Forecasting Deflation Probability in the EA: A Combinatoric Approach *

Luca Brugnolini

Central Bank of Malta & University of Rome Tor Vergata

October 9, 2018

Abstract

I develop a two-step subset selection procedure to extract the best-performing predictors from a large dataset and combine them to identify a set of best-performing models. I apply the methodology to build an index to forecast the probability of having the euro area year-on-year inflation below the 2% level in a medium-term horizon—i.e., the Deflationary Pressure Index (DPI). I compare the index with the probabilities reported in the European Central Bank Survey of Professional Forecasters (ECB SPF) and show that, although the indexes are comoving, the DPI is more operationally convenient and timely in catching the inflation turning points. As a final exercise, in a real out-of-sample forecast, the index shows that having medium-term inflation above the 2% level before 2019 is unlikely.

JEL classification: C25, C63, E3, E58

Keywords: inflation, prediction, index, euro area, ECB, ROC

I 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. I am grateful to Antonello D'Agostino, Giuseppe Ragusa, Giovanni Ricco, Giulio Nicoletti, Marco del Negro, Domenico Giannone, and Gregor von Schweinitz for their precious advice. Also, I would like to thank the participants of the annual meeting of the Society for Computational Economics (CEF 2018) and the International Association for Applied Econometrics (IAAE 2018), the I Vienna Workshop on Forecasting, and colleagues at the Central Bank of Malta's Research Department for useful comments. All errors are mine.

Corresponding author: Luca Brugnolini brugnolinil@centralbankmalta.org, senior economist at the Central Bank of Malta's Research Department.

1 Introduction

Stabilizing prices is an arduous task. On the one hand, central banks cannot rely only on contemporaneous inflation measures, as monetary policy actions exerted today transmits to prices solely in future periods (Friedman, 1961, 1972). On the other, the correct implementation of monetary policy actions is related to a general medium-term orientation (see, e.g., Bernanke et al., 2018). Therefore, the interval for achieving price stability has to be extremely general. 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—i.e., the *Deflationary Pressure Index* (DPI). 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 (Estrella and Hardouvelis, 1991; Estrella and Mishkin, 1998). In this respective, I create an index to predict the likelihood of having inflation below the 2% level in the next two years by averaging the forecasted probabilities from the best horizon-calibrated models.

The reason for forecasting probabilities is related to the prominent role achieved by density forecast in the last decades (J.P. Morgan, 1996; Diebold et al., 1997). In general, supporting point forecasts with probabilities provides a quantitative assessment of the forecaster uncertainty and helps policymakers in the decision process—a prominent example is the Bank of England *fan chart* (Britton et al., 1998). However, probabilities are intrinsically informative and can be a primary source of knowledge. For example, in the United States, since 1968 the Survey of Professional Forecasters (SPF) asks respondents to provide density forecasts of inflation and output (Zarnowitz and Lambros, 1987). In Europe, the history is much shorter. However, the European Central Bank (ECB) has directly collected density forecasts since 1999. Although surveys are often accurate in predicting inflation (Faust and Wright, 2013), there are at least two reasons for developing a novel tool to forecast probabilities. First, surveys are deterministically released and cannot be timely updated. Secondly, median survey movements are not easily interpretable, as the model specification of the forecasters is unknown. On the contrary, the index I propose, by having a known specification and being directly updatable, overcomes both disadvantages.

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, 2002a). In this work, I partially build on this discussion.

However, I also builds on the machine learning literature of *best subset selection* (see, e.g., Liu and Motoda, 2007). Moreover, some specific choices are related to the particular application of medium-term inflation forecasting. Hence, to find the best-performing model at each predicted horizon, I develop a two-step procedure to select and combine an optimal subset of variables extracted from a large dataset. In a first step, I discard the variables producing a poor out-of-sample density forecast using a binomial probability model. Then, in a second step, I combine the best-performing variables in a multivariate framework to select the best performing models.

The highlighted procedure is significantly different from the problem of extracting unobserved factors from a large pool of variables. In particular, not involving any data transformation, the method allows for a more substantial degree of interpretability. Secondly, while the inclusion of non-linearities in the model specification does not increase the complexity for best subset selection, this feature has a different effect on factor analysis. The reason is linked to the fact that in a hypothetical state-space model the measurement equation would be nonlinear, implying that the standard Kalman filter cannot be used to derive the likelihood. Finally, the choice of a discrete probability model makes the DPI very different from standard leading indicators, as the outcome is a density and not a point forecast.

In this sense, the constructed index resemble more closely the indicators developed in the literature of the early-warnings (Kaminsky et al., 1998; Kaminsky and Reinhart, 1999; Reinhart, 2002), with the main difference that it is developed to forecast the probability of having inflation around a central bank target. This feature also implies that a multi-horizon prediction has to be designed to deal with the general medium-term orientation. Finally, with respect to continuous models capable of producing density forecasts and dealing with large datasets, as Bayesian shrinkage (De Mol et al., 2008; Bańbura et al., 2010; Giannone et al., 2014, 2015), the use of discrete models allows to exploit specific loss functions for variable selection based on the model ability to generate true and false signals independently of the selected threshold. The most prominent example is the *Area Under the Receiver Operating Characteristics* (AUROC), which I describe and exploit in the next sections.

The methodology I propose is new in this field, and because there are few 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 only on point forecast, then they include the extracted factors in a discrete model to predict the

probabilities. In this sense, the probability model is not directly calibrated. Also, the same model is used to predict each horizon, without considering that longer and shorter horizons would benefit from different specifications. On the contrary, the present paper gives prominent importance to each model and prediction used to construct the index.

In particular, I apply the procedure to the euro area case and develop a tool to forecast the probability that the *Harmonized Index of Consumer Prices* (HICP) inflation exceeds the ECB target in the medium-term. Finally, to assess the predictive ability of the index, I compare it to an *ad hoc* measure constructed from the ECB SPF probabilities. I show that the two indexes comove, even if the SPF often fails in capturing the turning points between the two states—i.e., inflation above and below the target. Also, I show that both indexes predict that the medium-term probability of having inflation higher than 2% before March 2019 is extremely low.

The rest of the paper is organized as follows. Section 2 describes the two-step model selection procedure. Section 3 apply the methodology to the euro area case, builds the *Deflationary Pressure Index* and compares it with a probability measure built from the ECB SPF. Finally, section 4 concludes.

2 Methodology

The literature has developed many approaches to deal with large datasets. In general, these techniques either exploit dimensionality reduction, as factor models (Forni et al., 2000; Stock and Watson, 2002b,a), or employ parameter selection/shrinkage, by including an L_1, L_2-penalty function, or a combination of the two into the maximization problem—prominent examples are the lasso estimator, Tikhonov regularization and elastic net. A similar goal can also be achieved by Bayesian spike-and-slab prior (Mitchell and Beauchamp, 1988; Madigan and Raftery, 1994; George and McCulloch, 1997) or Bayesian Model Averaging (Hoeting et al., 1999). However, although extremely appealing in various contexts, one of the main drawbacks of these approaches are related to the usual trade-off between flexibility and interpretability. Especially in forecasting, being able of understanding the specific variables causing a change in the predictions is essential to judge the reliability of the forecast itself. In this sense, the best subset selection approach allows for the maximum degree of interpretability among all the candidates. In particular, I develop a two-step selection procedure focusing exclusively on the out-of-sample performance of each of the model; in a first step, I directly discard the variables producing a poor out-of-sample density forecast using a binomial probability model.

Then, in a second step, I combine the best-performing variables in a multivariate framework to select the best-performing models. Therefore each model is tailored to predict a specific horizon. Assessing the predictive power of different variables at different horizons is one of the appealing features of out-of-sample metrics. In fact, it is well-known among forecasters that different variables have different predictive power along different horizons. For example, the yield curve slope—i.e., the difference between short to long-term yields—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 for the medium to long-term forecasts.

Along with this line, I test different variables to select a pool of predictors with a proved forecasting power along different horizons. Finally, the predictions obtained from the best models are be averaged along a time-dimension to create the index. In what follows, I describe the two main steps of the model selection procedure.

2.1 First step

Equation (1) presents the model specification. In a first step, I project the discrete variable Y_{t+h} on the space spanned by each variable x_t^i in the full dataset using a univariate probability model.

$$Y_{t+h} = G(x_t^i) + \epsilon_t, \quad h = 1, \dots, H, \quad i = 1, \dots, K \tag{1}$$

In the notation, $G(\cdot)$ is a cumulative distribution function, ϵ_t is the regression error, $h = 1, \dots, H$ is the forecast horizon and K is the number of predictors in the dataset. The models are initially estimated on a pre-sample up to time T_p. From T_p+1 to T each model is recursively estimated by adding a data point at each iteration and computing the direct forecast for each horizon h. At the end of the procedure, the predictive ability of each model is assessed against one or a pool of loss functions $\mathcal{L}(\cdot)$. In particular, I select the standard L_1 and L_2 losses, besides the Area Under the Receiver Operating Characteristics. The last criterion is tailored to signal extraction and ranks the predictive ability of a model according to its ability to report true and false signals irrespectively of the selected threshold. Appendix A provides a detailed review of the AUROC. Out of the first step, $K_1 < K$ variables are selected by picking the best M predictors according to all criteria $\mathcal{L}^s$ and horizons h. These variables are then combined in a second step to find the best performing models at each horizon. There are different ways in which the hyperparameter M can be chosen. However, in what follows, we select it according to computational reasons only. In particular, there is a direct mapping between

$M \to K_1$ which depends on the particular dataset X, the loss functions $\mathcal{L}^s$, and the horizons h. Therefore, M can be selected in a way that exactly K_1 predictors are kept in the second step. The reason is related to the fact that there are $2^{K_1} - 1$ possible models in the second step, and, unless using parallel computing, a researcher is often bound in a range of fifteen to twenty predictors. However, in recent times, a particular form of parallel computing, as grid computing, allows researchers to relax this constraint by simultaneously employing thousand of processors simultaneously[1]. The first step procedure is summarized in algorithm 1.

Algorithm 1: first step of the selection procedure.

for $i = 1$ **to** K **do**
  Pre-estimating the model in (1) up to time T_p ;
  for $t = T_p + 1$ **to** T **do**
    Re-estimating the model and computing the direct forecast for each horizon h;
  end for
  Selecting and pooling the M *best-performing models according to* $\mathcal{L}^s$;
end for

2.2 Second step

In the second step, I fit a separate multivariate discrete probability model to all combinations of the K_1 predictors selected in the first stage, as shown in equation (2).

$$Y_{t+h} = G(x_t^c) + \epsilon_t, \quad h = 1, \dots, H, \quad c = 1, \dots, C^j, \quad j = 2, \dots, K_1 \tag{2}$$

Where c is a combination of j variables. As in the previous step, the models are pre-estimated on a sample up to time T_p, and from $T_p + 1$ to T each model is recursively estimated by adding a data point at each iteration. Then, the direct forecasts for each horizon h are computed. At the end of the procedure, the predictive ability of each model is assessed by the loss-functions $\mathcal{L}^s$. As a twist, in this step, each model can be augmented with a counterpart specification which includes a common factor F_t extracted from the full dataset. This procedure increases the computational burden, by adding a costly operation at each iteration and increasing the number of models to $2^{K_1+1} - 2$. However, in some application, adding a common factor might include information useful for the forecast. A specific case is when the presence of few strong-correlated predictors characterizes the full dataset. In the application, the common factor is estimated non-parametrically via principal component. In particular, I estimate the factor

[1] An example is the Techila grid.

from the eigenvector corresponding to the largest eigenvalue of the variance-covariance matrix of the demeaned full dataset. However, different procedures can easily be embedded in the algorithm. Naturally, including the models augmented with a common factor does not force the algorithm to select them, as in the out-of-sample prediction they might underperform their counterpart specifications without the factor. The procedure described in the second step can be summarized in algorithm 2.

Algorithm 2: second step of the selection procedure.

for $j = 1$ **to** K_1 **do**
  Compute the number of combinations $C^j = \binom{K_1}{j}$ *with* j *variables*;
  for $c = 1$ **to** C^j **do**
    Pre-estimating the model $Y_{t+h} = G(x_t^c) + \epsilon_t, \quad h = 1, \dots, H, \quad c = 1, \dots, C^j$ *up to time* T_p ;
    if *Factor* **then**
      Pre-estimating the model $Y_{t+h} = G(x_t^c, F_t) + \epsilon_t, \quad h = 1, \dots, H, \quad c = 1, \dots, C^j$ *up to time* T_p ;
    end if
    for $t = T_p + 1$ **to** T **do**
      Re-estimating the model and computing the direct forecast for each horizon h;
      if *Factor* **then**
        Re-estimating the model and computing the direct forecast for each horizon h;
      end if
    end for
    Storing the model scores in terms of loss-functions $\mathcal{L}^s$;
  end for
  Selecting the $B^{h,j}$ *best-performing models for each horizon according to* $\mathcal{L}^s$;
end for

In many applications, having K_1 different variables may imply that the number of possible combinations is extremely high. Also, in some cases, a researcher would prefer to have a large number of predictors to combine, but only allowing a certain number of regressors to enter into the model. This choice may reflect a trade-off in terms of overfitting and model flexibility, but also considerations in terms of computational burden. This problem can be easily handled by adding a second hyperparameter $K_2 \leq K_1$ into the algorithm to set the upper bound of the number of regressors in a single model. K_2 restricts the number of combinations from $2^{K_1} - 1$ to $\sum_{j=1}^{K_2} \binom{K_1}{j} = 2^{K_1} - \sum_{j=K_2+1}^{K_1} \binom{K_1}{j} - 1$. Naturally, as $K_2 = K_1$ all the $2^{K_1} - 1$ combinations are included. From an operational point of view, there are different ways in which K_2 can be selected. However, as all the models included in $K_2^{(1)} << K_2^{(2)}$ are also considered in $K_2^{(2)}$, a leading choice in setting this parameter should be related to reducing the computational burden. For example, selecting $K_1 = 20$ predictors and allowing for a maximum of $K_2 = 10$

regressors in the models, approximately reduces the number of models to estimate from one million to half.

2.3 Building the index

After selecting the best models according to the selected $\mathcal{L}^s$ loss functions, an index can be constructed by averaging over a time dimension the forecasts produced at each point in time by each best models. In particular, at each point t, each best model can be used to generate a single forecast for its tailored horizon h, and predictions averaged and recorded at time t. Following this approach, each point of the index reports the probability of an event over the next H periods, as shown in equation (3).

$$I_t^{\mathcal{L}} = \frac{1}{H} \sum_h \omega_h \hat{Y}_{t+h}^{\mathcal{L},h}, \quad h = 1, \ldots, H \tag{3}$$

Where the superscript $\mathcal{L}$ highlights that, depending on the number of loss-functions applied for the evaluation, a corresponding number of indexes can be produced. This feature is related to the fact that, at the end of the second step, each loss-function produces its own set of best models for the horizons $h = 1, \ldots, H$. Also, ω_h represent a set of weights that can be used to enhance/reduce the credibility towards certain models. Operationally, the weights can be selected using statistical methods, or to reflect some researcher priors. For example, by weighting fewer predictions at longer horizons. With the appropriate standardization, a second set of weights $\omega_{\mathcal{L}}$ can be also included in equation (3), in order to average both across horizons and loss-functions. However, in the application presented in the next sections, I avoid averaging the predictions across criteria to cross-benchmark the extracted indexes. Nevertheless, depending on the application, this feature can be easily embedded in the algorithm.

3 Application

3.1 Model selection

In this section, I apply the methodology described in the previous one to a large dataset of national and euro area indicators to forecast the probability of having the EA HICP year-on-year below the 2% level over the next two years. The variables contained in the dataset and their transformations are reported in appendix B, while appendix C reviews the construction of a discrete dependent variable for inflation. In the first step, for each variable i, I pre-estimate

Table 1: first step, all best predictors M, all horizons h, and loss-functions $\mathcal{L}^j$.

Horiz.	**Criteria**					
	AUROC		*MAE*		*RMSE*	
	first best	*second best*	*first best*	*second best*	*first best*	*second best*
$h = 1$	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 criterion 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 activity 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.

the model in equation (1) on the period between 1999M1 to 2007M2. In the particular application, I assume normally distributed errors ϵ_t to characterize $G(\cdot)$ as a Gaussian CDF. Then, for each horizon $h = [1, 3, 6, 9, 12, 15, 18, 24]$, I re-estimate the model recursively from 2007M3 to 2017M3 (121 months), increasing the sample size by one data-point at each iteration. Finally, I evaluate each model $\mathcal{M}_{i,h}$ with three different loss-functions $\mathcal{L}$. I select the AUROC as a natural candidate for this type of analysis, but also the mean absolute error (MAE) and the root mean squared Error (RMSE) as standard in the forecasting literature. Finally, I select and pool the best $M = 2$ predictors to match exactly $K_1 = 20$ variables in the second step. Table 1 shows the best selected variables. Although the table shows an intricate 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). Secondly, following the 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, but might also be linked to the unconventional monetary policy programs undertaken by both countries in the last years. In particular, asset purchase programmes were tailored to target longer maturities. For example, the euro area Public Security Purchase Programme (PSPP) has a weighted average maturity close to eight years. On the contrary, the AUROC predictors are more heterogeneous. From six months to one-year-ahead,

Table 2: first step, unique variables. All criteria, horizons and loss-functions.

Price	Interest rate	Real	Monetary	Survey
HICP DE	EA3Y	Interm. 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 declared 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 activity 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.

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 excellent 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, surveys as the industrial confidence indicators and expectations show an excellent forecasting ability. 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, a German inflation measure dominates.

Table 2 summarizes the 20 unique predictors delivered by the first selection step. Not surprisingly, the selected predictors approximately coincide with variables considered as the main determinants of inflation by established economic relationships—e.g., the New Keynesian Investment-Saving (NKIS) curve, the New-Keynesian Phillips curve (NKPC), and the Taylor rule embedded in standard three equations Dynamic Stochastic General Equilibrium (DSGE) models—as inflation itself, inflation expectations, interest rates, and output measures. The strong linkage between the selected variables and economic theory builds confidence in the procedure and benefits from the interpretation point of view. Also, it is valuable to notice that similar findings are common in the inflation forecasting literature. In fact, various forms of the NKPC are often used as a proper reduced-form model to forecast inflation (Faust and Wright, 2013).

In the second step, I fit a separate probit model to all possible combinations of the 20 predictors selected in the first stage, plus a counterpart model for each combination augmented

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.*

with the first principal component extracted from the full dataset. Having twenty different variables plus the augmented models implies that the number of possible combinations is extremely high. Therefore, I set an upper bound of $K_2 = 10$ maximum predictors. This restriction leads the algorithm to estimate around 1.2 million models. However, to understand the true running time of the algorithm, the recursive structure of the out-of-sample exercise should be considered. In particular, with $T^{out} = 121$ points in the out-of-sample estimation and $H = 8$ forecasted horizons, the algorithm estimates approximately 1.2 billion models[2]. 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. Therefore, the second stage of the selection process returns a set of H models for each loss-function. 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 the inclusion of a common factor. Also, to have a benchmark for the comparison, I build a *naive model* fitting a binomial

[2]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 intensive problems. Together with the paper, the ForecastingCombinations.jl package is released on my GitHub page at the following link https://github.com/lucabrugnolini/NFP.jl. The package allows using both linear and probability models in both steps, and the example shows an application on predicting US Non-farm payrolls with a linear model at different h-steps ahead. Also, different loss functions can be easily added to the procedure. All the hyperparameters as horizons or number of best variables to select are customizable.

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).*

probit model on 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. The tables show some common characteristics among the selected specifications which 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 defeat. For longer horizons, the selected models are usually more precise. 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 regarding the number of selected variables, while the MAE is the least. Fourth, on average, it seems that all models use predictors coming from different classes, implying that different information is useful in improving the prediction. Figure 1 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 the chosen 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 criterion, the curve is stable or slightly

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

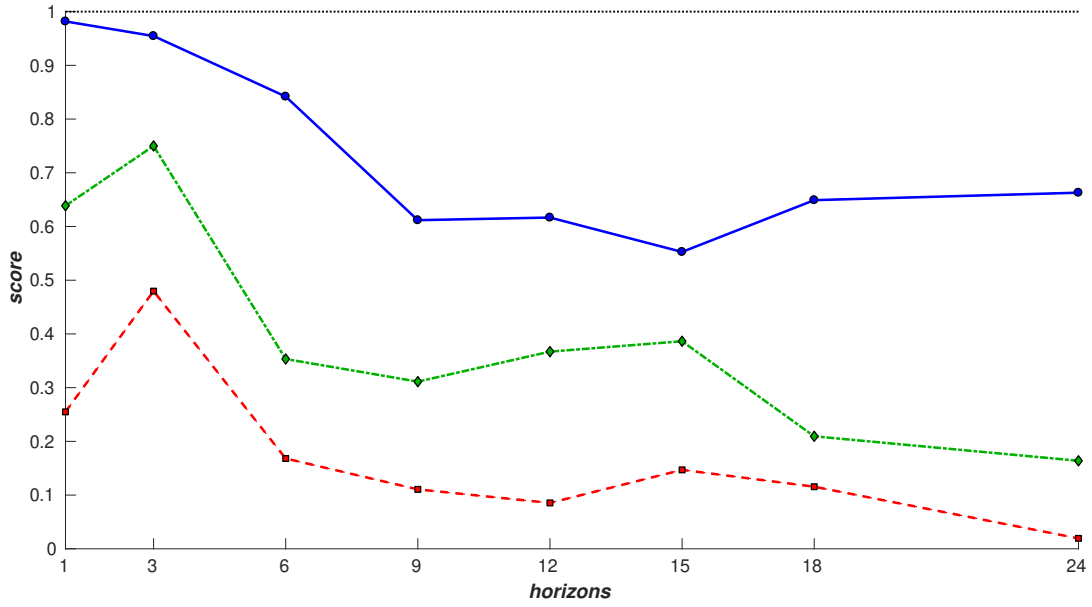
Note: the table shows the best model for each horizon h, selected among $\sum_{k=1}^{10}\binom{20}{k}$ *models. The variables are used as regressors in a multivariate probit model for forecasting the probability of having inflation below the 2% level according to the MAE criteria. The table also shows 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).*

Figure 1: relative model loss in terms of AUROC, MAE, and RMSE.



Note: the figure shows the 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.

increasing/decreasing.

3.2 Forecasting performance

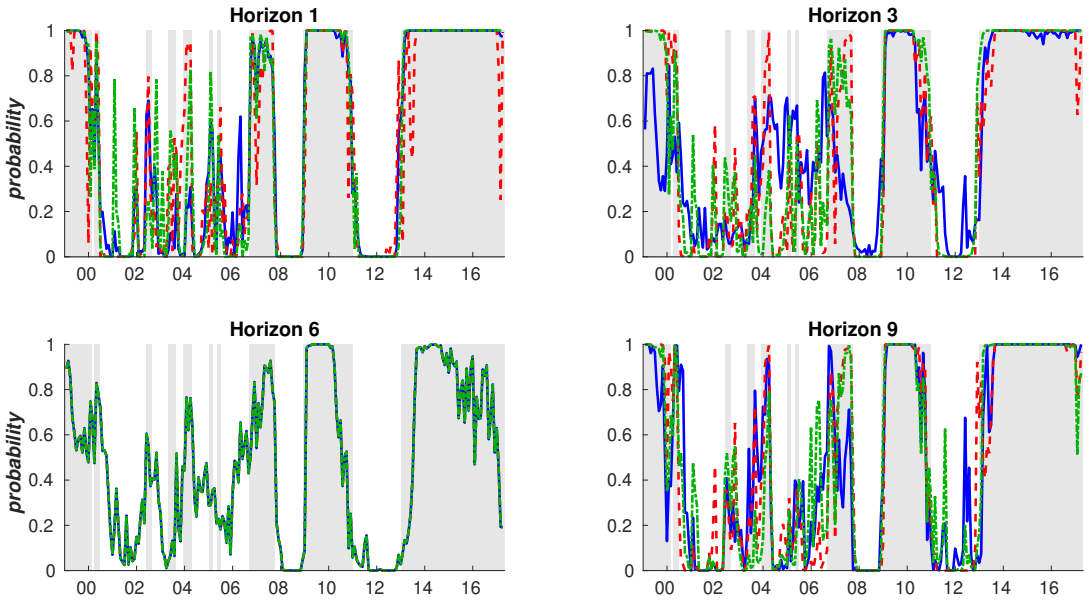
In this section, I analyze the in-sample and out-of-sample performance of the models selected in the previous section. I use the in-sample fit to get a general feeling of the overall performance of the models. The reason is also related to the lack of a long time series for the particular application. In fact, 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 fit, I evaluate the out-of-sample performance of the models. Also, in a true out-of-sample exercise, I predict the probability of having low inflation from March 2017 to March 2019.

3.2.1 In-sample analysis

I start analyzing 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 2 and 3 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 shaded area represent periods of inflation lower than 2%.

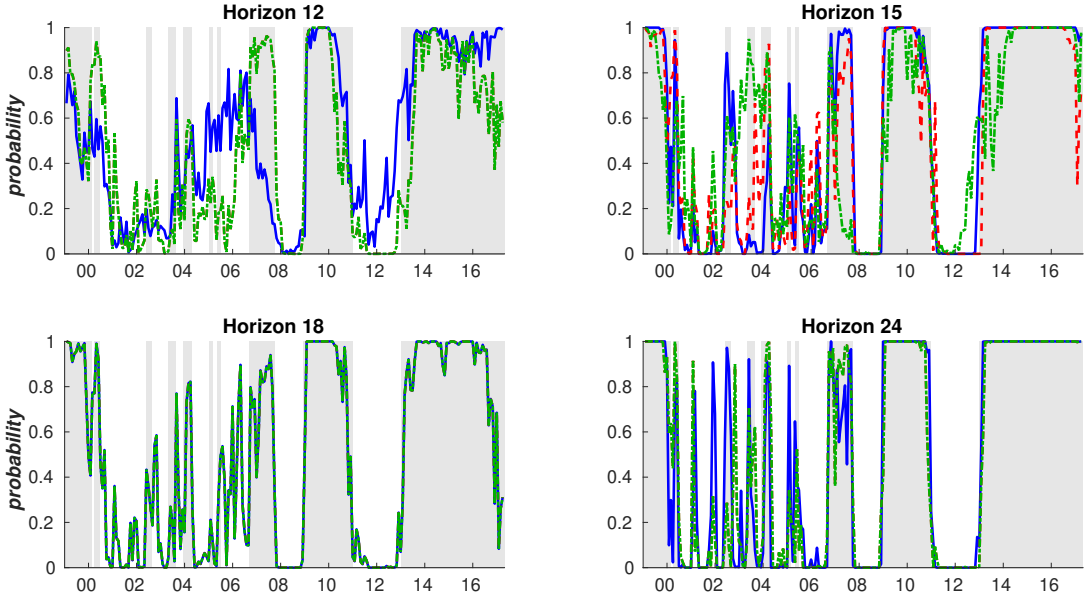
Except for very short periods, the first part of the sample is mainly characterized by having inflation above the threshold. However, the rapid changes in the regimes create many false signals in the estimated probabilities. This feature is evident, especially at shorter horizons. Starting from 2008 the series is characterized by prolonged periods of inflation above/below the threshold, which substantially reduce false signals. Overall, the models have a satisfactory

Figure 2: in-sample model fit, horizons 1, 3, 6, and 9.



Note: 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 shaded area represent periods with inflation below the 2% level.

Figure 3: in-sample model fit, horizons 12, 15, 18, and 24.



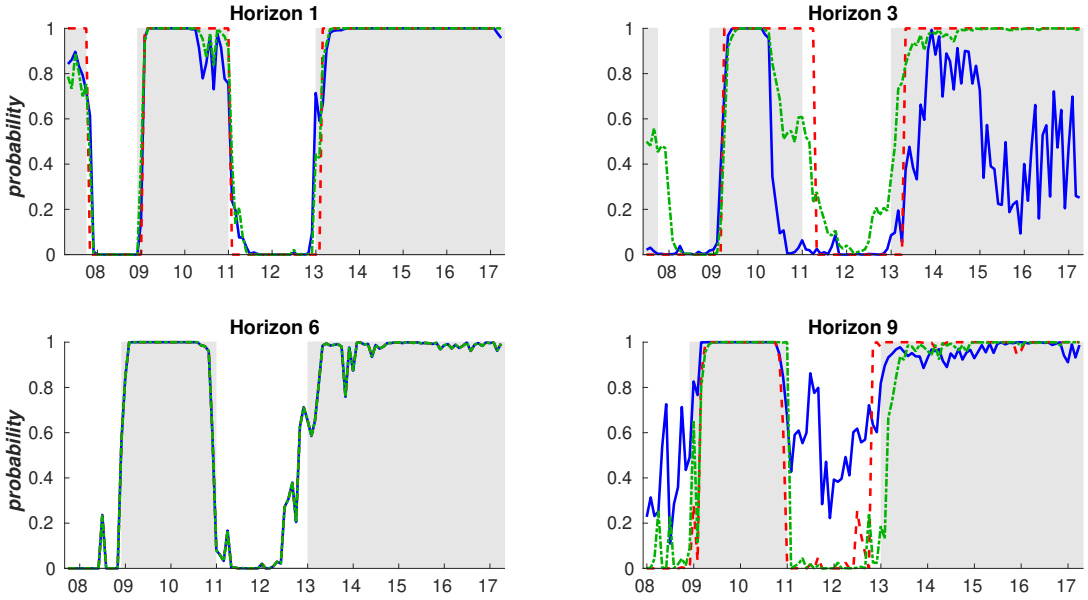
Note: In-sample-model-fit for horizons $h = 12$ to $h = 24$. Each panel shows the best model selected using AUROC (solid blue lines), MAE (dash-dot green lines) and RMSE (dotted red lines). The shaded area represent periods with inflation below the 2% level.

in-sample prediction ability. Also, the models chosen with the three criteria are very similar, and on many occasions, the estimated probabilities overlap—e.g., horizons $h = 6$ and $h = 18$.

3.2.2 Out-of-sample analysis

After assessing the in-sample fit of the selected models, I evaluate their out-of-sample performance. For each set of variables, I pre-estimate the model from January 1999 to February 2007 and compute a direct forecast up to the end of the sample (March 2017). Figure 4 and 5 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 to 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 a visual inspection, the out-of-sample predictions do not show

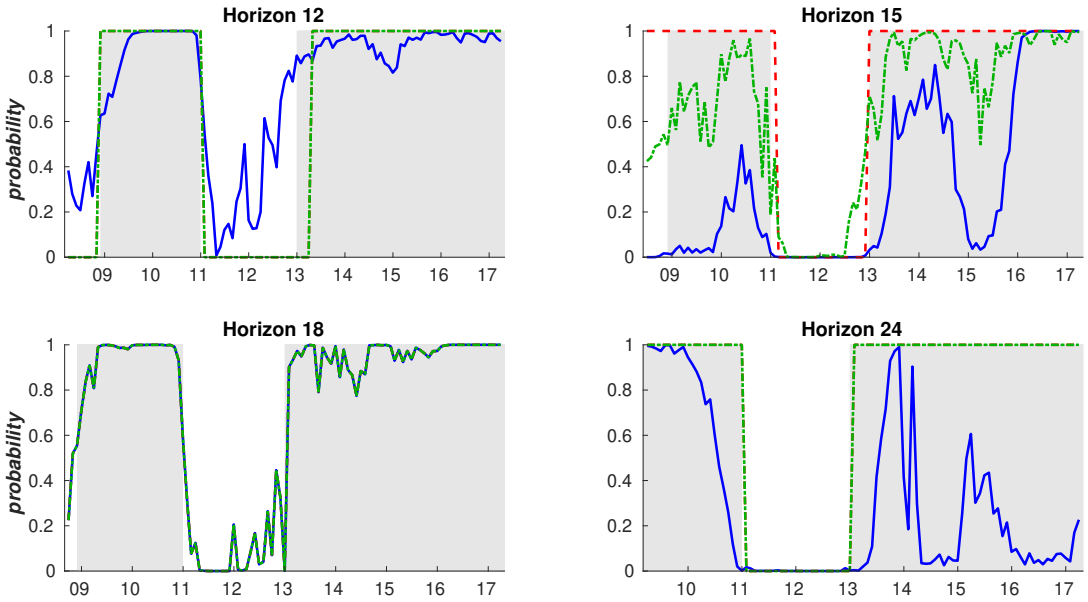
Figure 4: out-of-sample forecast, horizons 1, 3, 6, and 9.



Note: 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 shaded area 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, and timely catch 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 last one. The MAE model is slightly delayed with respect to the turning points, while the AUROC model delivers a poor job. At horizon six and eighteen the

Figure 5: out-of-sample forecast, horizons 12, 15, 18, and 21.



Note: 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 shaded area represent periods with inflation below the 2% level.

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 models are the best performers, and in many occasions, the variables selected by the two coincide. From the other side, the AUROC is poor performing, especially at horizons $h = 15$ and $h = 24$. The main problem seems to be related to the high autocorrelation of the estimated probabilities which miss the turning points. At the opposite, the other models do a solid job in capturing the switch from one state to the other.

To investigate this finding, 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. The labels AUROC, MAE, and RMSE highlight 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 (first panel, first column—blue), for the MAE panels is the MAE column (second panel, second column—green) and for the RMSE panels is the RMSE column (third panel, third 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 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 by the naive model (second and third panel). The same happens to very

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

Horizon	AUROC			MAE			RMSE		
	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 chosen 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.

short horizons for the MAE in terms of AUROC (first panel, 0.99 and 0.96). However, the distance between the two models is much smaller (one to four percentage points).

3.3 The 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 (4) shows the DPI. The index is a simple average ($\omega_h = 1$) over the best model forecasts at each horizon h. In particular, the index is built having in mind a researcher that at time t is forecasting from one to two years ahead, and averaging 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 over the next two years.

$$DPI_t^{\mathcal{L}} = \frac{1}{H} \sum_h \hat{Y}_{t+h}^{\mathcal{L},h} \quad h = [1, 3, 6, 9, 12, 15, 18, 24], \quad \mathcal{L} = \{AUROC, MAE, RMSE\} \tag{4}$$

I avoid averaging across criteria to compare the three different indexes. Figure 6 shows the constructed Deflationary Pressure Indexes against the periods in which inflation is below the 2% level. The chart presents three prominent features; first, the movements in the indexes are extremely close. This characteristic can be considered as a signal of robustness, as the three are created using different loss functions. Secondly, the indexes seem to have good predicting

Figure 6: Deflationary Pressure Index.

Note: Deflationary Pressure Index. This indicator 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 shaded area represent periods with inflation smaller than 2%.

power. This feature is evident in the period between 2009 and 2013, as they start moving in the correct direction before the regime switch in inflation. Finally, as the three indexes signal a very high probability of having inflation below the 2% level, I 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 from the ECB SPF probabilities.

3.4 ECB Survey of Professional Forecasters: a comparison

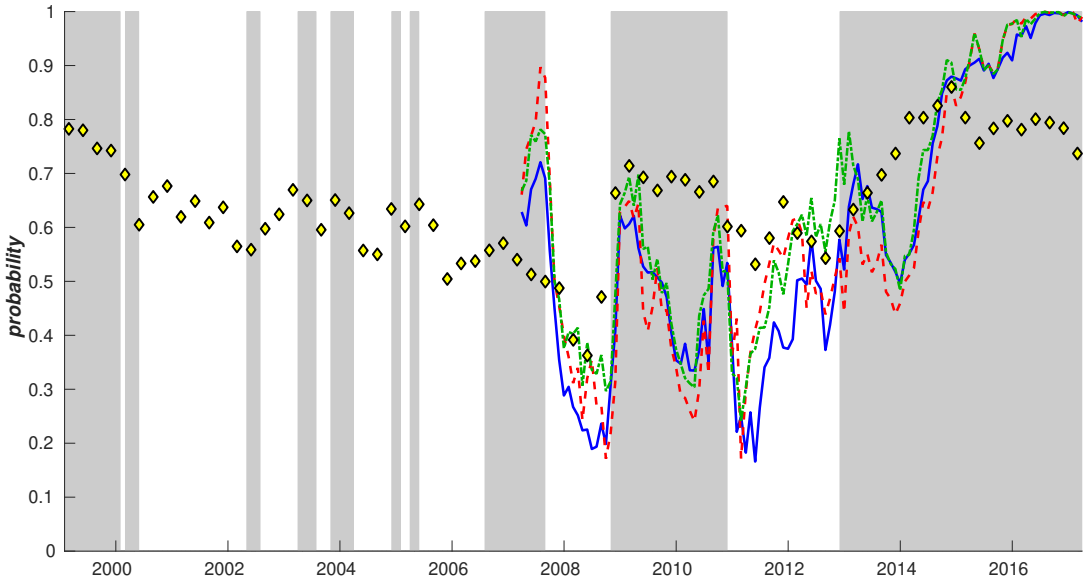
In this section, I compare the DPI against the ECB *Survey of Professional Forecasters*. Starting from December 2000 the SPF is quarterly collected by surveying more than 80 professional forecasters[3]. In a standard survey, forecasters express their point forecasts for inflation (as well as GDP growth and unemployment) over a specific time horizon. 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 EA 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 the final measure is an average of all the forecasts.

To create an index comparable to the DPI, I construct a new measure by cumulating the probabilities of having inflation below the 2% level. These range between -1% and 2%

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

and refer to a 24 months horizon. However, given that the SPF is quarterly collected, the comparison involves mixed frequency. Therefore, figure 7 shows the quarterly SPF median survey forecast (yellow diamonds) against the monthly Deflationary Pressure Index. Due to

Figure 7: Deflationary Pressure Index against ECB SPF.



Note: Deflationary Pressure Index. This indicator 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 shaded area represent periods with inflation smaller than 2%.

the lower-frequency nature, the SPF is more autocorrelated than the DPI. This feature is also responsible for missing the turning points in the inflation regimes. The strong autocorrelation is especially evident in the last two transition periods. However, apart from this characteristics, the two measures are incredibly similar. This feature is true especially between 2013 and 2017. Finally, it is interesting to notice that in the final part of the series, both indexes display a rapid increase in the probabilities. However, at the end of the sample, the two measures slightly diverge, with the SPF measure declining to 75%. Nevertheless, both models confirm a high probability of inflation below the 2% level up to 2019.

4 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 propose a methodology based on best subset selection to build an index to assess and forecast

the probability of having inflation around a certain threshold. In particular, I apply it to the euro area case, and create an index to predict the probability of having EA HICP below the 2% level over the next two years—i.e., the *Deflationary Pressure Index* (DPI). The main idea related to this measure is that capturing the probability of having inflation below the target can support policymakers regarding 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 actions. In the present context, the index shows that it is unlikely to have medium-term inflation above the 2% level before March 2019.

References

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

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.

Bernanke, B. S., Laubach, T., Mishkin, F. S., and Posen, A. S. (2018). *Inflation targeting: lessons from the international experience*. Princeton University Press.

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

Britton, E., Fisher, P., and Whitley, J. (1998). The inflation report projections: understanding the fan chart. *Bank of England. Quarterly Bulletin*, 38(1):30.

De Mol, C., Giannone, D., and Reichlin, L. (2008). Forecasting using a large number of predictors: Is bayesian shrinkage a valid alternative to principal components? *Journal of Econometrics*, 146(2):318–328.

Diebold, F. X., Tay, A. S., and Wallis, K. F. (1997). Evaluating density forecasts of inflation: the survey of professional forecasters.

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.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *Review of Economics and statistics*, 82(4):540–554.

Friedman, M. (1961). The lag in effect of monetary policy. *Journal of Political Economy*, 69(5):447–466.

Friedman, M. (1972). Have monetary policies failed? *The American Economic Review*, 62(1/2):11–18.

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica sinica*, pages 339–373.

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.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *The Review of Economics and Statistics*, 97(2):436–451.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401.

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

J.P. Morgan (1996). Riskmetrics—technical document.

Kaminsky, G., Lizondo, S., and Reinhart, C. M. (1998). Leading indicators of currency crises. *Staff Papers*, 45(1):1–48.

Kaminsky, G. L. and Reinhart, C. M. (1999). The twin crises: the causes of banking and balance-of-payments problems. *American economic review*, 89(3):473–500.

Liu, H. and Motoda, H. (2007). *Computational methods of feature selection*. CRC Press.

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

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

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.

Reinhart, C. M. (2002). Default, currency crises, and sovereign credit ratings. *the world bank economic review*, 16(2):151–170.

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

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

Stock, J. H. and Watson, M. W. (2002b). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460):1167–1179.

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.

Zarnowitz, V. and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political economy*, 95(3):591–621.

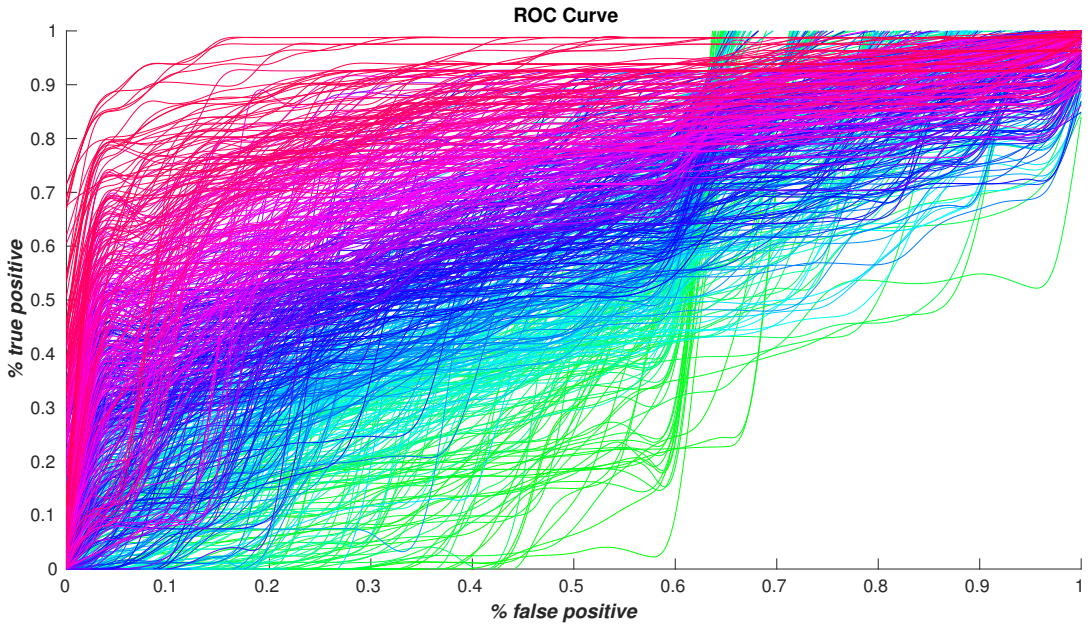
A Appendix — AUROC review

In a discrete context, as the case of forecasting inflation probability, the prediction exercise is closer to a classification problem. Therefore, given the estimated probabilities, a researcher assigns each forecast to the correct class $\{0, 1\}$. For this class of problems appropriate loss functions have been extensively studied, and, among these, a particularly well-tailored criterion for binary classification problems is the Area Under the Receiver Operating Characteristics (AUROC) function. The AUROC ranks models according to their ability to classify observations among the entire spectrum of cutoff points. In particular, in a first step, 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*) is evaluated for all possible thresholds $\mathcal{T}^i$. The set of thresholds is approximated by a discrete variable bounded between zero and one, as showed in equation (A.1).

$$\begin{cases} \hat{Y}_t = 0 \quad if \quad \hat{Y}_t < \mathcal{T}^i \\ \hat{Y}_t = 1 \quad if \quad \hat{Y}_t \geq \mathcal{T}^i \end{cases} \tag{A.1}$$

Where $\mathcal{T}^i \in [0, 1]$, $i = 1, 2, ..., I$. Therefore, in the first step, a researcher estimates a discrete dependent variable model and compare the predictions against a threshold $\mathcal{T}^i$. Depending on the threshold, each prediction can be classified, and the classification can be compared against the true one. Repeating this process for all possible $\mathcal{T}^i$ allows assessing the model classification ability. The result can be represented in a plane having the percentage of false positive and true positive on the x and y axis ($FP(\mathcal{T}^i)$, $TP(\mathcal{T}^i)$). The line connecting the points is called Receiver Operating Characteristics (ROC) function. Figure A.1 shows the ROC curve for different univariate probit models estimated with the discretized EA HICP inflation year-on-year as the dependent variable (as in appendix C), and all the variable in the dataset described in appendix B as regressors. The chart can be interpreted as follows; first, 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 better performing model—i.e., red curves. Secondly, 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 allowed to increase the rate of true positives. Third, in the ($FP(\mathcal{T}^i)$,$TP(\mathcal{T}^i)$) plane, the 45-degree line is a random guess equivalent, and it is often used as a reference line—i.e., 50% probability of having TP and FP. When the ROC curve is below the 45-degree line, a researcher should revert the classification scheme—i.e., green and light-blue lines in the

Figure A.1: ROC curve.



Note: 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.

figure. Reverting the classification scheme flips the curve symmetrically around the 45-degree line. Fourth, 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, as shown in equation (A.2).

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

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 AUROC ranks models from the one with the largest area to the one with the smallest, and it ranges from 0.5 to 1. The advantage of using the AUROC is that models can be evaluated without selecting an arbitrary threshold. Despite its proven ability, the AUROC has only been used in recent times in the economic literature (Berge and Jordá, 2011; Liu and Moench, 2016).

B Appendix — Data

I build a large dataset comprising around 100 monthly variables at the national and euro area level starting in January 1999 and ending in March 2017 (219 observations). Table B.1 shows

the complete monthly dataset, the respective identification codes and transformations. All the data are provided by Thomson Reuters Eikon and Datastream[4]. 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.

[4]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.

Table B.1: complete monthly dataset. Variable names, identification and transformation codes.

Variable	RIC/DS ID	Trans.	Variable	RIC/DS ID	Trans.
US10Y	US10YT=RR	1	HICP EA	aXZCPIHICP/C	3
EURIBOR3M	EURIBOR3MD=	1	IP EA	aXZCINDG/CA	3
EURIBOR6M	EURIBOR6MD=	1	Consumers good	aXZPDAGCGS/A	3
EURIBOR1Y	EURIBOR1YD=	1	Durable	aXZPDAGCDRB/A	3
DE stock	.GDAXI	3	Non durable	aXZPDAGCNDR/A	3
ES stock	.IBEX	3	Intermediate	aXZPDAGINTG/A	3
FR stock	.FCHI	3	Energy	aXZPDAGENE/A	3
IT stock	.FTMIB	3	Capital	aXZPDAGCAPG/A	3
DE2YT	DE2YT=RR	1	Construction	aXZIPCON/A	3
DE5YT	DE5YT=RR	1	Manufacturing	aXZIPMAN/A	3
DE10YT	DE10YT=RR	1	Unemploy. rate EA	aXZUNR/A	1
ES5YT	ES2YT=RR	1	Credit gen gov	aXZCRDGOV/A	3
ES10YT	ES10YT=RR	1	Car regist	aXZCRDRG/A	3
FR2YT	FR2YT=RR	1	Business climate	aXZBUSCLIM	6
FR5YT	FR5YT=RR	1	Consumer conf.	aXZECOSE	6
FR10YT	FR10YT=RR	1	Industrial conf.	aXZBSMFGCI/A	6
IT2YT	IT2YT=RR	1	Retail conf.	aXZBSSVRTCI/A	6
IT5YT	IT5YT=RR	1	Construction conf.	aXZBSCSCI/A	6
IT10YT	IT10YT=RR	1	Service conf.	aXZBUCFM/A	6
NL2YT	IE2YT=RR	1	Core CPI ea	aXZCCORF/C	3
NL5YT	NL5YT=RR	1	Eonia	aXZONIA	1
NL10YT	NL10YT=RR	1	M1	aXZM1	3
EA short repo	RC2AALM	1	M2	aXZM2	3
EA2Y	EMECB2Y.	1	M3	aXZM3	3
EA3Y	EMECB3Y.	1	Neer	aXZINECE/C	3
EA5Y	EMECB5Y.	1	US ff rate	aUSFEDFUND	1
EA7Y	EMECB7Y.	1	IP DE	aDECINDG/A	3
EA10Y	EMGBOND.	1	IP ES	aESCINDG/A	3
Loans nonfin	EMEBMC..A	3	IP FR	aFRCINDG/A	3
Loans hsld	EMEBMH..A	3	IP IT	aITCINDG/A	3
Loans non-mfi	EMEBMEO.A	3	Unemploy. DE	aDECUNPQ/A	1
Loans mfi	EMECBXLMA	3	Unemploy. ES	aESCUNPQ/A	1
IT CPI SA	ITCCPI..E	3	Unemploy. FR	aFRCUNPQ/A	1
IT core CPI SA	ITCCOR..E	3	HICP DE	aITUNRM/A	3
ES CPI SA	ESCCOR..E	3	HICP ES	aESHICP	3
DE CPI SA	BDCONPRCE	3	HICP FR	aFRHICP	3
DE CPI core SA	BDUSFG10E	3	HICP IT	aITHICP	3
FR CPI SA	FRCONPRCE	3	Core CPI DE	aDECCORF/C	3
FR core SA	FRCPUNDEE	3	Core CPI FR	aESCCORF/C	3
Price trends 12M	EMZEWCP.R	6	Core CPI IT	aITCCORF/C	3
Econ12M	EKTOT4BSQ	6	EA stock	.STOXX50E	3
Unemp.12M SA	EKTOT7BSQ	6	EA bank stock	.SX7P	3
REER	EMI..RECE	3	US stock	.SPX	3
US crude	USSCOP.BP	3	US vol	.VIX	3
Shadow	-	1	Crude	LCOc1	3

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.

C Appendix — Discrete inflation measure

In this section, I describe the process employed to build the dependent variable. In fact, the main difference concerning predicting inflation probabilities and more standard variables—e.g., predicting recession probabilities—is the choice of the dependent variable. Models tailored to predict the recession probabilities use a binary measure as the dependent variable. This measure is computed by independent research organizations which assess and release a discrete variable to track recession periods.[5] For inflation, a clear counterpart does not exist. Nevertheless, a very satisfying and intuitive alternative can be created by clustering inflation realizations above and below the central bank target. In particular, within the euro area, many inflation metrics exist. However, the ECB definition of the inflation target is in terms of year-on-year change in the EA *Harmonized Index of Consumer Prices* (HICP). There are other popular indicators[6]. However, as the paper focuses on forecasting inflation from a central bank viewpoint and the ECB target is in terms of HICP, I will only focus on this measure. To discretize the inflation measure and create the binary dependent variable (Y_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:

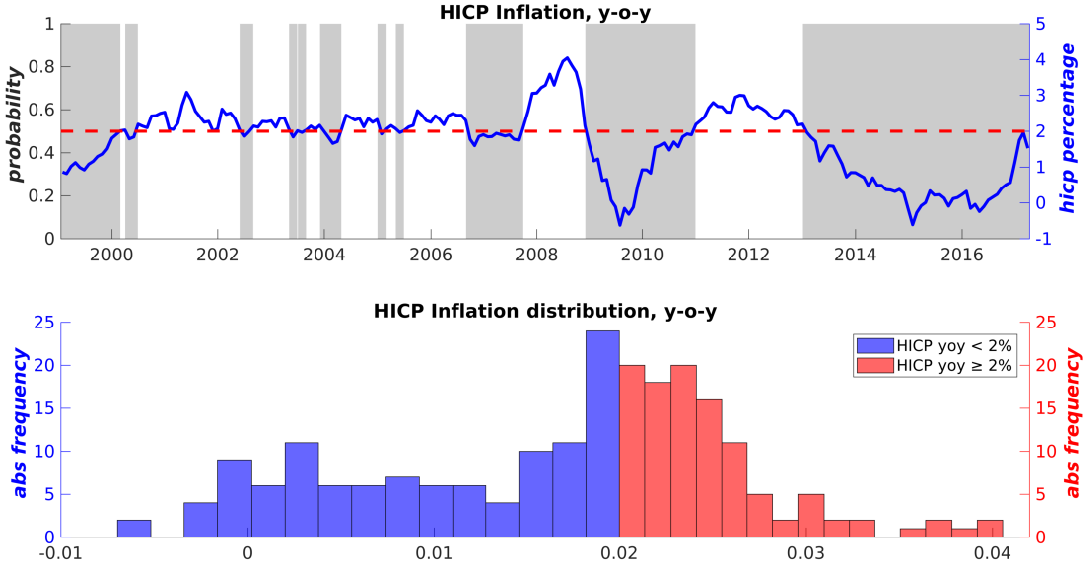
- Inflation below the confidence zone ($Y_t = 1 \; if \; \pi_t < 2\%$).
- Inflation above the confidence zone ($Y_t = 0 \; if \; \pi_t \geq 2\%$).

The first panel of figure C.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 shaded are highlight periods in which inflation is below the 2% level. By contrast, the others show periods in which HICP is above or equal to 2%. The second panel of figure C.1 shows the HICP sample distribution. The colors highlight the composition of the discretized HICP variable using 2% as a cutoff point. The blue bars (left-hand-side) show the portion of the distribution below the threshold. The red bars (right-hand-side) display the observations above or equal to it. The y-axis shows the absolute frequency of each bin. It is easy to notice that the binary variable for inflation is well balanced along the entire sample. It displays 112 observations below the threshold and 107 above. From the chart, it is easy to notice that the mass tends to locate

[5] 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.

[6] 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.

Figure C.1: EA HICP, year-on-year.



Note: 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.

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 since the great recession.