesian* model. Also, I assess the goodness of the models using three different loss functions: the *Mean Absolute Error* (MAE), the *Root Mean Squared Error* (RMSE) and the *Area Under the Receiver Operating Characteristics* (AUROC). The results are reasonably consistent among the three criteria. Finally, I compute an index averaging the forecasts to assess the probability of being in a deflation state in the next two years. The index shows that having inflation above the 2% level before March 2019 is extremely unlikely.

Keywords: inflation, prediction, index, Euro Area, ECB, ROC

JEL: C25, C63, E3, E58

Contents

1 Introduction **4**

2 A motivational example **8**

3 Dataset description **9**

4 Model evaluation **10**

4.1 Discretization of inflation 10

4.2 Loss functions: the ROC curve 12

4.2.1 The ROC and the AUROC 14

5 Methodology **16**

5.1 First step 17

5.2 Second step 21

6 Results **25**

6.1 In-sample analysis 26

6.2 Out-of-sample analysis 27

7 Deflationary pressure index **31**

7.1 Survey of Professional Forecasters: a comparison 32

8 Discussion **33**

9 Conclusion **36**

A Additional table **40**

1 Introduction

Money predates history and price systems have been around as long as there has been money. In contrast, institutions with a stabilizing price mandate as central banks are newcomers, and as such, they continually need modern tools and new ideas to fill this gap. In fact, due to the complex nature of price changes, the achievement of the stability mandate is an arduous task, and central banks need to be very well-equipped to deliver. Although central banks can directly measure inflation with a regular frequency, the prototypical central bank toolbox is mainly composed by forecasting models; the reason is that central banks cannot rely only on contemporaneous inflation measures due to the lagged effects of their interventions. In fact, monetary policy actions exerted today have a lagged impact which transmits to output and prices solely in future periods. Against this background, forecasting is the leading alternative to ensure monetary policy to be timely and effective. Early assessing inflation deviations from the target allows central banks to react. Such complexity also explains why central banks employ an army of talented researchers and a tremendous quantity of resources in economic monitoring and forecasting: monetary policy is all about timing.

Given the extreme difficulty of their tasks, central banks have to revise and update their forecasting tools regularly. Among these, the most important are the models specifically developed to predict the inflation evolution. In the central bank parlance, the inflation evolution is commonly measured as the annualized percentage change in a consumer good and service market-based index, and it is critical relative to the central bank inflation target. Within the Euro Area, the institution dealing with the price stability mandate is the *European Central Bank* (ECB). For the ECB, the price stability objective is established by the *Governing Council* (GC) which is the decision-making body of the ECB itself. The original definition of price stability was an annualized increase in the Harmonised Index of Consumer Prices (HICP) for the Euro Area of below 2%. However, in 2003 the Governing Council has clarified that in reaching price stability it intends to maintain inflation rates *below, but close to, 2% over the medium term*. In achieving this target, the ECB rests on the investigation of the information based on

the economic and monetary analysis[1], and among the tools employed, forecasting inflation is the one toward which more attention is devoted. For example, the ECB together with National Central Banks (NCBs) quarterly produces macroeconomic projections (MPE/BMPE). In particular, for forecasting inflation there are at least two different class of models; the first refers to fully micro-funded structural models as the *New Area-Wide Model* (NAWM) by Christoffel et al. (2008) and the *New Multi-Country Model* (NMCM) by Dieppe et al. (2011) and Dieppe et al. (2012). While the second refers to time series models as *Vector Autoregressive* (VAR) model, Bayesian VAR (Giannone et al., 2014), and *Dynamic Factor* (DF) models. The first class of models, which often goes under the name of *Dynamic Stochastic General Equilibrium* (DSGE), has the advantage of being capable of "telling stories" about the exogenous forces acting as drivers of the business cycle. However, as a drawback, the heavy structure imposed by the model assumptions often produce unreliable results (Chari et al., 2009). On the opposite, the second class of models is endowed with less story-telling power than DSGE. Nevertheless, not restricting the model parameters often results in better forecasts.

However, there is a third approach with which the ECB assesses the inflation evolution, and this is by surveying forecasters. In fact, the Bank interviews more than 80 professional forecasters every quarter (so-called *Survey of Professional Forecasters* - SPF). The professional forecasters are members of financial/non-financial institutions within the European Union[2]. In a typical survey, they are required to express their point forecast about inflation (as well as GDP growth and unemployment) over specific time horizons. Also, they are asked to provide probabilities for different inflation outcomes. For example, they are asked to give the probability that the year-on-year (y-o-y) HICP inflation will be below, in between or above certain thresholds. The final forecast measure is the average of all the forecasts among forecasters. Although surveys are often accurate and many papers highlight their predictive ability (Faust and Wright, 2013), there are at least two main differences between them and proper in-house models; first, there is an availability limit; the SPF is deterministically released every quarter and cannot be updated as-soon-as there is the need as for in-house models. Secondly, there is an interpretation

[1] This approach is known under the name of *two pillars strategy* and describes the information set under which the Governing Council takes its decisions.

[2] A detailed list of the participating organizations is available on the ECB website - SPF list.

limit; in fact, the median of the prediction distribution is the most informative measure within the survey. This measure summarizes different models with different loss functions and would give hard times to any economist that try to interpret its movements. On the contrary, in-house models, having a known specification can be easier to interpret. From an operational point of view, these two characteristics make in-house models more attractive.

Against this background, in this paper, I propose a third way between the in-house forecasting models employed by the ECB and the probabilities measured in the SPF. In particular, I tailored a model to predict the inflation probabilities directly. Concerning the SPF, this tool has the advantage to be extremely easy to interpret given that the model specification is known. Thus, when a forecast displays some curious behavior, the forecaster can trace back the variable which causes it. Secondly, the model has the advantage to be updatable as soon as new variables get released. While concerning DSGE, VAR and factor models, it has the benefit to be tailored for density forecast. In fact, with continuous dependent variable models, a forecaster calibrates the model only on point forecasting, then, in case there is the need, he also computes the predictive density. Instead, with discrete models, a forecaster calibrates the model directly on density forecast. Secondly, using discrete models, in addition to standard metrics as the *Mean Absolute Error* (MAE) and the *Root Mean Squared Error* (RMSE), a forecaster can employ specific model selection criteria tailored for this class of models. The most prominent example is the *Area Under the Receiver Operating Characteristics* (AUROC).

In this paper, to assess the probabilities of over/undershooting the inflation target, I investigate the predictive power of a large dataset of macroeconomic variables at different future horizons. In dealing with the dataset dimension, I perform a two-step variable selection procedure, and I use a combinatorial approach to retrieve the best model for each forecast horizon considered. In setting up the empirical exercise, I directly forecast probabilities using a binomial probit model. I choose the 2% inflation level as a natural cutoff point. In fact, many central banks have this level as the inflation target, and the ECB approximately follow this rule ("approximately" because the precise objective is *"below, but close to 2%"*). In this respect, I forecast the probability of having inflation above/below the target at short and medium horizons.

In setting up the exercise, the main issue is that a forecaster needs to know the precise horizons to predict. On the contrary, the correct implementation of monetary policy actions is related to a general medium-term orientation. The reason is that fluctuations in prices due to exogenous shocks make impossible to secure inflation at any point in time. Therefore, the interval for achieving price stability has to be extremely general. The lack of a precise definition join to the delayed effect of the monetary policy actions makes the forecasters' life much more involved, imposing the need for a set of models calibrated for different horizons. With that in mind, I propose a tool constructed by averaging the estimates of a set of forecasting models tailored for a grid of short to medium term horizons. The main idea connected to this choice is that macroeconomic as well as financial variables have different predictive power at distinct horizons, and a single model unlikely produces the best forecast at different steps-ahead. In this respective, I average the forecasted probabilities from the best horizon-calibrated models, and I create an index to predict the likelihood of having inflation below the 2% level in the next two years. The index shows that the probability of having inflation higher than 2% before March 2019 is extremely low. Finally, to benchmark the index, I compare it to a measure built from the inflation probability forecasts of the ECB SPF. I show that the two are broadly in line, even if the SPF often fails in capturing the turning points between the two states.

The rest of the paper is organized as follows. Section 2 presents a motivational example and shows why it is desirable to forecast the inflation probabilities. Section 3 sketches the dataset used in the paper. Section 4 highlights the loss functions used to evaluate the models and presents the ROC/AUROC. Section 5 describes the forecasting methodology along with the two-step selection procedure. Section 6 presents the results of the in-sample and out-of-sample forecast for all the selected models at different horizons. Section 7 describes the *Deflationary Pressure Index* and compares it with a probability measure built from the ECB SPF. In section 8 I discuss the results of the paper. Finally, section 9 concludes.

2 A motivational example

In the last decade, density forecast has become a prominent tool to assess the possible outcomes of macroeconomic indicators. For example, the Bank of England *fan chart* has been used extensively to present the forecast distribution of inflation. In general, supporting point forecasts with probabilities provides a quantitative assessment of the forecaster uncertainty and can help policymakers in taking decisions. However, probabilities are intrinsically informative and can be a primary source of knowledge. For example, knowing with which probability inflation undershoots the central bank target can help policymakers to decide on interest rate cuts. The argument can be heuristically formalized with the help of a forward-looking Taylor rule as in Equation (1)[3].

$$i_t = \phi_\pi(\mathbb{E}_t \pi_{t+h} - \pi^*) \tag{1}$$

Where i_t is the interest rate under the control of the central bank, π_t is the inflation rate, π^* is the inflation target, ϕ_π is the central bank reaction coefficient, and $\mathbb{E}_t \equiv \mathbb{E}(\cdot|\Omega_t)$ is the expectation operator given the information set at time t (Ω_t) . The Taylor rule determines the central bank interest rate direction in response to price deviations from the target. According to Equation (1), when $\mathbb{E}_t \pi_{t+h} > \pi^*$ the central bank increases the interest rate while when $\mathbb{E}_t \pi_{t+h} < \pi^*$ the opposite happens. Following the rule, given ϕ_π, the point forecast is necessary and sufficient to know the magnitude of the interest rate adjustment. However, it is sufficient but not necessary to identify the direction of the policymakers' action. What is necessary is to know whether inflation will be above or below the target, and, indeed, this information is readily assessed through probabilities. Suppose that the central bank target is 2% ($\pi^* = 2$) and that the reaction coefficient is $\phi_\pi = 1.5$. Then, if the estimated inflation ($\hat{\pi}_{t+h}$) is 2.2, the interest rate has to increase by 0.3 percentage points. However, to know the direction of the interest rate change, it would have been sufficient to recognize whether $\hat{\pi}_{t+h}$ were higher or lower than the target. This knowledge would have triggered a reaction by the central bank in the same direction of the distance from the target.

[3]The Taylor rule may also depend on other variables, as the output-gap or lagged interest rate, but for an illustrative purpose, I am abstracting from these.

As this simple example shows, having a tool tailored to predict inflation probabilities around a meaningful threshold can be tremendously informative. This is especially true recalling that point forecast is by definition more susceptible to forecast errors than interval forecast. First, knowing in which direction the inflation will exceed the target, it is informative from the policy-makers point of view, as it directly communicates possible up-side or down-side risks. Secondly, it can be effective in informing external entities like banks or market makers on future central banks' actions. Accordingly, in this paper, I focus on building a tool to forecast the probability that inflation exceeds the central bank target, and particularly I tailor the model on EA data and the ECB monetary policy, as described in the following sections.

3 Dataset description

I build a large dataset comprising around 100 monthly variables at national and Euro Area level starting in January 1999 and ending in March 2017 (219 observations). Table 9 in the Appendix shows the complete monthly dataset and the respective identification codes. All the data are provided by Thomson Reuters Eikon and Datastream. The only exception is the Wu and Xia (2016) shadow rate measure for the Euro Area as in Wu (2017) which is available on their web-page. The dataset has five different broad categories:

1. Real indicators: these correspond to real economic activity measures as production, consumption, government spending, import and export activities for the EA and the largest European countries.

2. Price indicators: these correspond to seasonally and non-seasonally adjusted indexes of consumer prices comprising different aggregate categories at both EA and national level.

3. Monetary aggregates: these are the monetary aggregates M1, M2, and M3 which include currency in circulation, deposits and liquid financial products.

4. Financial variables: these include the European Overnight Index Average (EONIA), the

Euro Inter-Bank Offered Rate (EURIBOR) at different maturities, the Nominal and Real Effective Exchange Rate (NEER-REER), the US Fed Fund rate, European and US bonds, stock indexes, volatility indexes and oil prices.

5. Surveys: these correspond to confidence indexes and professional forecaster surveys.

Each series is transformed to be approximately stationary. All the transformations and respective codes are reported in Table 9.

4 Model evaluation

In this section, I describe the empirical methodology adopted in the paper. Firstly, I outline the process employed to build the dependent variable. In fact, the main difference concerning inflation probability forecast and standard recession prediction is in the choice of the dependent variable. Models tailored to predict the recession probabilities normally use as dependent variable a binary measure. This measure is computed by independent research organizations which assess and release a discrete variable to track recession periods [4]. For inflation, a clear counterpart does not exist. Nevertheless, a very satisfying and intuitive alternative can be found by clustering inflation realizations in points below and above the central bank target. Secondly, in this section, I outline the model evaluation procedure employed in this paper to select the best predictive variables at different horizons and in particular I describe the less-known AUROC metric.

4.1 Discretization of inflation

Within the Euro Area, many inflation metrics exist. However, these measure inflation as a continuous variable (π_t). In particular, the ECB definition of the inflation target is in terms of year-on-year change in the *Harmonized Index of Consumer Prices* (HICP). There are other popular measures[5]; however, as the paper focus on forecasting inflation from a central bank

[4] For example, in the Euro Area the recession indicator is computed by the *Centre for Economic Policy Research* (CEPR), which is an independent organization. In the United States, the *National Bureau of Economic Research* (NBER) performs the same task.

[5] For example, the *GDP deflator* or the *core inflation*. The former is the ratio between nominal and real GDP. The latter is the HICP excluding food and energy.

viewpoint and the ECB target is in terms of HICP, I will only focus on this measure.[6] To discretize the inflation measure and create the binary dependent variable (Π_t), I divide the HICP year-on-year change π_t into two different categories. I choose as a threshold the 2% level, as many central banks have this cutoff as a target, and I use it as an approximation for the ECB target. Thus the dependent variable looks as follows:

- Inflation below the confidence zone ($\Pi_t = 1 \; if \; \pi_t < 2\%$).
- Inflation above the confidence zone ($\Pi_t = 0 \; if \; \pi_t \geq 2\%$).

The first panel of Figure 1 shows the year-on-year HICP for the EA (π_t) from January 1999 to March 2017. The solid blue line shows the monthly level in percentage points; the vertical gray bars highlight periods in which inflation is below the 2% level, by contrast, the "white bars" show periods in which HICP is above or equal to 2%. The second panel of Figure 1

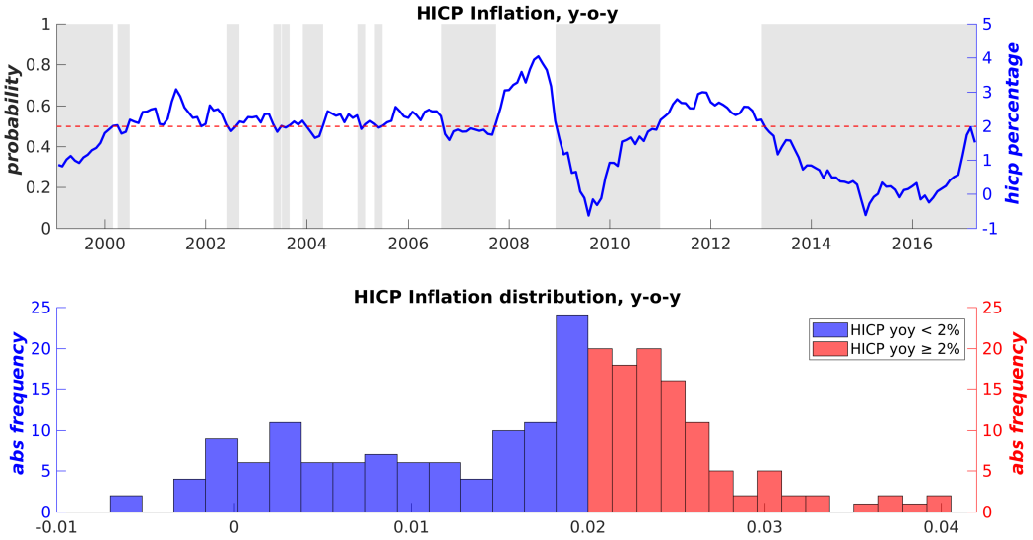

Figure 1: The upper panel shows the y-o-y HICP for the EA (solid blue line) and highlights the 2% inflation level (red dashed line). The lower panel shows the inflation distribution. Observations greater or equal than 2% are highlighted by red bars, while data points lower than 2% are reported as blue bars.

shows the HICP sample distribution. The colors highlight the composition of the discretized

[6]Notice that forecasting from a central bank viewpoint does not imply that only central bankers can benefit from this study. Indeed, also the agents interested in knowing the inflation direction from a central bank viewpoint can take advantage of this modeling strategy. In fact, from a market viewpoint, it is more profitable to predict the central bank inflation expectation then knowing the true value of inflation itself. This is because the central bank forecasts are the determinant of the central bank actions. In turn, CBs actions, are powerful market movers and can be exploited by the agents in the markets to secure some profits.

HICP variable using 2% as a cutoff point. The blue bars (left-hand-side) show the portion of the distribution below the threshold while the red bars (right-hand-side) display the observations above or equal to it. The y-axis shows the absolute frequency of each bin. As built, the binary variable for inflation is well balanced along the entire sample. It displays 112 observation below the threshold and 107 above. From the chart, it is easy to notice that the mass tends to locate around the cutoff point. Indeed, the mode is located slightly below the 2% level, consistently with the ECB mandate. Also, it is interesting that while the right tail of the distribution concentrates around the threshold, the left tail is longer and exhibits more dispersion. This characteristic is mainly due to the recent deflationary period experienced by the Euro Area, which has led inflation in negative territory for the first time after the great recession.

4.2 Loss functions: the ROC curve

Forecasting is a particular case of a decision theory problem (Elliott and Timmermann, 2016), as such, it follows some precise rules to evaluate possible outcomes (models). To evaluate a model, the most important concept that a forecaster has to keep in mind is the loss function. Different loss functions reflect different weights a forecaster puts on the same forecast error. And, of course, different weighting schemes attribute models with the same output different scores. This, in turn, affects model selection. Previous research has employed different metrics, but, the most selected are symmetric loss functions as the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE). The reason why researchers often select standard loss functions is that, especially in economics, the choice of the loss function is often disregarded. This fact is true even if the selection of an appropriate loss function is intrinsically related to the problem faced by the forecaster. In fact, the only forecaster exempted from an accurate choice of the loss function is the one that commits no errors. Unfortunately, those who not belong to this class need a metric to evaluate the distance of their predictions from the true realizations. The MAE and the RMSE are extremely valid loss functions when a forecaster has symmetric disutility in over/underestimating the outcome, however, this is not always the case. Also, the MAE and the RMSE differ from the fact that the former penalizes errors more than the latter. Then, models selected using this two standard loss functions often choose different variables.

Equation (2) and (3) present the MAE and the RMSE for a discrete variable model.

$$MAE = \frac{1}{T}\sum_{t=1}^{T}\left|\hat{\Pi}_{t+h|T} - \Pi_{t+h}\right| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\Pi}_{t+h|T} - \Pi_{t+h}\right)^2} \tag{3}$$

Where $\hat{\Pi}_{t+h|T}$ is the predicted probability estimated from a discrete model. However, these loss functions are tailored for continuous regression models and evaluate the model fit. In a discrete context, as the case of forecasting inflation probability, the prediction exercise is closer to a classification problem. Thus, given the estimated probability, a researcher assigns each forecast to the correct class $\{0, 1\}$.

For this class of problems more appropriate loss functions have been extensively studied in many scientific fields. Among these, a particularly well-tailored criterion for binary classification problems is the Receiving Operator Characteristics (ROC). The ROC is a metric computed in several steps, and in a nutshell, it attributes a score to different models based on their ability to correctly classify observations among the whole spectrum of the possible cutoff points. The reader should notice that the cutoff described in this section has nothing to do with the cutoff used to construct the dependent binary variable. In this context, the cutoff means the point in the estimated probability above which an observation is classified as a one or zero. In this sense, the main feature of the ROC is the ability to evaluate models along with all possible cutoffs and to attribute a score according to the best split. In this way, models can be evaluated without selecting a unique threshold. Despite its proved ability, it has only been used in recent times in the economic literature (Berge and Jordá, 2011; Liu and Moench, 2016) and still it is unclear whether it provides a better assessment than more standard methodology. I take a step in this direction by comparing in the exercise different models based on both the standard and the ROC methodology. In what follows I provide a brief description of the ROC computation procedure.

4.2.1 The ROC and the AUROC

The first step implies evaluating the model ability to assign an observation to the correct class (*True Positive*, TP – also called *sensitivity*) or to the wrong class (*False Positive*, FP – also called *fall-out*) for all possible thresholds in the estimated probability. The set of thresholds is approximated by a discrete variable bounded between zero and one. Equation (4) and (5) show the difference between these thresholds and the one used to compute the discrete dependent variable.

$$\begin{cases} \Pi_t = 1 \quad if \quad \pi_t < 2\% \\ \Pi_t = 0 \quad if \quad \pi_t \geq 2\% \end{cases} \tag{4}$$

$$\begin{cases} \hat{\Pi}_t = 0 \quad if \quad \hat{\Pi}_t < C_i \\ \hat{\Pi}_t = 1 \quad if \quad \hat{\Pi}_t \geq C_i \end{cases} \tag{5}$$

Where $C_i \in [0,1]$, $i = 1, 2, ..., I$. For example, in the first step, a researcher estimates a discrete dependent variable model. In a discrete model, computing the conditional expectation corresponds to estimate the entire model density. Therefore, the probability of having a particular outcome can be compared against a threshold C_i. Then the outcome can be classified accordingly. Repeating this process allows assessing the model classification ability. This is achieved by comparing the classification in (4) and (5). The result can be represented in a plane having the percentage of *TPs* on the y-axis and *FPs* on the x-axis (*FP(C_i),TP(C_i)*). Figure 2 shows the ROC for a probit model estimated with the discrete version of inflation as the dependent variable and using as regressors all possible variables in the dataset.

By moving along each curve, a researcher can gather the model trade-off between true and false positives. Moving from left to right tells the percentage of false positives that have to be tolerated to increase the rate of true positives. Also, it is crucial to mention the following characteristics:

1. In the (*FP(C_i),TP(C_i)*) plane, the 45-degree-line is a random guess equivalent, and it is often used as a reference line (50% probability of having both *TP* and *FP*).

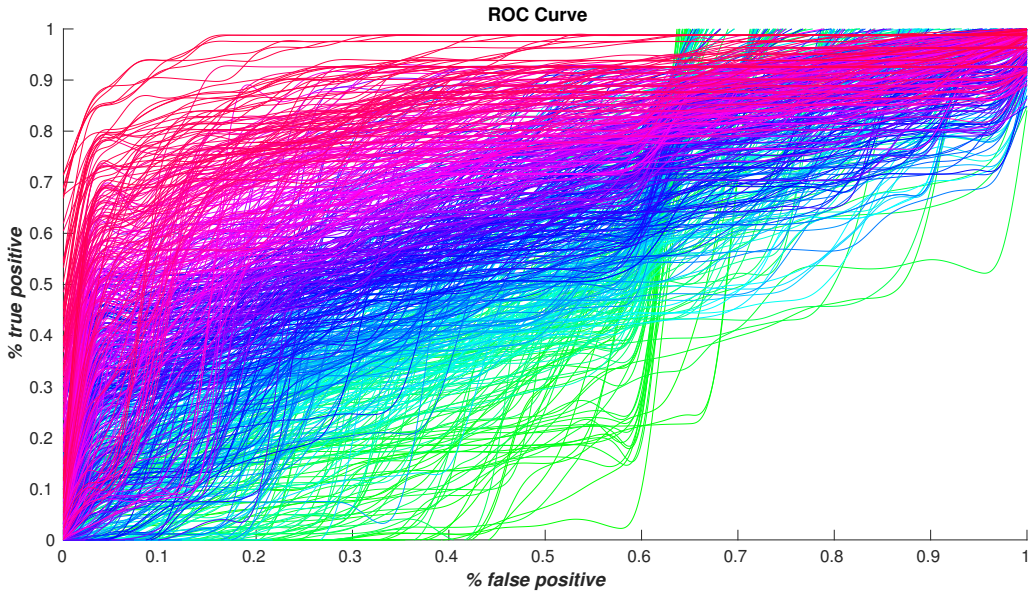

Figure 2: ROC curve computed with the estimated probability of having inflation below the 2% level by using all the variables available in the dataset and a constant in a univariate binomial probit model. Curves closer to the upper-left corner highlight models with a better performance.

2. The ROC curve below the 45-degree-line signals that a researcher should revert the classification scheme (figure 2 green and light-blue lines). This symmetrically flips the curve around the 45-degree-line.

3. The best model attains 100% *TP* and 0% *FP,* which is the upper-left corner of the chart. This point gives the direction toward which the curve should increase to have a more performing model.

A scalar measure of the goodness of the model is the area under the ROC curve. A larger area implies a better model. A commonly used estimator of the ROC area is the non-parametric *AUROC* estimator shown in Equation (6):

$$A\hat{U}ROC = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \left(X_i > Z_j + \frac{1}{2} \left(X_i = Z_j \right) \right), \quad A\hat{U}ROC \in [0.5, 1] \tag{6}$$

Where n_0, n_1 are respectively the zeros and ones according to the correct classification. X_i is the estimated probability corresponding to the correct ones and Z_j corresponding to the

correct zeros. The $AU\hat{R}OC$ ranks models from the one with the largest area to the one with smallest, and it ranges from 0.5 to 1.

5 Methodology

The literature has developed many approaches to deal with large datasets. In general, these techniques either exploit dimensionality reduction such as factor models or employ parameter selection/shrinkage as the lasso estimator. However, the main drawbacks of these approaches are in terms of interpretability. Especially in forecasting, understanding which variables cause a change in the predictions is essential to judge the reliability of the forecast itself. For example, an experienced forecaster might recognize that a variable is released higher than its consensus value and decide to include some form of judgment in his estimates. An example may help to clarify this point; suppose you want to forecast the quarterly y-o-y change in the US GDP. It happens that after the *non-farm-payroll* (NFP) release published by the US Department of labor, your forecasting model revises its prediction by a large number (let's say 0.5 percentage points). It could be the case that you are slightly puzzled by such a significant increase (even if there could be nothing wrong with that). Then, a valuable check can be looking to the value of the *ADP national employment report*, which is released a couple of days before the NFP and extremely correlated with the NFP itself. If you see a significant discrepancy between the two releases, you can put some probability on a measurement error that would be revised in next rounds. Therefore, you may think about lowering your prediction. This process is as plain as described only because of model interpretability. Unfortunately, this excellent characteristic gets lost in more sophisticated approaches. For these reasons, given that this paper focuses on forecasting the inflation probability, I build a direct subset selection approach which allow for the maximum possible degree of interpretability. Also, due to the prediction target, I focus on the out-of-sample performance of the model. The reason is that out-of-sample prediction and in-sample fit are not always strictly linked. The best explanation for this is regarding the usual trade-off between bias and variance. In general, increasing the number of predictors in a model increases the in-sample fit by reducing the bias. However, this comes at the cost of overfitting the data, which often translates into a poor out-of-sample performance.

Having this statement in mind, in the present paper I focus exclusively on the out-of-sample performance of each of the model to evaluate their prediction performance. In particular, I develop a two steps selection procedure; in the first step, I perform a recursive out-of-sample analysis in a univariate regression framework. Meaning, for each variable in the dataset I fit a binomial probit model using Π_t as the dependent variable and that variable as the independent one. In selecting the forecast horizon, having in mind that the primary objective of the paper is an average prediction across short to medium period, I test each variable predictive power at eight different horizons. Being allowed to assess the predictive power of different variables at different horizons is an additional appealing feature to evaluate models using an out-of-sample metric. In fact, it is well-known among forecasters that different variables have different predictive power along different horizons. For example, the yield curve slope, defined as the difference between short to long-term yields, it is a variable which presents these characteristics. The yield curve slope is renewed in the literature for being a powerful predictor of recessions (Estrella and Hardouvelis, 1991; Estrella and Mishkin, 1998). However, its predictive ability is evident only in the forecast at medium to long term. On the contrary, its short-term predicting power is not particularly good. Along with this line, I test different variables at different horizons to select a pool of predictors with a proved forecasting power along different horizons. In what follows I describe the two main steps of the model selection procedure.

5.1 First step

Equation (7) presents the model specification. In the first step, I regress $\Pi_{T+h|T}$ on each i variable in the dataset in a univariate probit model (including a constant). In the notation, $h \equiv [1, 3, 6, 9, 12, 15, 18, 24]$ is the forecast horizon, and $\Pi_{T+h|T}$ is conditioned on the information set available at time T. For each variable, I pre-estimate the model from January 1999 to March 2007; then, I recursively compute the direct forecast up to the end of the sample (March 2017).

$$\Pi_{T+h|T} = G(x_t^{(i)}) + \epsilon_t, \quad \epsilon_t \sim \mathbb{N}(0, \sigma^2), \quad h = [1, 3, 6, 9, 12, 15, 18, 24], \quad i = 1, ..., K \quad (7)$$

Where ϵ_t is an i.i.d. Normally distributed error term with variance restricted to $\sigma^2 = 1$. The

normality assumption on the error term characterize the model as a probit model and attributes to the link function $G(\cdot)$ the interpretation of the cumulative normal distribution function[7]. K is the number of predictors in the dataset. The first step procedure can be summarized as follows:

1. For each variable i, pre-estimating the model coefficients on the period between 1999M1 to 2007M3.

2. For each horizon h, re-estimating the model recursively from 2007M3 to 2017M3 (121 months), increasing the sample size by one data-point for each iteration.

3. For each pair (i, h), assessing the model by comparing the differences between the estimated probability and the true probability using the *Area Under the Receiver Operating Characteristics* (AUROC), the *Mean Absolute Error* (MAE) and the *Root Mean Squared Error* (RMSE).[8]

4. Selecting the two best variables for each horizon h and each criterion (AUROC, MAE, RMSE).

Table 1 shows the best selected variables. Although it shows a complex pattern, it is possible to rationalize the results along with some common lines. First, it is evident that the best predictors for very short horizons are direct measures of inflation (*Consumer Price Indexes*, CPI, and HICP). Second, following MAE and RMSE, the best predictors for short-medium horizons are yields. In particular, the Euro Area interest rate between 3 to 10-years to maturity. It is interesting to notice that also the ten years German and US government bond yields have some predictive power. This fact is likely due to the strong co-movements in the yields among markets. On the contrary, the AUROC predictors are heterogeneous. From six months to one-year-ahead, the best predictors are real variables as the industrial production for Germany and France. However, also intermediate goods and capital for the Euro Area seems to have outstanding predictive power, especially between twelve and eighteen months ahead. Also, it is fascinating to notice that beside real variables, also monetary variables as the M3 aggregate,

[7] For a review of the probit model the reader can refer to Wooldridge (2010) chapter 15.

[8] What I call true probability here is the probability of observing a specific outcome lying in a particular set after the outcome is observed. For example, this implies that when $\pi_t = 1.5\%$, its probability of being in the set $\Pi_t < 2\%$ is one.

Table 1: Results from the first step of the variable selection procedure.

Horiz.	**AUROC**		**MAE**		**RMSE**	
$h = 1$	FR CPI SA	HICP FR	IT CPI SA	HICP IT	IT CPI SA	FR CPI SA
$h = 3$	FR CPI SA	HICP FR	IT CPI SA	HICP IT	FR CPI SA	IT CPI SA
$h = 6$	IP DE	EA7Y	EA7Y	EA3Y	EA7Y	EA3Y
$h = 9$	IP DE	Price trends 12M	EA7Y	EA5Y	EA7Y	EA5Y
$h = 12$	IP FR	Intermediate	EA7Y	EA10Y	EA10Y	EA7Y
$h = 15$	Intermediate	Industrial conf.	EA10Y	DE10Y	EA10Y	US10Y
$h = 18$	M3	Capital	US10Y	EA10Y	US10Y	EA10Y
$h = 24$	DE CPI SA	HICP DE	HICP DE	DE CPI SA	M1	DE CPI SA

Note: the table shows the two best predictors for each horizon. These are selected among the entire dataset using a univariate probit model for forecasting the probability of having inflation below the 2% level. The predictions are evaluated according to three different criteria (AUROC, MAE, RMSE) and the name of the selected variables is reported. The first column under each criteria highlights the best predictor, while the second displays the second best. EA, FR, DE, IT and US are the country abbreviation for Eura Area, France, Germany, Italy and United States. M1 and M3 are the monetary aggregates. IP stands for industrial production. CPI and HICP are price indexes. "Intermediate" and "Capital" are real alctivity measures of intermediate good production and capital. "Industrial conf." is a survey measure of industrial confidence. Finally, the country abbreviations reported beside the number of years as "DE10Y" stand for benchmark yields with a particular maturity.

surveys as the industrial confidence indicators and expectations show an excellent forecasting power. Finally, for longer horizons, even if the monetary aggregate M1 shows some predictive power, the best predictors are some direct inflation measures as for shorter horizons. It is interesting to notice that for shorter horizons, the best predictors were the inflation measures of France and Italy, while for longer horizons, the German inflation measure dominates. For this last finding, I do not have a clear answer. However, I suspect that the reason could be related to the way in which the EA HICP is computed in terms of the disaggregated national price indexes.

Table 2 summarizes the 20 unique predictors delivered by the first selection step. Not surprisingly, the selected predictors coincide with variables considered the main determinants of inflation by established economic relationships. For illustrative purposes, consider a small-scale *New Keynesian model* as described by Equation (8) to (10) [9]. The first relation is the *New Keynesian IS curve* (NKIS) while the second describes the *New Keynesian Phillips Curve* (NKPC). The third relation is the Taylor rule introduced in section 2 and reported here for convenience.

$$\hat{y}_t = \mathbb{E}_t\left[\hat{y}_{t+1}\right] + \frac{1}{\sigma}\left(i_t - \mathbb{E}_t\left[\pi_{t+1}\right]\right) + \epsilon_t \tag{8}$$

[9] See for example Galí (2015).

Table 2: Results from the first step of the variable selection procedure. All criteria and horizons.

Price	Interest rate	Real	Monetary	Survey
HICP DE	EA3Y	Intermediate goods	M1	Industrial confidence
HICP FR	EA5Y	Capital	M3	Price trends 12M
HICP IT	EA7Y	IP FR		
DE CPI SA	EA10Y	IP DE		
FR CPI SA	DE10Y			
IT CPI SA	US10Y			

Note: the table shows the best predictors for all the horizons. These are selected among the entire dataset using a univariate probit model for forecasting the probability of having inflation below the 2% level. The predictions are evaluated according to three different criteria (AUROC, MAE, RMSE) and the name of the selected variables is reported. Each variable is reported only once, and it is allocated in one of the five macro-categories (Price, Interest rate, Real, Monetary, Survey). EA, FR, DE, IT and US are the country abbreviation for Eura Area, France, Germany, Italy and United States. M1 and M3 are the monetary aggregates. IP stands for industrial production. CPI and HICP are price indexes. "Intermediate" and "Capital" are real alctivity measures of intermediate good production and capital. "Industrial conf." is a survey measure of industrial confidence. Finally, the country abbreviations reported beside the number of years as "DE10Y" stand for benchmark yields with a particular maturity.

$$\pi_t = \beta \mathbb{E}_t \left[\pi_{t+1}\right] + \kappa \hat{y}_t + \eta_t \tag{9}$$

$$i_t = \phi_\pi (\mathbb{E}_t \pi_{t+1} - \pi_t^*) + \varepsilon_t \tag{10}$$

Where $\hat{y}_t$ is the *output-gap*, which is the difference between current output (y_t) and output at full employment (y_t^n), also called *potential output*. π_t is the inflation at time t, $\mathbb{E}_t \left[\pi_{t+1}\right]$ is the expected inflation given the information set at time t, and i_t is a measure of monetary policy stance controlled by the central bank. ϵ_t, η_t, and ε_t are random variables referred to *technology*, *price markup*, and *monetary policy shock*. For our purpose, the most relevant equation is the NKPC, which is a relation between current and expected inflation. If one believes in this simplified structure, selecting the best predictors among many variables should indeed return some proxy of the variables entering in the entire NK model. Table 2 shows precisely this point. First, the NKPC is a function of the expected inflation and the output gap. Among the selected variables, the expected price trend in the next twelve months can be considered a good proxy for the inflation expectations. Also, the NKPC links the price variation to the output-gap (which is a latent variable). For quarterly data, the best proxy for the output gap is the difference between

current and potential GDP level. However, when a researcher deals with monthly variables, the best proxy for the output-gap is indeed constructed from industrial production. Indeed, from the variable selection procedure, I retrieve exactly this variable for both France and Germany. Nevertheless, in the NK model, the output-gap is a function of the real interest rate, which is the difference between nominal interest rate and expected inflation. It is true that the interest rate included in NK models refers to the interest rate directly under the control of the central bank; however, due to strong co-movements among yields, government bonds can be considered good proxies for that variable. The selection procedure highlighted includes yields at different maturities among the best predictors for short to medium horizons. Finally, also the money supply is reported among the best predictors. This variable is often used in theoretical economic models as an alternative instrument under the control of the central bank and can readily fit into an NK model.

This strong linkage between the selected variables and three established economic relationships builds confidence in the procedure and benefits from the interpretation point of view. Finally, it is valuable to notice that similar findings are common in the inflation forecasting literature. In fact, various forms of the NKPC are often estimated and used as a proper reduced-form model to forecast inflation. In the second step, I employ the twenty selected variables as an input for the model selection procedure as described in the next paragraph.

5.2 Second step

In the second step, I perform a procedure similar to *best subset selection*. I fit a separate probit model to all possible combinations of the $K_2 = 20$ predictors selected in the first stage. Having twenty different variables implies that the number of possible combinations is extremely high. Then, I restrict the number of maximum regressors in the model to $K_1 = 10$. This is the set which contains the largest number of combinations. Equation (11) shows the total number of possible combinations:

$$C = \sum_{k=1}^{K_1} \binom{K_2}{k} = 2^{K_2} - \sum_{k=K_1+1}^{K_2} \binom{K_2}{k} - 1 \tag{11}$$

As $K_1 = 10$ and $K_2 = 20$ in our setting, this leads to $C = 616,665$ models to estimate. Moreover, given the recursive structure of the out-of-sample exercise, to understand the total number of estimated models, C has to be multiplied by the number of data points by which the model is re-estimated. Those are $T^{out} = 121$. Also, C has to be multiplied by the number of horizons for which the models are re-estimated ($H = 8$). This process leads to the estimation of $M \approx 600,000,000$ models. Finally, for each estimated model, I compute an alternative model augmented with a common factor extracted from the complete dataset. The common factor is estimated non-parametrically via principal component. In particular, I estimate the factor from the eigenvector corresponding to the largest eigenvalue of the variance-covariance matrix of the demeaned dataset. The reason to have the augmented models is that many authors, starting from Stock and Watson (2002) have shown the predictive ability of common factors[10]. Introducing this twist, from one side it helps to explore relevant information that could have been left out of the first step procedure (for example, due to the presence of strong predictors). From the other, it massively increases the computational burden. In fact, adding a factor-augmented counterpart doubles the number of models to estimate ($M \approx 1.2bn$). Also, to avoid including information from the future, the principal component is recursively estimated each time a data point is added. To deal with such complexity, I write the entire code in Julia Language (Bezanson et al., 2017), and I perform estimation parallelizing the code on an octa core processor. Julia is a modern and flexible open source language, which easily allows to perform parallel computing and to deal with computationally intense problems. Table 3 shows the time employed for the combinations of each variable group, including the time for the out-of-sample performance and the principal component analysis.

I evaluate each model for each horizon according to the AUROC, MAE, and RMSE. Also, to have a benchmark for the comparison, I build a *naive model* fitting only the first lag of the EA HICP and compare the performance of each model against this one. The score is reported as a ratio between the two models. For the AUROC, a ratio larger than one implies that the selected model outperforms the naive. For the MAE and RMSE, the opposite is true. Finally, I select only models with maximum AUROC and minimum MAE and RMSE for each horizon.

[10] Stock and Watson (2016) survey this class of models and their use in macroeconomics.

Table 3: Number of variables, combinations and computational time for each combination in the second step.

#Variables	#Combinations	Time
1	$20 \times H \times T^{out} \times 2$	13s
2	$190 \times H \times T^{out} \times 2$	117s
3	$1,140 \times H \times T^{out} \times 2$	12m
4	$4,845 \times H \times T^{out} \times 2$	48m
5	$15,504 \times H \times T^{out} \times 2$	2h34m
6	$38,760 \times H \times T^{out} \times 2$	6h28m
7	$77,520 \times H \times T^{out} \times 2$	13h02m
8	$125,970 \times H \times T^{out} \times 2$	19h20m
9	$167,960 \times H \times T^{out} \times 2$	28h30m
10	$184,756 \times H \times T^{out} \times 2$	31h20m
Total	616,665	$\approx$ 100h

*Note: the table shows the number of variables used as regressors in a multivariate probit model (univariate when **#Variables** is equal to one) for forecasting the probability of having inflation below the 2% level. The total number of variables used is twenty and were selected in a univariate framework in a previous step. When the **#Variables** is equal to K, there are $\binom{20}{K}$ possible combinations. H is equal to eight and $T^{out} = 121$. The number of combinations is multiplied by two to account for the alternative models including the first principal component of the entire dataset. The table also shows the amount of time employed by the Julia code for each particular number of combinations.*

The second stage of the selection process returns a set of $3H = 24$ models. The best-selected models are reported in Table 4 to 6. Depending on the criterion chosen, the results are mostly heterogeneous, both in the number of selected variables and in the inclusion of a common factor. Also, according to different criteria, the difference between the naive model and the selected models is weaker or stronger. However, some common characteristics are worth to highlight. First, according to all criteria, the selected models are always able to outperform the naive model. However, for shorter horizons, the naive model is more difficult to beat. For longer horizons, the selected models perform much better. Secondly, for some specific horizons, the three criteria agree on both the number and the variables to include. Two clear example are the horizons $h = 6$ and $h = 18$. Thirdly, The AUROC and the RMSE are more parsimonious criteria in terms of the number of selected variables, while the MAE is the least. Forty, on average, it seems that all models use predictors coming from different classes, implying that those can bring different information useful in improving the prediction. Figure 3 shows the score as the ratio between each of the selected model against the naive model. The AUROC is reported in terms of reciprocal to enhance comparability. A score lower than one implies that

Table 4: Results from the second step of the variable selection procedure. AUROC criteria, all horizons.

	$h=1$	$h=3$	$h=6$	$h=9$	$h=12$	$h=15$	$h=18$	$h=24$
AUROC	1.02	1.05	1.19	1.63	1.64	1.81	1.54	1.51
Factor	0	1	1	1	1	0	1	0
#Var	5	9	8	10	8	7	8	8
	Capital	M1	Capital	Inter.	Inter.	Ind.Conf.	Capital	M3
	Ind.Conf.	M3	M3	Capital	Capital	M1	M1	HICP DE
	IP FR	IP DE	IP DE	Ind.Conf.	Ind. Conf.	M3	M3	HICP FR
	IT CPI SA	IP FR	IP FR	M3	M3	EA10Y	IP DE	HICP IT
	FR CPI SA	US10Y	HICP IT	HICP DE	IP FR	IT CPI SA	IP FR	DE10YT
		DE10YT	EA7Y	US10Y	US10Y	FR CPI SA	US10Y	EA3Y
		EA3Y	DE CPI SA	DE10YT	DE10Y	PRICE 12M	EA10Y	EA5Y
		EA7Y	PRICE 12M	EA10Y	EA10Y		FR CPI SA	FR CPI SA
		EA10Y		DE CPI SA				
				PRICE 12M				

Note: the table shows the best model for each horizon h, selected among $\sum_{k=1}^{10}\binom{20}{k}$ models. The variables are used as regressors in a multivariate probit model for forecasting the probability of having inflation below the 2% level according to the AUROC criteria. The table also shows the score of the models reported as a ratio between the selected model and the naive. For the AUROC, a ratio larger than one implies that the selected model outperforms the naive. For the MAE and RMSE the opposite is true. Also, the table highlights the number of variables in the model and whether the forecast improves including the first principal component of the original dataset ("Factor" equal to one implies that including the factor enhances the prediction).

Table 5: Results from the second step of the variable selection procedure. MAE criteria, all horizons.

	$h=1$	$h=3$	$h=6$	$h=9$	$h=12$	$h=15$	$h=18$	$h=24$
MAE	0.25	0.39	0.17	0.11	0.09	0.15	0.12	0.02
Factor	1	1	1	1	1	1	1	1
#Var	10	10	8	10	10	10	8	10
	Inter.	Inter.	Capital	Inter.	Inter.	Inter.	Capital	Capital
	Capital	Capital	M3	M1	Ind.Conf.	Ind.Conf.	M1	M1
	M3	Ind.Conf.	IP DE	M3	M3	M1	M3	M3
	HICP DE	M3	IP FR	IP FR	IP DE	M3	IP DE	IP DE
	DE10YT	HICP FR	HICP IT	HICP DE	IP FR	HICP DE	IP FR	IP FR
	EA3Y	HICP IT	EA7Y	HICP FR	HICP DE	HICP FR	US10Y	EA5Y
	EA5Y	EA3Y	DE CPI SA	HICP IT	DE10YT	US10Y	EA10Y	EA7Y
	EA7Y	EA5Y	PRICE 12M	EA3Y	IT CPI SA	EA7Y	FR CPI SA	EA10Y
	IT CPI SA	EA7Y		EA7Y	DE CPI SA	EA10Y		DE CPI SA
	DE CPI SA	DE CPI SA		PRICE 12M	FR CPI SA	FR CPI SA		FR CPI SA

Note: the table shows the best model for each horizon h, selected among $\sum_{k=1}^{10}\binom{20}{k}$ models. The variables are used as regressors in a multivariate probit model for forecasting the probability of having inflation below the 2% level according to the MAE criteria. The table also shows the score of the models reported as a ratio between the selected model and the naive. For the AUROC, a ratio larger than one implies that the selected model outperforms the naive. For the MAE and RMSE the opposite is true. Also, the table highlights the number of variables in the model and whether the forecast improves including the first principal component of the original dataset ("Factor"equal to one implies that including the factor enhances the prediction).

Table 6: Results from the second step of the variable selection procedure. RMSE criteria, all horizons.

	$h=1$	$h=3$	$h=6$	$h=9$	$h=12$	$h=15$	$h=18$	$h=24$
RMSE	0.64	0.75	0.35	0.31	0.37	0.39	0.21	0.16
Factor	0	0	1	1	1	0	1	1
#Var	8	4	8	8	10	6	8	10
	Ind.Conf.	Capital	Capital	Inter.	Inter.	Capital	Capital	Capital
	M1	Ind.Conf.	M3	M1	Ind.Conf.	IP DE	M1	M1
	M3	EA7Y	IP DE	M3	M3	IP FR	M3	M3
	IP FR	FR CPI SA	IP FR	HICP IT	IP DE	HICP DE	IP DE	IP DE
	US10Y		HICP IT	DE10YT	IP FR	EA10Y	IP FR	IP FR
	EA5Y		EA7Y	EA10Y	HICP DE	PRICE 12M	US10Y	EA5Y
	EA10Y		DE CPI SA	FR CPI SA	DE10YT		EA10Y	EA7Y
	FR CPI SA		PRICE 12M	PRICE 12M	IT CPI SA		FR CPI SA	EA10Y
					DE CPI SA			DE CPI SA
					FR CPI SA			FR CPI SA

Note: the table shows the best model for each horizon h, selected among $\sum_{k=1}^{10}\binom{20}{k}$ *models. The variables are used as regressors in a multivariate probit model for forecasting the probability of having inflation below the 2% level according to the RMSE criteria. The table also shows the score of the models reported as a ratio between the selected model and the naive. For the AUROC, a ratio larger than one implies that the selected model outperforms the naive. For the MAE and RMSE the opposite is true. Also, the table highlights the number of variables in the model and whether the forecast improves including the first principal component of the original dataset ("Factor" equal to one implies that including the factor enhances the prediction).*

the selected model outperforms the naive. As expected, the naive model is performing better in the very short horizons. For $h = 1$ the best model selected according to the AUROC is only slightly better. However, since $h = 3$ the models selected using different criteria considerably outperform the naive model and present massive increases until $h = 9$. Then, depending on the criteria, the curve is stable or slightly increasing/decreasing.

6 Results

In this section, I analyze the in-sample and out-of-sample performance of the models selected in the previous section. The in-sample performance gives a general framework to assess the overall performance of the models. The reason is that, given the lack of a long time series for the EA, there are only a few separate periods in the out-of-sample exercise in which inflation is below the 2% level. However, as I am mainly interested in forecasting, after assessing the in-sample performance, I evaluate the out-of-sample performance of the selected models, and predict in a true out-of-sample the probability of having low inflation from the end of the sample to March 2019. Finally, in the next section, I build an index averaging all the forecasts from one month to two years and assess the index prediction against the ECB SPF.

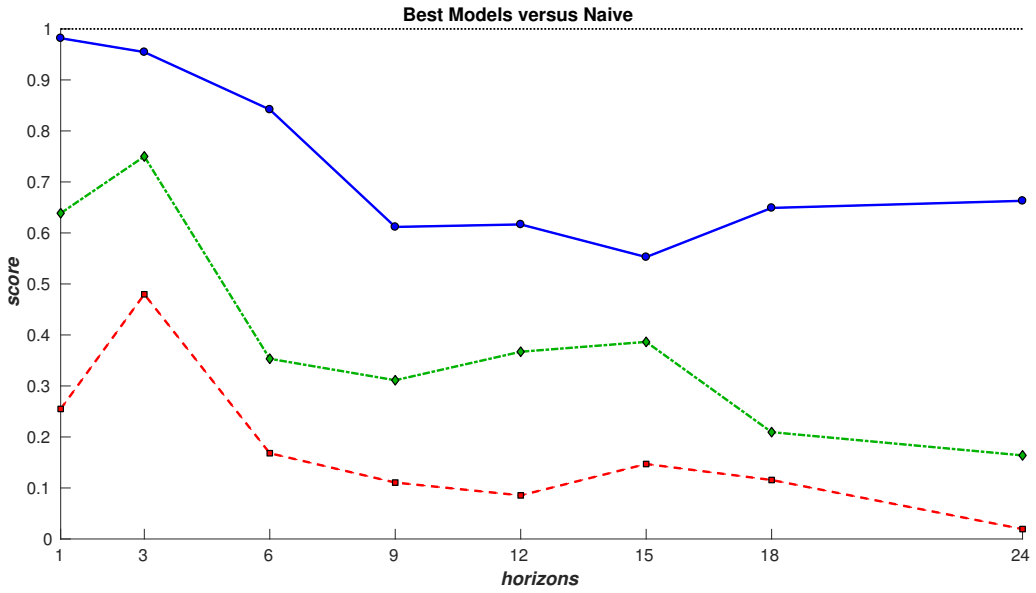

Figure 3: Score of the models reported as a ratio between the selected model and the naive according to the AUROC reciprocal (solid blue line), MAE (dash-dot green line) and RMSE (dashed red line). For all criteria, a ratio below one implies that the selected model outperforms the naive.

6.1 In-sample analysis

I start analysing the in-sample results. I focus mainly on the probability that inflation is below the 2% level. The reason is that the downside risk seems the major concern in the EA. However, given that I am forecasting the whole density of the process, the upside risk can always be computed as the complement of the downside probability. Figures 4 and 5 show the in-sample fit of the models. For each horizon, I plot the three best models selected according to AUROC, MAE, and RMSE. The gray bars represent periods of inflation lower than 2%. Except for very short periods, the first part of the sample is mainly characterized by having inflation above the threshold. However, the rapid changes in the regimes creates many false signals in the estimated probabilities. This is evident, especially at shorter horizons. Starting from 2008 the series is characterized by prolonged periods of inflation above/below the threshold, which substantially reduce false signals.

Overall, the models have a satisfactory in-sample prediction ability. Also, the models chosen with the three criteria are very similar, and on many occasions, the estimated probabilities overlap as for $h = 6$ and $h = 18$.

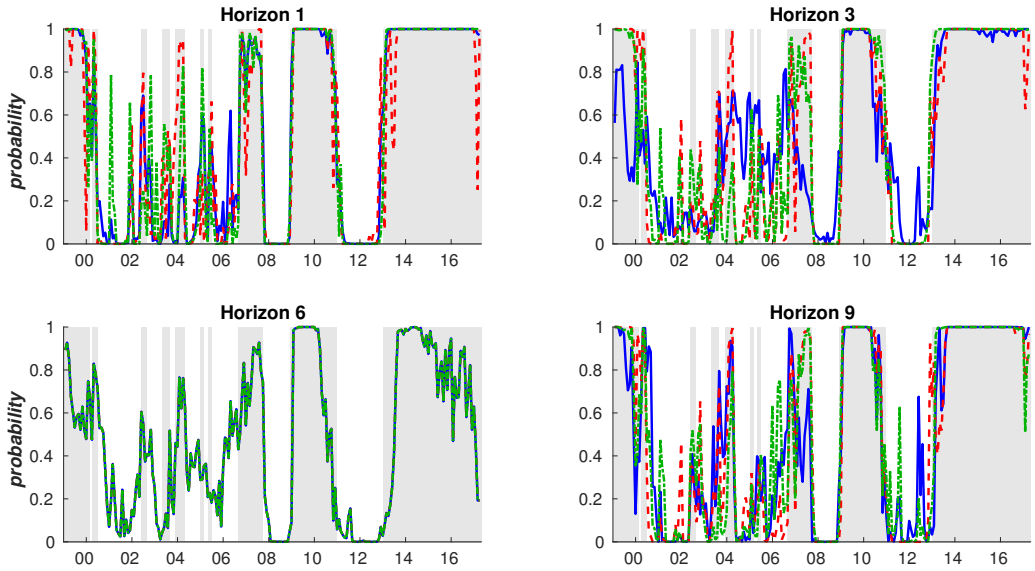

Figure 4: In-sample-model-fit for horizons $h = 1$ to $h = 9$. Each panel shows the best model selected using AUROC (solid blue lines), MAE (dash-dot green lines) and RMSE (dotted red lines). The gray bars represent periods with inflation below the 2% level.

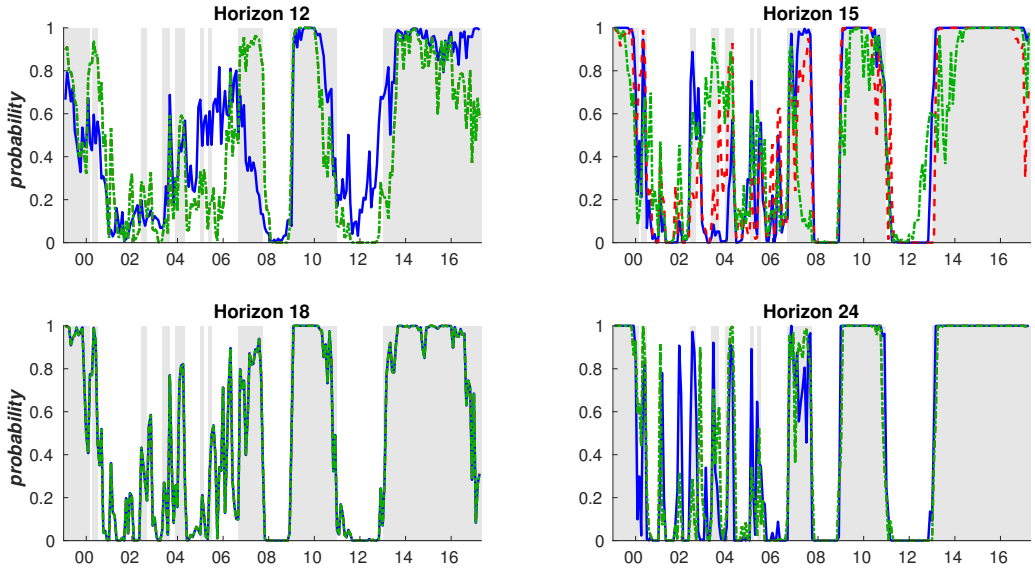

Figure 5: In-sample-model-fit for horizons $h = 12$ to $h = 24$. Each panel shows the best model selected using AUROC (solid blue lines), MAE (dash-dot green lines) and RMSE (dotted red lines). The gray bars represent periods with inflation below the 2% level.

6.2 Out-of-sample analysis

After assessing the in-sample fit of the selected set of models, I evaluate their out-of-sample predictions. For each set of variables, I pre-estimate the model from January 1999 to March

2007 and compute a direct forecast up to the end of the sample (March 2017). Figure 6 and 7 show the out-of-sample estimates of the different models for all the horizons. The sample-period used for the estimation presents two extended periods of inflation below the 2% level divided by an interval of inflation above or equal 2%. These periods partially overlap with the great recession (inflation below the target starts in December 2008) and to the post-European debt crisis (January 2013). From visual inspection, the out-of-sample predictions do not show

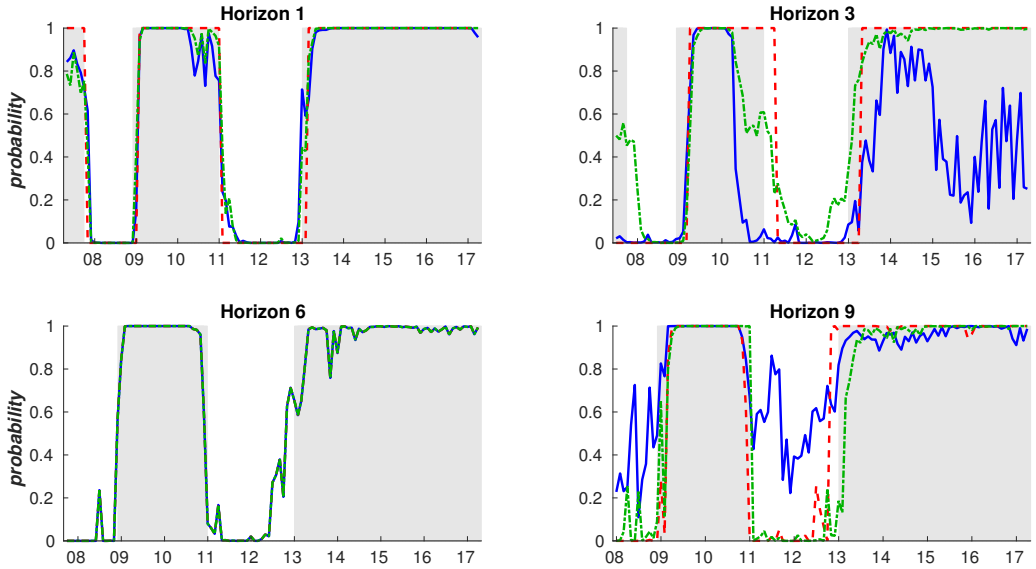

Figure 6: Out-of-sample forecast for horizons $h = 1$ to $h = 9$. Each panel shows the best model selected using AUROC (solid blue lines), MAE (dash-dot green lines) and RMSE (dotted red lines). The gray bars represent periods with inflation below the 2% level.

severe lacks, and the overall fit is pretty good. However, as usual in out-of-sample forecasting, the goodness of the projections is a function of the horizons, and overall results are heterogeneous. At one-step-ahead, the three models have a performance which is exceptionally close. The models do an excellent job in timely catching the turning points in the inflation probability. At $h = 3$, the RMSE is the best model, succeeding in capturing the first-period of low inflation and start rising slightly in advance with respect to the third one. The MAE model is sligtly delayed with respect to the turning points, while the AUROC model delivers an extremely meager job. At horizon six and eighteen the three criteria have selected the same variables among all possible combinations. This choice implies that the selection is exceptionally robust across the different loss functions. For what concern the remaining horizons, the MAE and RMSE mod-

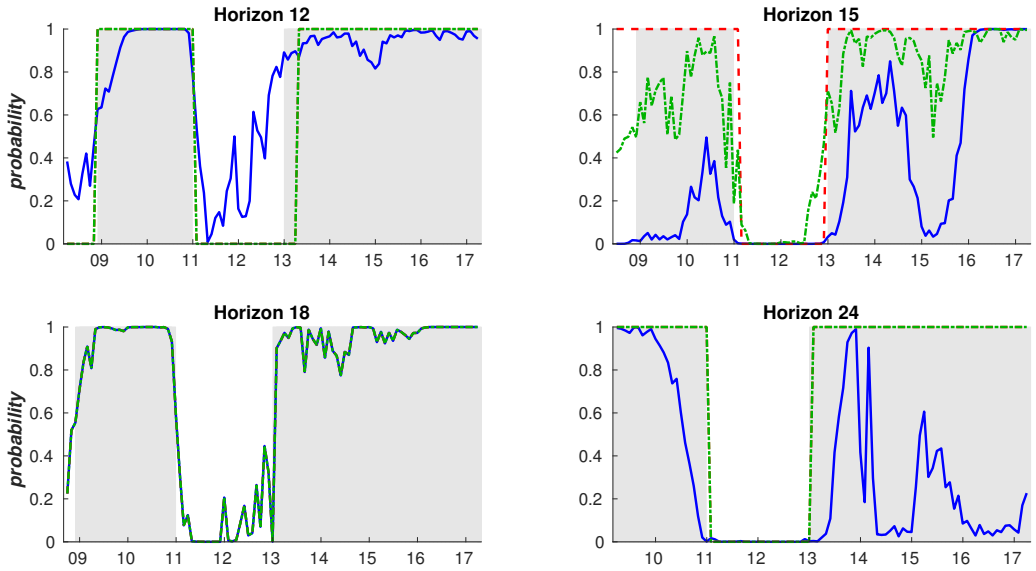

Figure 7: Out-of-sample forecast for horizons $h = 12$ to $h = 24$. Each panel shows the best model selected using AUROC (solid blue lines), MAE (dash-dot green lines) and RMSE (dotted red lines). The gray bars represent periods with inflation below the 2% level.

els are the best performers, and in many occasions, the variables selected by the two coincide. From the other side, the AUROC is not delivering, especially at horizons $h = 15$ and $h = 24$. The main problem seems to be the high autocorrelation in the estimated probabilities which miss the turning points. At the opposite, the other models do a solid job in timely capturing the switch from one state to the other.

Table 7 summarizes the results of all the models, criteria, and horizons with respect to the naive model. The three panels show the results for the three criteria. AUROC, MAE, and RMSE highlights the set of models chosen to maximize/minimizing these criteria. Therefore, by definition, for the AUROC panel, the set of models which attains the best results is the AUROC column (blue), for the MAE panels is the MAE column (green) and for the RMSE panels is the RMSE column (red). It is interesting to notice that while MAE and RMSE models are very close to the AUROC models in terms of the score, the opposite is not true, and AUROC chosen models are very distant from MAE and RMSE models. Oddly, at $h = 3$, according to both the MAE and RMSE criteria the AUROC model is outperformed even from the naive model. The same happens to very short horizons for the MAE in terms of AUROC. However,

Table 7: Results from the out-of-sample forecast. All models, criteria, and horizons.

	AUROC			MAE			RMSE		
Horizon	**AUROC**	**MAE**	**RMSE**	**AUROC**	**MAE**	**RMSE**	**AUROC**	**MAE**	**RMSE**
$h = 1$	1.02	0.99	1.02	0.46	0.25	0.41	0.68	0.73	0.64
$h = 3$	1.05	0.96	1.04	1.41	0.48	0.59	1.43	0.99	0.75
$h = 6$	1.19	1.19	1.19	0.17	0.17	0.17	0.35	0.35	0.35
$h = 9$	1.63	1.59	1.6	0.39	0.11	0.11	0.58	0.38	0.31
$h = 12$	1.62	1.57	1.57	0.3	0.09	0.09	0.46	0.37	0.37
$h = 15$	1.81	1.55	1.79	0.69	0.15	0.28	0.9	0.46	0.39
$h = 18$	1.54	1.54	1.54	0.12	0.12	0.12	0.21	0.21	0.21
$h = 24$	1.51	1.48	1.48	0.86	0.02	0.02	1.0	0.16	0.16

Note: the table shows the scores for each model selected according to AUROC, MAE, and RMSE and for each horizon h. The total number of selected models is 24, and each selected model is evaluated according to all the three possible criteria. The reported number is the ratio of the score of the selected model over the score of a univariate probit model which uses as regressor the first lag of the HICP inflation (naive model). For the AUROC, a score higher than one implies that the selected model outperforms the naive model, while for MAE and RMSE the opposite is true.

Table 8: Best model predictions at different horizons.

Date	**AUROC**	**MAE**	**RMSE**
Apr 2017	0.87	0.99	0.99
Jun 2017	0.99	0.95	0.95
Sep 2017	0.98	0.98	0.98
Dec 2017	0.99	0.98	0.97
Mar 2018	0.99	0.99	0.99
Jun 2018	0.99	0.99	0.99
Sep 2018	0.99	0.99	0.99
Mar 2019	0.99	0.99	0.99

Note: the table shows the forecasted probability of having inflation below the 2% level in the next 24 months for each model selected according to AUROC, MAE, and RMSE.

the distance between the two models is much smaller (one to three percentage points). As a final exercise, I use all estimated models to predict the probability of having low inflation in a true out of sample forecast. I predict all the horizons of interest starting from the end of the sample. I use each model to forecast only the single horizon for which the model is tailored. Table 8 shows the predicted probability of having low inflation in the 24 months after the model is calibrated. All the model predictions are extremely close and signal that in the next future the inflation upside risk is not a concern.

7 Deflationary pressure index

As a final exercise, in this section, I create an index to signal the probability of having low inflation in the medium run. I call the index *Deflationary Pressure Index* (DPI) given it signals the average probability of moving toward an undesired inflation territory from the downside. This index can supply a valid in-house alternative to the SPF probability forecasts and help in dealing with the generic medium-term horizon considered by the European Central Bank to undertake policy actions. Equation 12 shows the DPI. The index is a simple average over the AUROC, MAE and RMSE best model forecasts $\hat{\Pi}^{*(C)}_{T+h|T}$ at horizons $h = [1, 3, 6, 9, 12, 15, 18, 24]$. In particular, the index is built having in mind a researcher that at time T forecasts the horizons from one to twenty-four months ahead, and averages "horizontally" the predictions. The value of the index is recorded at time T. In this way, each point of the DPI represents the probability of having inflation below the 2% level in the next two years.

$$DPI^{(C)}_t = \frac{1}{H}\sum_h \hat{\Pi}^{*(C)}_{T+h|T}, \quad C = \{AUROC, MAE, RMSE\} \tag{12}$$

I avoid averaging across criteria in order to have three different indexes to compare. Figure 8 shows the constructed deflationary pressure indexes against the periods in which inflation is below the 2% level. The chart presents two prominent features; first, the movements in the three indexes are incredibly close. This characteristic is an excellent signal given that the three were created using three different loss functions. Secondly, the indexes, accordingly to the true out-of-sample estimates, do not signal any upside risk for inflation in the next two years.

Two main features are worth to be mentioned. First, all the three indexes seem to have good predicting power. This feature is evident in the period between 2009 and 2013, as the indexes start moving in the correct direction before the inflation switch in the regime. Secondly, as all the three signal a very high probability of having inflation below the 2% level, we decide to investigate whether other measures support this finding. For this reason, in the next section, I compare the DPI against a probability measure constructed starting from the SPF probabilities.

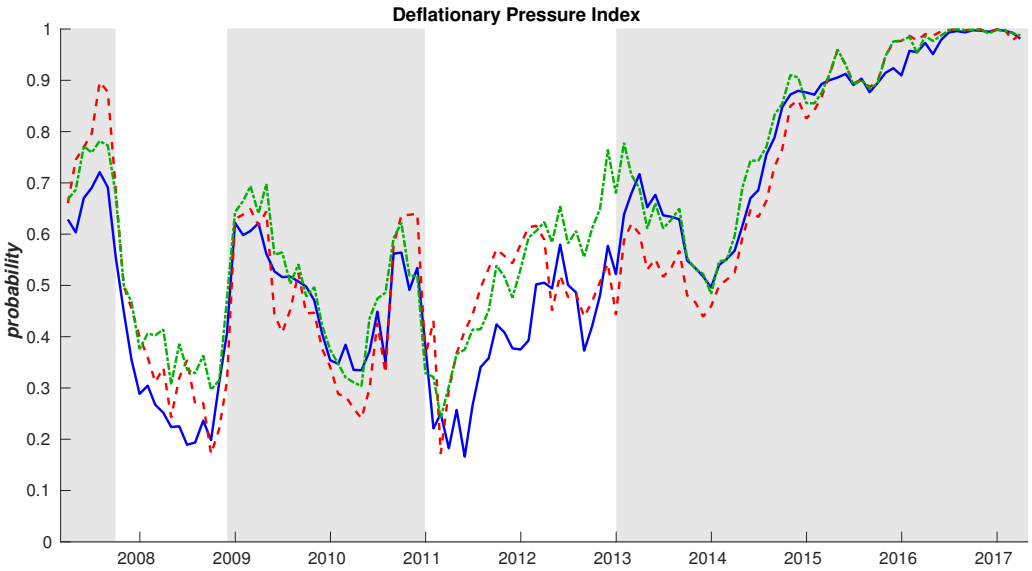

Figure 8: Deflationary pressure index. This is the simple average of the out-of-sample forecasts for all horizons. AUROC (solid blue line), MAE (dash-dot green line), RMSE (dotted red line) highlight the index made from models selected independently with these three criteria. The gray bars represent periods with inflation smaller than 2%.

7.1 Survey of Professional Forecasters: a comparison

In this section, I compare the DPI against the ECB *Survey of Professional Forecasters*. Starting from December 2000 the SPF is regularly collected every quarter from surveying more than 80 professional forecasters. In a typical survey, forecasters are required to express their point forecasts for inflation (as well as GDP growth and unemployment) over specific time horizons. Also, they are asked to provide their probabilities for different inflation outcomes. For example, they are asked to report the likelihood that the year-on-year HICP inflation is below, in between or above certain thresholds. The thresholds range from -1% to 4% stepping by 0.4%, for a total of 12 bins. Probabilities have to sum to one, and as the final forecast measure, the average of all the forecasts among forecasters is used.

To create an SPF index comparable to the DPI, I construct a measure by cumulating the probabilities of having inflation below the 2% level between -1% and 2% in the next 24 months. Also, given that the SPF is quarterly collected, the comparison involves mixed frequency. Figure 9 shows the quarterly SPF median survey forecast (yellow diamonds) against the monthly

deflationary pressure index. From the figure, it is evident that the SPF is strongly autocorrelated

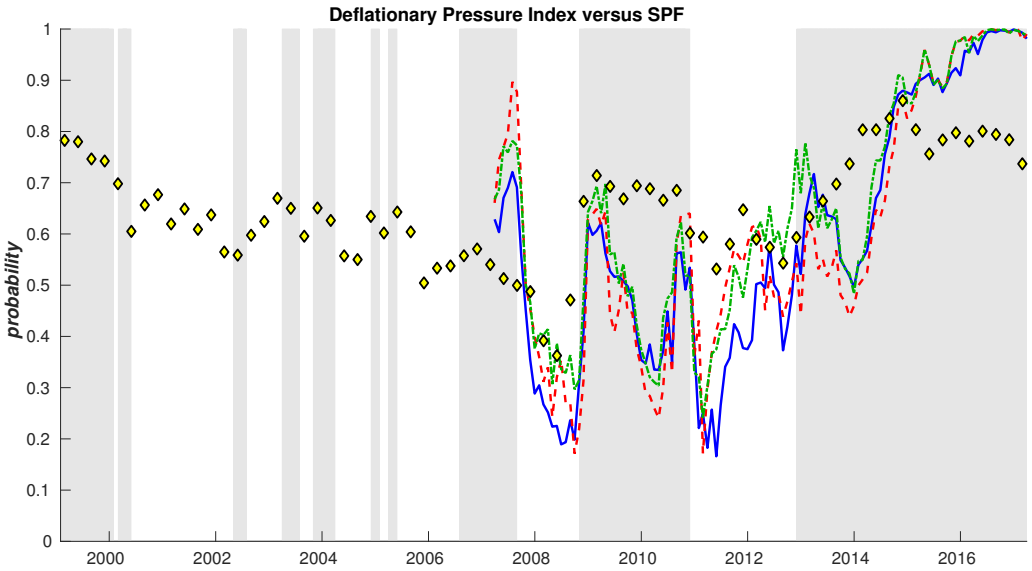

Figure 9: Deflationary pressure index. This is the simple average of the out-of-sample forecasts for all horizons. The indexes are plotted against the ECB Survey of Professional Forecasters (SPF) 24 months ahead predictions (yellow diamonds). AUROC (solid blue line), MAE (dash-dot green line), RMSE (dotted red line) highlight the index made from models selected independently with these three criteria. The gray bars represent periods with inflation smaller than 2%.

and moves slower than the DPI. This is also the reason why they often miss to catch the turning points. The autocorrelation is especially evident in the last two transition periods of inflation above/below the 2% threshold level. However, apart from this characteristics, the two measures are extremely similar. This feature is true especially between 2013 and 2017. In the final part of the series they both present increasing probabilities, however at the end of the sample the two measure slightly diverges, with the SPF measure declining to 75%. However, both models predict a very high probability of having inflation below the 2% level in the two years following the end of the sample.

8 Discussion

The construction of an index to interpret and forecast business cycle conditions is a tale of a long tradition in economics. The seminal paper by Mitchell and Burns (1938) has spawned voluminous literature on coincident and leading indicators, and many influential articles have followed

(Stock and Watson, 1989, 2002). In this paper, I partially build on this discussion. However, there are some significant differences between the DPI and the leading indexes typically built in the literature. First, these indexes are traditionally created by extracting unobserved factors from a pool of variables. This procedure is very different from the methods I am using in this paper. Secondly, as the article focuses on predicting probabilities rather than points, I include some mild forms of non-linearities in the model specification. Therefore, extracting factors with standard techniques is harder than in a simple linear models. The reason is that in a hypothetical state-space model the measurement equation would be non-linear, implying that a modified version of the Kalman filter or the particle filter has to be used to derive the likelihood. A second alternative would be using a two-step procedure to first extract factors and then plug them as predictors in the probability model. However, both these processes would make interpreting changes in the forecasts much harder. The third main difference is strictly connected to the particular problem addressed by the paper. As described, central banks targeting inflation have to rely on forecasts, since monetary policy affects output and prices with a lag. Also, they have a general medium-term orientation. Thus, in creating the DPI, I also average model predictions along horizons. From one side, this procedure creates an additional complication, since a forecaster has to select the best model to use for each horizon. From the other, it generates the condition to undertake a combinatoric procedure to understand which variables are the best predictors at different horizons.

This methodology is new in this field, and because there are not many studies on forecasting inflation probabilities, there is no established benchmark for comparisons. An article with a similar outcome is the one presenting the St. Louis Federal Reserve *Price Pressure Measure* (Jackson et al., 2015). This series measures the probability that the expected *personal consumption expenditures price index* (PCEP) inflation rate over the next 12 months will exceed the 2.5% level. However, in the paper, the authors calibrate a factor model to predict the y-o-y change in inflation, then use the extracted factors to predict the inflation probabilities. Most of the work is on choosing how many factors to use in the prediction of the inflation point forecast. The authors do not perform a direct calibration of the probability model. Also, the same model is used to predict all the different horizons. Therefore, the price pressure measure is constructed

indirectly from the starting model.

This methodology is new in this field, and because there are not many studies on forecasting inflation probabilities, there is no established benchmark for comparisons. An article with a similar outcome is the one presenting the St. Louis Federal Reserve *Price Pressure Measure* (Jackson et al., 2015). This series measures the probability that the expected *personal consumption expenditures price index* (PCEP) inflation rate over the next 12 months will exceed the 2.5% level. However, in the paper, the authors calibrate a factor model to predict the y-o-y change in inflation, then use the extracted factors to predict the inflation probabilities. Most of the work is on choosing how many factors to use in the prediction of the inflation point forecast. The authors do not perform a direct calibration of the probability model. Also, the same model is used to predict all the different horizons. Therefore, the price pressure measure is constructed indirectly from the starting model.

On the contrary, the present paper gives prominent importance to each component used in constructing the index. This result is accomplished by selecting each predictor and model at "micro-level" using the combinatoric approach. Also, appropriate attention is devoted to the choice of the loss function, focusing on different metrics. Naturally, these two components do not come without any drawbacks. First, there is a risk in employing a combinatoric approach blindly. In particular, a forecaster could "overfit the out-of-sample". The issue is strictly connected to the variables to combine when selecting the best model. When the number starts increasing wildly, it is possible to fit the out-of-sample shape of the data perfectly. In turn, this feature can create the same issues related to the "in-sample overfitting". Therefore, a forecaster should always be careful in setting-up the combinatoric exercise. Secondly, although I evaluate models using different loss functions, I restrict the choice among symmetric ones. However, predicting turning points as the beginning of a recession would probably benefit from the use of an asymmetric loss function. The reason behind this statement is extremely naive and is that a forecaster would always prefer a model which predicts a recession before it happens. Therefore, a model which forecast a probability equal to one approaching a recession should be preferred to one which predicts zero recession probability when the recession is starting. Symmetric loss functions disregard this simple behavior, and future research is needed to address this topic.

9 Conclusion

Central Banks worldwide target an optimal inflation level to maintain price stability. In this respect, they face two main challenges. First, they have to rely on forecasts, since monetary policy affects output and prices with a lag. Secondly, they cannot rely on a predetermined horizon, since price stability has a medium-term orientation. Against this background, I build on the EA case and propose an index to assess and forecast the probability of having inflation below the ECB target level in the next two years. The *Deflationary Pressure Index*. To accomplish this task, I use a broad set of macroeconomic variables and develop a two-step methodology building on combinatorial approach and exploiting parallel computation in Julia language. I first select twenty-four different models using three different loss functions (AUROC, MAE, RMSE) and specialize each of them to forecast a particular horizon. Then, I average the forecasts from all the models to get a meaningful out-of-sample index. The main idea is that an index capturing the probability of having inflation below the target can help in taking monetary policy decisions. In fact, central banks can be interested in the medium-run probability of deviating from the target as an additional measure to build confidence in their policy decisions. In the present context, the index shows that it is very unlikely to have inflation above the 2% level before March 2019.

References

Berge, T. J. and Jordá, Ó. (2011). Evaluating the Classification of Economic Activity into Recessions and Expansions. *American Economic Journal: Macroeconomics*, 3(2):246–277.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, 59(1):65–98.

Chari, V. V., Kehoe, P. J., and McGrattan, E. R. (2009). New Keynesian Models: Not Yet Useful for Policy Analysis. *American Economic Journal: Macroeconomics*, 1(1):242–266.

Christoffel, K. P., Coenen, G., and Warne, A. (2008). The new area-wide model of the euro area: a micro-founded open-economy model for forecasting and policy analysis. *ECB Working Paper, 944*.

Dieppe, A., Pandiella, A. G., Hall, S. G., and Willman, A. (2011). The ECB's New Multi-Country Model for the euro area: NMCM-with boundedly rational learning expectations. *ECB Working Paper Series, 1316*.

Dieppe, A., Pandiella, A. G., and Willman, A. (2012). The ECBs New Multi-Country Model for the euro area: NMCM Simulated with rational expectations. *Economic Modelling*, 29(6):2597–2614.

Elliott, G. and Timmermann, A. (2016). *Economic Forecasting*. Princeton University Press, Princeton.

Estrella, A. and Hardouvelis, G. A. (1991). The Term Structure as a Predictor of Real Economic Activity. *The Journal of Finance*, 46(2):555–576.

Estrella, A. and Mishkin, F. S. (1998). Predicting U.S. Recessions: Financial Variables as Leading Indicators. *Review of Economics and Statistics*, 80(1):45–61.

Faust, J. and Wright, J. H. (2013). Forecasting Inflation. In Elliott, G. and Timmermann, A., editors, *Handbook of Economic Forecasting*, volume 2A, chapter 1, pages 2–56. Elsevier.

Galí, J. (2015). *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*. Princeton University Press.

Giannone, D., Lenza, M., Momferatou, D., and Onorante, L. (2014). Short-term inflation projections: A Bayesian vector autoregressive approach. *International Journal of Forecasting*, 30(3):635–644.

Jackson, L. E., Kliesen, K. L., and Owyang, M. T. (2015). A Measure of Price Pressures. *Review*, 97(1):25–52.

Liu, W. and Moench, E. (2016). What predicts US recessions? *International Journal of Forecasting*, 32(4):1138–1150.

Mitchell, W. C. and Burns, A. F. (1938). Statistical indicators of cyclical revivals, NBER Bulletin 69, New York, Reprinted as: In Moore G. H. editor, Business Cycle Indicator, NBER Book Series Studies in Business Cycles. volume 1, chapter 6, pages 184–260. Princeton: Princeton University Press. 1961.

Stock, J. and Watson, M. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Stock, J. and Watson, M. (2016). Dynamic Factor Models Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics. In Taylor, J. B. and Uhlig, H., editors, *Handbook of Macroeconomics*, volume 2A, chapter 8, pages 415–525. Elsevier.

Stock, J. H. and Watson, M. W. (1989). New indexes of coincident and leading economic indicators. *NBER macroeconomics annual*, 4:351–394.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press, Cambridge (Massachusetts).

Wu, J. C. (2017). Time-Varying Lower Bound of Interest Rates in Europe. *Chicago Booth Research Paper, No. 17-06*.

Wu, J. C. and Xia, F. D. (2016). Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound. *Journal of Money Credit and Banking*, 48(2-3):253–291.

A Additional table

Table 9: Complete monthly dataset. Variable names, identification and transformation codes.

Variable	RIC/DS ID Code	Transformation Code
HICP EA	aXZCPIHICP/C	3
IP EA	aXZCINDG/CA	3
Consumers good	aXZPDAGCGS/A	3
Durable	aXZPDAGCDRB/A	3
Non durable	aXZPDAGCNDR/A	3
Intermediate	aXZPDAGINTG/A	3
Energy	aXZPDAGENE/A	3
Capital	aXZPDAGCAPG/A	3
Construction	aXZIPCON/A	3
Manufacturing	aXZIPMAN/A	3
Unemploy. rate EA	aXZUNR/A	1
Credit gen gov	aXZCRDGOV/A	3
Car regist	aXZCRDRG/A	3
Business climate	aXZBUSCLIM	6
Consumer confidence	aXZECOSE	6
Industrial confidence	aXZBSMFGCI/A	6
Retail confidence	aXZBSSVRTCI/A	6
Construction confidence	aXZBSCSCI/A	6
Service confidence	aXZBUCFM/A	6
Core CPI ea	aXZCCORF/C	3
Eonia	aXZONIA	1
M1	aXZM1	3
M2	aXZM2	3
M3	aXZM3	3
Neer	aXZINECE/C	3
US ff rate	aUSFEDFUND	1
IP DE	aDECINDG/A	3
IP ES	aESCINDG/A	3
IP FR	aFRCINDG/A	3
IP IT	aITCINDG/A	3
Unemploy. DE	aDECUNPQ/A	1
Unemploy. ES	aESCUNPQ/A	1
Unemploy. FR	aFRCUNPQ/A	1
HICP DE	aITUNRM/A	3
HICP ES	aESHICP	3
HICP FR	aFRHICP	3
HICP IT	aITHICP	3
Core CPI DE	aDECCORF/C	3
Core CPI FR	aESCCORF/C	3
Core CPI IT	aITCCORF/C	3
EA stock	.STOXX50E	3
EA bank stock	.SX7P	3
US stock	.SPX	3
US vol	.VIX	3
Crude	LCOc1	3
US10Y	US10YT=RR	1
EURIBOR3M	EURIBOR3MD=	1
EURIBOR6M	EURIBOR6MD=	1
EURIBOR1Y	EURIBOR1YD=	1
DE stock	.GDAXI	3
ES stock	.IBEX	3
FR stock	.FCHI	3
IT stock	.FTMIB	3
DE2YT	DE2YT=RR	1

DE5YT	DE5YT=RR	1
DE10YT	DE10YT=RR	1
ES5YT	ES2YT=RR	1
ES10YT	ES10YT=RR	1
FR2YT	FR2YT=RR	1
FR5YT	FR5YT=RR	1
FR10YT	FR10YT=RR	1
IT2YT	IT2YT=RR	1
IT5YT	IT5YT=RR	1
IT10YT	IT10YT=RR	1
NL2YT	IE2YT=RR	1
NL5YT	NL5YT=RR	1
NL10YT	NL10YT=RR	1
EA short repo	RC2AALM	1
EA2Y	EMECB2Y.	1
EA3Y	EMECB3Y.	1
EA5Y	EMECB5Y.	1
EA7Y	EMECB7Y.	1
EA10Y	EMGBOND.	1
Loans to nonfin corps	EMEBMC..A	3
Loans to hsld	EMEBMH..A	3
Loans to non-mfi	EMEBMEO.A	3
Loans to mfi	EMECBXLMA	3
IT CPI SA	ITCCPI..E	3
IT core CPI SA	ITCCOR..E	3
ES CPI SA	ESCCOR..E	3
DE CPI SA	BDCONPRCE	3
DE CPI core SA	BDUSFG10E	3
FR CPI SA	FRCONPRCE	3
FR core SA	FRCPUNDEE	3
Price trends 12M	EMZEWCP.R	6
Econ 12M	EKTOT4BSQ	6
Unemployment 12M SA	EKTOT7BSQ	6
REER	EMI..RECE	3
US crude	USSCOP.BP	3
Shadow	-	1

Note: the table shows the entire dataset along with the Thomson Reuters Eikon and Datastream identification codes for each variable. The transformation codes from 1 to 6 correspond to level, monthly difference, annual difference, log level, monthly log difference and annual log difference. The shadow rate for the EA is provided by Wu (2017) and available on her web-page.