BANK OF ENGLAND

# Staff Working Paper No. 915

# Forecasting UK inflation bottom up

Andreas Joseph, Eleni Kalamara, George Kapetanios and Galina Potjagailo

March 2021

BANK OF ENGLAND

# Staff Working Paper No. 915
# Forecasting UK inflation bottom up

Andreas Joseph,[1] Eleni Kalamara,[2] George Kapetanios[3] and
Galina Potjagailo[4]

## Abstract

We forecast CPI inflation in the United Kingdom up to one year ahead using a large set of monthly disaggregated CPI item series combined with a wide set of forecasting tools, including dimensionality reduction techniques, shrinkage methods and non-linear machine learning models. We find that exploiting CPI item series over the period 2011–19 yields strong improvements in forecasting UK inflation against an autoregressive benchmark, above and beyond the gains from macroeconomic predictors. Ridge regression and other shrinkage methods perform best across specifications that include item-level data, yielding gains in relative forecast accuracy of up to 70% at the one-year horizon. Our results suggests that the combination of a large and relevant information set combined with efficient penalisation is key for good forecasting performance for this problem. We also provide a model-agnostic approach to address the general problem of model interpretability in high-dimensional settings based on model Shapley values, partial re-aggregation and statistical testing. This allows us to identify CPI divisions that consistently drive aggregate inflation forecasts across models and specifications, as well as to assess model differences going beyond forecast accuracy.

**Key words:** Inflation, forecasting, machine learning, state space models, CPI disaggregated data, Shapley values.

**JEL classification:** C32, C45, C53, C55, E37.

(1) Bank of England. Email: andreas.joseph@bankofengland.co.uk
(2) King's College London. Email: eleni.kalamara@kcl.ac.uk
(3) King's College London. Email: george.kapetanios@kcl.ac.uk
(4) Bank of England. Email: galina.potjagailo@bankofengland.co.uk

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to the participants of the 2020 Conference on Real-Time Data Analysis, the 2020 Conference on Methods, and Applications, Modelling with Big Data & Machine Learning, the 2020 Banca d'Italia and Federal Reserve Board Joint Conference on Non-traditional Data & Statistical Learning, the Office for National Statistics, and the participants of internal seminars for highly valuable comments.

The Bank's working paper series can be found at www.bankofengland.co.uk/working-paper/staff-working-papers

Bank of England, Threadneedle Street, London, EC2R 8AH
Email enquiries@bankofengland.co.uk

ISSN 1749-9135 (on-line)

# 1 Introduction

Forecasting inflation accurately in the near and medium term can have large implications for economic policy choices as well as business decisions in the wider economy. Central banks regularly publish consumer price inflation forecasts that form the basis for policy decisions and communications within their inflation targeting mandates. They increasingly base their forecasts on large and granular sets of indicators that provide a broad view on price dynamics across different sectors. From the academic side, advances in data availability and computing power have promoted forecasting methods that incorporate large sets of predictors as well as non-linear features such as time-varying parameters or, with increasing popularity, non-parametric machine learning tools.

In this paper, we use a unique set of 581 monthly disaggregated CPI item series, as well a set of 43 macroeconomic series, to forecast aggregate monthly CPI headline, core, and service inflation in the United Kingdom at horizons of 1-12 months ahead. We evaluate a wide range of forecasting methods that exploit the large data set in different ways: dimensionality reduction techniques (Dynamic Factor Model (DFM), Principal Component Analysis (PCA), partial least squares (PLS)), shrinkage methods (Ridge regression, Lasso, Elastic Net), as well as non-linear machine learning tools (Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forests). We consider the period 2011m2-2019m12 for which CPI items are available, evaluating the models in pseudo out-of-sample forecasts against an autoregressive benchmark. For comparison, we also consider an extended sample period 2000-2019 for which only macroeconomic predictors are available.

The contribution of the paper is three-fold. First, we employ a large set of disaggregated item-level predictors that the United Kingdom's Office for National Statistics (ONS) makes publically available. Why can we expect forecasting gains from considering item level information? Item-level prices (e.g. "cereal bar", "light bulb", "cinema admission") directly relate to the aggregate consumer price index. The ONS constructs aggregate CPI inflation from the item indices, that are themselves aggregations of price quotes collected in shops and centrally collected prices. But the dynamics and inter-dependencies of disaggregated price items are complex and the distributional moments of item indices do not necessarily translate linearly to the aggregate level. As such, prices of different items or sectors can behave asynchronously, the frequency and dispersion of price adjustments can vary across items and over time, and the characteristics of certain groups of items can be over-represented in the aggregate (Chu et al., 2018; Petrella et al., 2019; Stock and Watson, 2019). This suggests that by incorporating item indices directly into a flexible model the forecaster is able to exploit a rich set of information (Hendry and Hubrich, 2011). The use of disaggregated information can also help to communicate adjustments to forecasts based on dynamics observed in different sectors.

Second, we run a horse race between a wide range of forecasting models that represent different approaches to tackle the large dimension and high degree of disaggregation of our forecasting setup. We compare well-established linear approaches that deal with large data, such as factor models and shrinkage methods, with machine learning tools that are potentially stronger in detecting turning points and complex dynamics in the item data due to their flexibility to learn unknown functional forms.

Third, we provide a model-agnostic and flexible approach to address the "black box" critique of machine learning models while comparing the signals from a diverse set of models in a uniform manner. We measure the contribution of individual items, aggregated to the main CPI sub-sectors, to the forecast using Shapley values (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017), and we test the statistical significance of the sectoral components for individual predictions. These steps allow us to provide direct measures of forecast contributions of classes of CPI items and their statistical reliability for forecasting aggregate inflation.

Our findings are as follows. First, adding a large set of predictors into the forecaster's information set significantly improves forecast accuracy. Using CPI items or macroeconomic data for forecasting aggregate inflation in the UK strongly improves forecast accuracy at horizons of 3-12 months for all models, with improvements against an autoregressive benchmark of 30-70%. In this, disaggregated CPI item series predictors yield stronger gains than the macroeconomic predictors for the sample period 2011-2019 for all models except the DFM. Second, Ridge regression performs well across horizons, targets and specifications. The Ridge is particularly strong when combined with CPI items and over the 2011-2019 sample period: it yields improvements against the benchmark of 20-30% at short horizons and of 55-70% at horizons of 9-12 months. But also the PCA, PLS and the Lasso perform well, whereas the DFM model performs poorly when fed with CPI item series alone, but much better when macroeconomic series are added to the information set. Third, SVM and the ANN show the best forecast accuracy for headline inflation over the longer sample period 2000-2019, for which only macroeconomic predictors are available. The two models capture turning points and episodes of changes in volatility in aggregate inflation dynamics particularly well throughout the extended sample.

Our findings contribute to the debate in the forecasting literature on the use of sparse versus dense models. Whereas we do not discuss formal definitions of sparsity, which are not straight-forward in the non-linear machine learning setting, we can derive considerations from our results in terms of comparing results from different models. For specifications that use macroeconomic predictors, we confirm previous findings of superiority of dense models for forecasting accuracy (Giannone et al., 2017). However, when using disaggregated CPI items as predictors, sparse methods such as the Lasso are among the best performing models too. According to our results, efficient shrinkage and penalisation of the large information set is particularly relevant when exploiting disaggregated

CPI items, but whether the shrinkage method is a sparse or a dense one is secondary. On the other hand, exploiting dynamic co-movement in the data, as it is done by the DFM, appears to work less well with CPI item data compared to the macroeconomic context. Many CPI items show discrete jumps and non-standard dynamics and some of those co-move, without their common component being reflected in aggregate inflation dynamics. With such data at hand, it helps to downgrade this information directly through shrinkage instead of modelling co-movement dynamically.

Finally, our findings provide insights on the forecasting performance of linear models compared to non-linear, non-parametric machine learning tools. In our analysis, machine learning models have the highest forecast accuracy for headline inflation over the longer sample period, where they are particularly strong in capturing turning points and changes in inflation dynamics, such as the"missing disinflation" during the Great Recession, early on at the 9-12 months forecasting horizon. This makes them a promising tool for crisis episodes in particular. This said, Ridge regression performs almost as well as the machine learning tools over the longer sample period, while it beats them over the shorter period. The better interpretability of Ridge regression makes it a relevant forecasting model for central banks and institutions that aim to forecast inflation with large datasets.

Our analysis relates to various strands of the forecasting literature. A vast literature focuses on forecasting inflation using a wide range of approaches such as Philips curve-based models (Stock and Watson, 1999, 2008), univariate unobserved component models (Stock and Watson, 2007), aggregation of forecasts of sub-components (Hubrich, 2005), Bayesian VARs (Koop, 2013; Domit et al., 2019), dimensionality reduction (Kim and Swanson, 2018) and medium-sized DSGE models (Carriero et al., 2019), to name a few. Dynamic factor models have been highly successful for nowcasting and forecasting GDP due to their capacity to incorporate a large information sets and deal with ragged-edged data (Giannone et al., 2008; Bańbura et al., 2013; Thorsrud, 2018). For inflation, fixed-parameter dynamic factor and FAVAR models tend to show smaller and less stable improvements relative to simple benchmarks or compared to univariate unobserved component models (Faust and Wright, 2013). Stock and Watson (2016) find that a time-varying dynamic factor model that extracts a multivariate trend from 17 components of US inflation provides sizeable forecast gains for aggregate inflation.

With regard to machine learning tools, earlier studies find that inflation forecasts with neural networks outperform linear autoregressive benchmarks at different horizons (Chen et al., 2001; McAdam and McNelis, 2005; Nakamura, 2005). More recently, Almosova and Andresen (2019) find that a long short-term memory recurrent neural network strongly outperforms a random walk model, but also seasonal or Markov switching autoregressive models and a fully-connected neural network in forecasting US CPI inflation. Closer to our approach, Garcia et al. (2017) and Medeiros et al. (2019) forecast Brazilian and US CPI inflation, respectively, using large sets of macroeconomic predictors within shrinkage methods, factor models, ensemble forecasts, and a Random Forest. For the US, the Random Forest is found to perform best. Compared to these papers, we use a wider range of non-linear machine learning tools and we add a large set of disaggregated CPI items as predictors.

Our analysis also relates to a rather small set of studies that have used disaggregated data to forecast aggregate series. Hernández-Murillo and Owyang (2006) and Owyang et al. (2015) use US state-level data to forecast national-level GDP while accounting for spatial interactions between the states, finding forecast gains relative to aggregate predictors. Hendry and Hubrich (2011) show that adding disaggregated sector-level information into forecast models improves forecast accuracy for aggregate US inflation. Aparicio and Bertolotto (2020) use combinations of high-frequency online price item series to forecast CPI one to three months ahead in ten advanced economies; their forecasts outperform benchmark models as well as surveys of forecasters by anticipating changes in official inflation rates. Closest related to our approach, Ibarra (2012) uses a factor model based on 243 CPI item series and 54 macroeconomic series to forecast aggregate CPI in Mexico, reaching a forecasting performance comparable to forecasts from expert surveys. Our analysis for the UK includes a larger set of CPI item series and a wider range of forecasting approaches to extract information from the data.

While our study connects to the literature on using machine learning and disaggregated data for forecasting, we are also the first to address the black box problem resulting from both the opaqueness of non-linear models and the high dimensionality of the input space. Shapley value-based inference provides a general and principled approach to extract signals of our models (Joseph, 2019; Bluwstein et al., 2020). The method decomposes each prediction into linear contributions from individual predictors, aggregated into classes, and tests for the statistical association between these classes and the aggregate inflation value. This enhances the interpretability of results, as it pins down for each model the most relevant item classes for prediction within a large-dimensional setting. Results indicate that predictive shares from CPI item classes for Random Forest predictions are overall comparable to those for Ridge prediction, despite the different nature of those models. But Ridge regression extracts information from a wider range of CPI item classes in terms of evenly distributed component shares and a larger number of significant components across forecast horizons.

The remainder of the paper is organized as follows. Section 2 describes the data used in the forecasting exercise and, in particular, introduces the CPI item series data set of predictors. Section 3 describes the forecasting set-up and gives a brief model overview. Section 4 presents the results and sensitivity checks. Section 5 addresses the black-box critique to our high-dimensional forecasting setting through Shapley value-based inference. Section 6 concludes. Supplementary material is given in the Appendix.

# 2 Data

We use the headline CPI index from the ONS, transformed to year-on-year inflation rates, as the main target variable in our forecasting exercise. Additionally, we compute year-on-year CPI core inflation that is based on the CPI headline index excluding the generally more volatile food and energy components. The third target, year-on-year CPI core service inflation is based on CPI indices of twelve service categories only, excluding goods and more seasonally volatile services.[1] Core inflation and service inflation are relevant targets since they represent the less volatile component of consumer prices, and are typically considered to be more closely linked to underlying and domestically generated price pressures, and therefore closely monitored by the Bank of England and other policy institutions. For prediction of the aggregate CPI series, we use a large set of CPI disaggregated item series published by the Office for National Statistics (ONS), which we describe in more detail below.

Additionally, we explore the content of a set of 43 macroeconomic series, selected to represent broad categories of economic and financial activity: unemployment and hours, real measures for retail trade, manufacturing and sales, international trade, labor costs, house price indexes, interest rates, stock market indicators, and foreign exchange measures for the UK economy. Several studies have shown their predictive power of such macroeconomic data sets in forecasting inflation (Stock and Watson, 2002a,b). The data also have the advantage of being readily available over longer sample periods and being continuously monitored by central banks and economists. Prior to estimation, the series are transformed to year-on-year log differences to achieve stationarity and are standardised (see Table B1 in Appendix B).

## 2.1 CPI item series

The consumer price index (CPI) measures the price of consumption goods according to the household expenditure on a representative basket of goods relative to a base date. Changes in CPI, i.e. price inflation, are a guide for changes in households' living costs. While the CPI and price inflation are both macroeconomic concepts, they are constructed from the prices of single items over time, i.e. prices observed through local collection in physical shops or online or central collection in case of national prices. That is, item prices connect the micro and macroeconomic level, which we exploit in this paper.

The UK CPI is constructed by the Office for National Statistics from an evolving set of 700 representative monthly item indices, weighted according to household expenditure patterns. At the lowest level, single item prices, or price quotes, are aggregated into

[1] The twelve services categories are household, health, miscellaneous, financial, accommodation, catering, recreational, communication, other housing, other transport, other services for personal transport equipment. Prices of airfares, package holiday, and education and rents since prices in these sectors tend to be volatile and have strong seasonal pattern.

item-level indices.[2] The item indices combine prices of products corresponding to an item using equal weights. For further aggregation, the items are weighted according to a representative consumption basket to produce prices of classes, groups, divisions, and finally the CPI based on the Classification of Individual Consumption according to Purpose (COICOP), an international classification framework.

Starting from 2011, all item indices are made publicly available by the ONS with one month publication delay. We use these data for the period February 2011 until December 2019. There are overall 878 items, of which 700 are used to construct the aggregate index at any point in time. This dynamic evolution of the CPI index implies that there are missing values for a number of item series. We drop series with missing values, which leaves us with 581 indices that we use to forecast CPI headline inflation. When forecasting core and service CPI inflation, we drop 162 CPI items that correspond to fuel and energy since these volatile components are also not included in the targets, leaving 419 series for these specifications.[3]

We chain-link the item indices, and we compute year-on-year log differences, which removes stochastic seasonality and smooths extreme observations through the log transform.[4] Item series are mean-variance standardised in line with the expanding window approach of our forecasting setting described in Section 3. Figure B1 in Appendix B plots a selection of the transformed item series for illustration.

## 2.2 Descriptive statistics

To better understand how the item series dynamics compare to the aggregate CPI index, we provide descriptive statistics for the disaggregated data we use. First, Table 1 assesses the representativeness of our sample of item series. It summarises statistics of year-on-year item-level index changes grouped by divisions, the twelve largest sub-categories of the CPI index using the final release classification (December 2019; see also (ONS, 2019)). Our item selection is very much in line with the full set of index series for all divisions. The series cover on average 84% of the item indices in each division.[5] The latter two columns show the mean and standard deviation of yearly changes of our chained-linked index series. The mean across items for most CPI divisions is comparable to the average

[2] A detailed description of the collection of prices and the construction of CPI is given by ONS (2019).

[3] We also exclude items for January 2011 due to relatively low coverage. A list of items used in our analysis are available upon request.

[4] The raw data do not come chain-linked, but are rather expressed relative to the January level for each year (since 2018: relative to the December level of the previous year).

[5] Numbers of all series by divisions are not integers due to series dropping in and out over time. A set of zero-weight indices not in the CPI have been added to Housing & Fuel (440249, 410201, 410701, 410703, 410801, 440202, 610307, 610308). Narrowing down the overlapping sampling period to three to five years and using a sliding window increases the number of series by only about 50, or less than 10%, at every point in time. Such an approach would be useful, however, when investigating the dynamic structure of the index on a longer sampling period, which is beyond the scope of the current study.

Table 1: Summary statistics of filtered UK CPI inflation item indices.

| | division | weight | # total | # included | coverage | mean | SD |
|---|---|---|---|---|---|---|---|
| 1 | Food & non-alc. bev. | 10 | 159.8 | 129 | 81 | 0.66 | 6.98 |
| 2 | Acl. bev. & tobacco | 4 | 27 | 20 | 74 | 1.13 | 4.68 |
| 3 | Clothing & footwear | 7 | 76.7 | 71 | 93 | 1.68 | 5.18 |
| 4 | Housing & fuels | 12 | 30.2 | 30 | 83 | 1.57 | 5.20 |
| 5 | Furnishing & house maint. | 7 | 70.8 | 55 | 78 | 1.29 | 4.53 |
| 6 | Health | 3 | 18.6 | 19 | 92 | 1.84 | 3.92 |
| 7 | Transport | 15 | 41.3 | 36 | 87 | 2.12 | 6.03 |
| 8 | Communication | 2 | 10.9 | 9 | 83 | 2.12 | 13.05 |
| 9 | Recreation & culture | 16 | 115.4 | 92 | 80 | 1.51 | 6.80 |
| 10 | Education | 1 | 3 | 3 | 100 | 10.48 | 8.18 |
| 11 | Restaurants & hotels | 12 | 51.3 | 44 | 86 | 2.49 | 1.80 |
| 12 | Misc. goods & services | 10 | 79 | 73 | 92 | 0.91 | 5.31 |
| – | Total | 100 | 684 | 581 | 84 | 2.32 | 5.97 |

Notes: Division-level summary statistics of year-on-year percentage changes of item series. CPI weights (%) are taken from COICOP weights for December 2019. The total number (#) refers to all item series available between February 2011 until December 2019 in that division. Note that this number does not need to be an integer because of items either entering or exiting the CPI basket. The number of included series are those reported throughout the whole period which enter our models. Coverage (%) is the fraction of our modelling set of series of all series. Mean and standard deviations (SD) are taken over all observations of all series in a division. Source: ONS & authors' calculation.

aggregate year-on-year price inflation, while the standard deviations around these values are relatively large.[6]

This can also be seen in Figure 1, which shows the distribution of item-level index changes for our sample. As previously documented (Klenow and Kryvtsov, 2008; Ozmen and Sevinc, 2011), micro price changes have a leptokurtic shape with a sharp peak and wide tails on both sides. That is, while most items do show only small price changes, some show very large changes. The mean item-level index change is close to average price inflation. A slight difference here to other studies is that we look at chained index series and not individual item price quotes. The latter's price changes are often centred around zero and hence more difficult to compare to aggregate price changes. In line with Table 1, average headline inflation (vertical red line) is close to the mean index change.

Finally, Figure 2 shows the evolution of the mean, median and standard deviation of year-on-year changes of all item-level index series in our sample, relative to changes in the overall CPI (blue line). The dynamics of the mean and median index values again are in line with with the overall index. The standard deviation of the changes in chained item-level indices changes (red line) does not show the same yearly patterns as the raw index data would do (purple). Chain-linking additionally removes large spikes in the data related to the yearly re-basing of raw index data each January (see (ONS, 2019)), and as such addresses two major issues in the raw data, while aligning our micro statistics with

[6] Education is an outlier here. However, its small weight and low number of series is not seen to affect our results.

the aggregate inflation rate indices, the forecast targets.

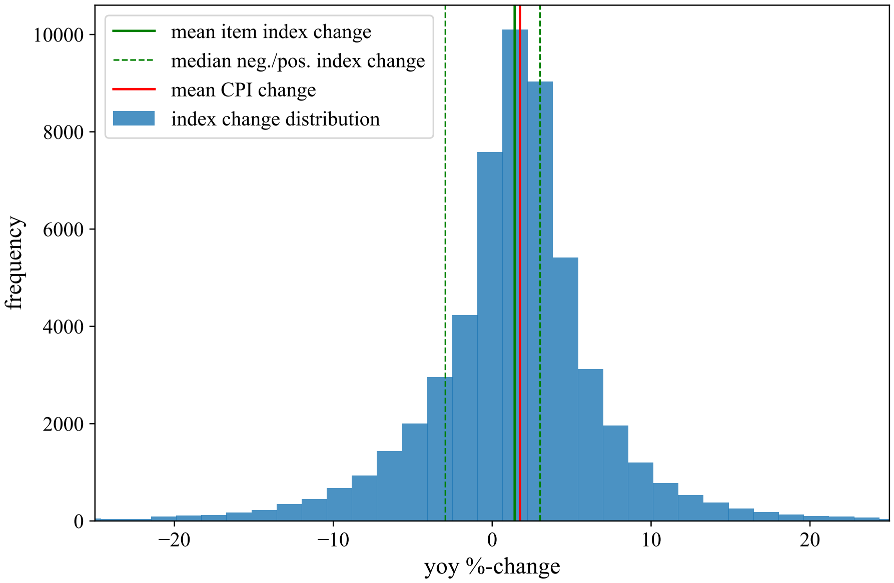

Figure 1: Distribution and moments of item-level CPI indices.
Notes: Chain-linked item series, distrobution of year-on-year changes for 581 selected items (blue bars). The depicted changes are limited to ±25% for clearer presentation with a small number of changes beyond this range. The solid green line shows mean year-on year changes in CPI items, whereas the green dashed lines distinguish between mean negative and mean positive changes. The red solid line shows the mean year-on-year CPI inflation for comparison. Source: ONS and authors' calculation.

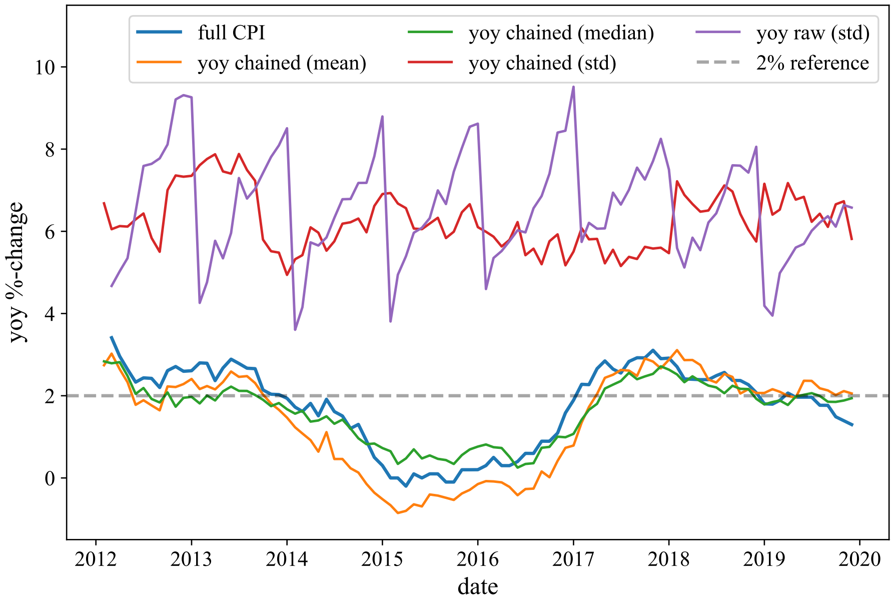

Figure 2: First and second moments of item-level CPI indices over time.
Notes: Mean (orange line), median (green line), and standard deviation (red line) of year-on-year changes of chain-linked item indices. For comparison: standard deviation of raw item indices (violet line), aggregate CPI year-on-year inflation (blue line). for selected item sample and raw indices. Source: ONS and authors' calculation.

# 3 Methodology

## 3.1 Forecasting set-up

We forecast monthly aggregate year-on-year CPI inflation series (headline, core, and core services) for the UK over forecasting horizons of $h = 1, \ldots, 12$ month. We use a large set of predictors including 581 CPI item series and/or 43 macroeconomic series. The sample period is 2011m2 to 2019m12, based on the availability of CPI item data.[7] In an alternative specification, we use macroeconomic predictors only for an extended period from 2000m1 to 2019m12. We run a recursive out-of-sample forecasting exercise with an expanding window. We estimate the model and tune hyperparameters over the initial period 2011m2 to 2015m4, i.e. 50 months, and evaluate forecasts over the remaining period.[8] As we move along the evaluation period, the training sample is sequentially extended one month at a time and the model is retrained. This makes sure that the latest available observations for each forecast are used and that as many observations as possible are exploited for training.[9]

Our benchmark model is an $AR(p)$ forecast which only accounts for lagged dynamics of the target variable, of the form

$$\hat{y}_{t+h} = \hat{\alpha} + \Sigma_{j=1}^{p} \hat{\gamma}_j y_{t-j+1} \tag{1}$$

where $y_t$ is the target variable, $h$ is the forecast horizon, and the number of lags $p$ is set to be maximum twelve and selected using the Bayesian Information Criterion (BIC). We evaluate the average precision of the forecasts from a range of models outlined below against the AR(p) benchmark, as well as against two of the dimensionality reduction models described below, based on relative root mean squared errors (RMSE). We test for statistical difference in forecast accuracy using the Diebold and Mariano (1995) test with Harvey's correction for short samples (Harvey and Newbold, 2000).

## 3.2 Overview of forecasting methods

All the forecasting methods we present here have the advantage that they can deal with large datasets and thus wider information sets. In particular, a large set of of predictor

[7] For the time being, we do not include the Covid-19 shock into our analysis. Such an analysis faces various challenges, including the unprecedented size of the shock and the fact that data for a substantial part of the CPI basket were not collected by the ONS during the first lockdown period in 2020. However, non-linear machine learning tools are a promising tool in capturing non-linearities and large shocks such as the Covid-19 episode, as discussed in section 4.3.

[8] Results remain very similar when the initial tuning period is extended until 2016m4 at at the expense of a shorter out-of-sample period. These results are available upon request.

[9] CPI aggregate and item series typically are not revised after first publication. Since the focus of this study lies in using CPI item series as predictors, we do not account for real-time data issues with macroeconomic data, and we use the final data release.

series $x_t$ used to forecast the target variable $y_t$.

We are given a dataset of a large number of predictors $x_t = (x_{1t}, \ldots, x_{Nt})\prime$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, and we are interested forecasting $y_t$ in period $t + h$, based on the past dynamics of $y_t$ and the set of predictors $x_t$[10]

$$\hat{y}_{t+h} = \hat{\alpha} + \Sigma_{i=1}^{N} \hat{\beta} x_t + \Sigma_{j=1}^{p} \hat{\gamma}_j y_{t-j+1}. \tag{2}$$

Estimating (2) directly with linear methods would go hand in hand with very high estimation uncertainty and a lack of degrees of freedom in typically relatively short sample periods, where the dimension of $\beta$ might be larger than $T$. Such a model thus suffers from over-parametrization, and some form of dimensionality reduction is required. The models we employ take different approaches to deal with this, either by reducing the dimensionality of the input space directly or via explicit or implicit weighting (shrinkage). The main ideas of the models are outlined below, with details presented in Appendix A.

### Dimensionality reduction techniques

These methods exploit the fact that economic series are often strongly correlated and thus can be summarized effectively in a small set of common components. This substantially reduces the number of parameters in the model, addressing over-parametrisation and degrees of freedom issues in rather short samples. In particular, a vector of $N$ indicator series $x_t$ is summarized by a vector $r \times 1$ of finite latent components $f_t$. We consider three approaches: Principal Component Analysis, Partial Least Squares, and a Dynamic Factor Model in state space form. Principal Component Analysis summarises the joint variability of predictors $x_t$ into a static factor which is added into a prediction regression as in equation 2. The Dynamic Factor models the dynamics of such a common component explicitly and provides estimates of the factors together with predictions via a state space representation of the model and using the Kalman filter. Finally, Partial Least Squares is a static model that combines predictors into a common component such that the covariance between the component and the target variable $y_{t+h}$ is maximised. All these models have in common that they use information densely, i.e. information from a wide range of available predictors is drawn upon by summarising them through common components.

### Shrinkage methods

The goal of shrinkage methods is, using different penalisation schemes, to reduce the dimension of the matrix of indicator series $x_t$. This produces linear combinations of the original regressors, where those coefficients that do not carry any predictive power

[10] We also experimented with including lags of the predictors $x_t$ into the models. Forecasts did not improve substantially, but estimation time increased considerably due to the larger number of parameters in models with lagged predictors. We therefore opt for a specification without lagged predictors.

for the target variable are assumed to approach zero or are set equal to zero, according to a shrinkage parameter $\lambda$, which differs across models. The Ridge regression shrinks the coefficients of predictors that contribute little to the predictive ability of the model towards zero, albeit they never become exactly zero—it is therefore a dense model which draws on all available information albeit to different degree. In the case of no shrinkage, i.e. $\lambda = 0$, Ridge regression becomes equivalent to a linear OLS regression. The Lasso regression, on the other hand, penalises the sum of squared residuals according to the sum of absolute coefficients which results in some of the coefficients being shrunk to exactly zero. It is thus a sparse model which performs shrinkage through variable selection. The Elastic Net is a hybrid approach which combines these two types of shrinkage: in a first step, it finds Ridge regression coefficients and, in a second-step, Lasso-type shrinkage i.e. variable selection is applied. A correction factor is applied to account for increased bias through double shrinkage.

### Non-Linear Machine Learning Models

The non-linear machine learning models that we use can be summarized as

$$\hat{y}_{t+h} = g\left(z_t, \beta^0\right) + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2) \tag{3}$$

where $\hat{y}_{t+h}$ are h-step ahead inflation forecasts, $z_t = [x_t, \Sigma_{j=1}^{p} y_{t-j+1}]$ is the set of $M = N + p$ predictors and lagged variables, $\beta^0$ is a $M \times 1$ vector of parameters, and $\varepsilon_t$ a vector of identically distributed errors with zero mean and variance $\sigma^2$. The relationship between the data matrix $z_t$ and the target $\hat{y}_{t+h}$ is captured by a non-linear matrix-valued function $g(\cdot)$ that varies with the model at hand. We use three types of machine learning models: Random Forests, Artificial Neural Networks, and Support Vector Machines.

Random Forests are collections of many decision trees, which in turn consecutively split the training dataset until an assignment criterion with respect to the target variable into a "data bucket" (leaf) is reached. The algorithm minimises the objective function within areas of the target space, i.e. these "buckets", conditioned on the input $z_t$. By averaging predictions over tree ensembles, random forests reduce the problem of overfitting by reducing the variance of model prediction, and typically performs better compared to individual trees. Tree models are mostly sparse as their hierarchical structure acts like a filter. That is, only variables which actually improve the fit are chosen during construction of each tree during training.

Artificial Neural Networks (ANN) consist of an input layer, at least one hidden layer, and an output layer. Layers are connected via the network weights $W$ representing the model parameters and pass through non-linear activation functions at each hidden layer. Note that, without hidden layer, an ANN becomes a linear function and is similar to solving the least squares problem. We use multilayer perceptrons (MLP), a form of feed-forward network, as ANN architecture. The activation function $g(z_t, W)$ acts as a gate for signals and introduce non-linearity into the model. Its functional form is subject to hyperparameter tuning. The variables $z_t$ in the input layer are multiplied by weight matrices $W$ at each layer, then transformed by an activation function in the hidden layers and passed on through the network until the linear output layer is reached resulting in a prediction $\hat{y}_{t+h}$. Deeper networks are generally more accurate but also require more data to train them due to the larger number of parameters in the weight matrices. The number of hidden layers, i.e. the depth of the network, and the number of neurons in each layer as well as appropriate weight penalisation in our ANN are hyper-parameters, and are determined by cross-validation as discussed below.

Support Vector Machines (SVM) identify a (small) set of training points, the support vectors, to either represent a boundary between classes (classification problem) or a line (regression problem). This representation becomes non-linear through the use of kernels for the joint processing of test observations in conjunction with the support vectors. We use the popular Gaussian kernel (radial basis function, RBF). Penalisation is introduced by allowing some wiggle room in situation where best fit lines or classification boundaries cannot be perfectly represented by the support vectors (see e.g. Friedman et al., 2001).

## 3.3 Tuning of hyperparameters

All of our models require some form of hyperparameter selection prior to estimation. In the case of dimensional reduction techniques, we use a form of information criteria (e.g AIC, BIC) to choose the lag length or number of common components. In cases where the derivation of information criteria is not feasible, such as the shrinkage methods and machine learning tools, we use cross-validation procedures.[11] Here, the strength of regularisation parameters, the number of nodes and layers for the ANN, or the choice of the kernel function for SVM is chosen among others. The main difference between information criteria and cross-validation methods is that the latter depends on out-of-sample performance, whereas information criteria are "in-sample" statistics.

K-fold cross-validation involves the assumption that samples are independent and identically distributed which results in unreasonable correlation between training and testing instances in the time series context. We therefore opt for a variant of K-fold cross-validation where the model is evaluated on "future" observations least like those that are used to train the model.[12] In each fold, test indices must be higher than before. We split the in-sample data in $k = 5$ folds as the train set and the $k + 1$-th fold as test set. This is consistent with our expanding window evaluation of the out-of-sample test forecasts. As a performance metric, we consider the average mean squared error over the test set.

[11] For a review of various cross-validation methods see Coulombe et al. (2019)

[12] We use the python package *TimeSeriesSplit* to perform time series cross-validation: it provides train/test indices to split data samples that are observed at fixed time intervals into train/test sets.

Table 2 reports the best performing tuning parameters selected by cross-validation for headline inflation forecasts based on different data specifications.[13] The hyper-parameters selected through cross-validation include the penalty imposed on shrinkage methods but also the maximum depth of trees for the Random Forest, the architecture of the ANN and the choice of the kernel function for SVM. Given that the estimation is done on a monthly basis with an expanding window, the values are the best parameters selected on the last, and longest, training period.[14] Overall, cross-validation favours quite similar parameters and model architectures across different data specifications. Differences mostly appear regarding the architecture of the ANN and the Random Forest. In particular, for larger data specifications, e.g. when we use both CPI items and macroeconomic predictors, the procedure selects deeper versions of the network and larger tree structures of the Random Forest. This suggests an increase in complexity as more data are involved in training the model. Regularisation seems to play an important role both for the linear and non-linear models. Notably, Ridge regression always imposes a heavy penalty which might explain the overall strong performance of this model.

Table 2: Hyper-parameter selection from cross-validation exercise.

| Model | Parameter | CPI items only | CPI items + Macro | Macro only | Macro only (long sample) |
|---|---|---|---|---|---|
| Ridge | $\alpha$ | 1.0 | 1.0 | 1.0 | 1.0 |
| Lasso | $\alpha$ | 0.1 | 0.1 | 0.01 | 0.1 |
| Elastic | $\alpha$ | 1.0 | 1.0 | 1.0 | 1.0 |
|  | L1-ratio | 0.1 | 0.1 | 0.1 | 0.1 |
| Forest | max. depth | 3 | 5 | 5 | 3 |
| ANN | activation | tanh | tanh | tanh | tanh |
|  | $\alpha$ | 1.0 | 1.0 | 1.0 | 1e-05 |
|  | hidden layer dim. | (10, 2) | (2, 3) | (10, 2) | (5, 2) |
| SVM | $C$ | 100 | 100 | 100 | 100 |
|  | $\epsilon$ | 0.5 | 0.5 | 0.5 | 0.5 |
|  | kernel | RBF | RBF | RBF | RBF |

Notes: For neural networks, the first component of "hidden layer dim." refers to the number of nodes $N_h$ and the number of hidden layers $L$, respectively. Thus, a network (10, 2) includes $L = 2$ hidden layers and $N_h = 10$ nodes in each. Grid sets: $\alpha \in \{1e-05, 0.0001, 0.001, 0.01, 0.1, 1.0\}$, L1-ratio $\in \{0.1, .5, .9, .95, 1\}$, max. depth $\in \{1, 2, 3, 5, 6, 7, 8, 9, 10\}$, hidden layer dimension $\in \{(2, 3), 10, 2), (20, 2), (2, 3), (20, 3), (5, 5)\}$, $C \in \{100, 10, 1000\}, \epsilon \in \{0.01, 0.1, 0.5, 0.9\}$, the activation function is chosen to be tanh or ReLU.

# 4 Results

We present the results of the forecasting exercise focusing on relative root mean squared errors (RMSE) and predicted value comparisons across the different forecasting models,

[13] Hyper-parameter choices are very similar for CPI core and service inflation targets and are available upon request.

[14] Results remain very similar when we look at the mode of the best parameter distribution over all training samples suggesting, a consistent selection of hyperparameters across estimation periods.

horizons and targets. We start with results for specifications where CPI items are included as predictors, either alone or in combination with macroeconomic series. These results are based on the sample period 2011 to 2019, in line with the availability of CPI item data. We then also present results using macroeconomic predictors only, where we can consider a longer sample period starting in the year 2000.

We show relative RMSE against the AR($p$) benchmark that only accounts for the lagged dynamics of the target. The comparison with the AR benchmark provides information on the improvement in the forecast through the inclusion of a large set of predictors. It also indicates a ranking between models in terms of the extent to which they outperform or lose out against the AR. The AR benchmark can be easy to beat, however, since it does not use any additional information apart from that in the time structure of the target itself. We therefore also present RMSE relative to the PCA and the DFM. This emphasizes how shrinkage and machine learning models that all exploit the same large information sets fare against quite standard dimensionality reduction methodologies. We assess the significance of forecast accuracy comparisons using Diebold and Mariano (1995) test statistics with Harvey's adjustment.

## 4.1 Forecast results using CPI item predictors

Table 3 shows results on forecasting accuracy in terms of relative RMSE for the specification with CPI item series used as predictors. Table 4 shows results for the specification where we include CPI items plus the set of macroeconomic series. Also, Table B2 in the appendix shows results of a specification only using macroeconomic series over the same sample period, 2011-2019. The two latter specifications allow us to check whether the disaggregated CPI item series are relevant predictors beyond the predictive power from macroeconomic dynamics, and whether certain models perform better when fed with different types of predictors. The tables are organised in six panels. The three horizontal panels show different benchmarks, i.e. RMSE relative to the AR benchmark (upper panel), relative to the PCA as benchmark (middle panel), and relative to the DFM as benchmark (lower panel). The vertical panels show results for different targets. The left panel shows relative RMSE for forecasts of CPI headline inflation one, three, six, nine, and twelve months ahead. The middle and right panels show the corresponding results for forecasts of CPI core inflation and service inflation one, six, and twelve months ahead.[15]

Various observations emerge. First, the autoregressive benchmark is clearly outperformed by most models at horizons of 3 months or higher for all three targets and specifications. Thus, adding a large set of predictors into the forecaster's information set significantly improves accuracy, independently of the model. Notably, exploiting CPI items provides stronger forecasting gains compared to macroeconomic series for all models (i.e.

[15] We show only these three horizons for core and service inflation due to space constraints and since results do not vary much across horizons for these two targets.

Table 3: Forecasting exercise results, CPI items series predictors.

**Benchmark: AR(p)**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| PCA | 0.91 | 0.73** | 0.56*** | 0.44*** | 0.41*** | PCA | 0.93 | 0.63** | 0.56*** | PCA | 0.89 | 0.64*** | 0.57*** |
| DFM | 1.04 | 0.97 | 0.85 | 0.72** | 0.8* | DFM | 1.09 | 1 | 0.82 | DFM | 1.18* | 1.11 | 1.23 |
| PLS | 0.76** | 0.57*** | 0.46*** | 0.41*** | 0.37*** | PLS | 0.89 | 0.56*** | 0.46*** | PLS | 0.84** | 0.59*** | 0.53*** |
| Ridge | 0.77 | 0.52*** | 0.37*** | 0.29*** | 0.29*** | Ridge | 0.73* | 0.52*** | 0.4*** | Ridge | 0.75** | 0.47*** | 0.44*** |
| Lasso | 0.77* | 0.59*** | 0.52*** | 0.42*** | 0.37*** | Lasso | 0.89 | 0.56*** | 0.47*** | Lasso | 0.77** | 0.6*** | 0.48*** |
| Elastic | 1.2 | 0.82 | 0.61** | 0.49*** | 0.44*** | Elastic | 1.01 | 0.59*** | 0.66*** | Elastic | 0.88 | 0.55*** | 0.5*** |
| SVM | 1.96*** | 1.22** | 0.85 | 0.71*** | 0.69*** | SVM | 1.22** | 0.68*** | 0.61*** | SVM | 1.03 | 0.72*** | 0.68*** |
| Forest | 1.18** | 0.83* | 0.57*** | 0.58*** | 0.57*** | Forest | 1.18 | 0.71*** | 0.62*** | Forest | 1.07 | 0.73*** | 0.69*** |
| ANN | 1.26 | 0.71* | 0.51*** | 0.31*** | 0.39*** | ANN | 1.21 | 0.6** | 0.6** | ANN | 0.95 | 0.55*** | 0.65** |

**Benchmark: PCA**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.09 | 1.36** | 1.78*** | 2.3*** | 2.44*** | AR | 1.07 | 1.6** | 1.79*** | AR | 1.12 | 1.57*** | 1.74*** |
| DFM | 1.14 | 1.32* | 1.51** | 1.65*** | 1.96*** | DFM | 1.17** | 1.6*** | 1.46*** | DFM | 1.32** | 1.73*** | 2.15*** |
| PLS | 0.84** | 0.78*** | 0.83* | 0.94 | 0.91 | PLS | 0.96 | 0.89 | 0.82*** | PLS | 0.94 | 0.93 | 0.93 |
| Ridge | 0.84 | 0.71*** | 0.66*** | 0.67*** | 0.71*** | Ridge | 0.79* | 0.82 | 0.71*** | Ridge | 0.84 | 0.73** | 0.78** |
| Lasso | 0.84 | 0.8** | 0.94 | 0.96 | 0.89 | Lasso | 0.96 | 0.9 | 0.84 | Lasso | 0.87 | 0.93 | 0.84 |
| Elastic | 1.31** | 1.12 | 1.09 | 1.13 | 1.08 | Elastic | 1.09 | 0.93 | 1.18 | Elastic | 0.98 | 0.85 | 0.87 |
| SVM | 2.14*** | 1.66*** | 1.52*** | 1.64*** | 1.68*** | SVM | 1.31** | 1.08 | 1.09 | SVM | 1.15* | 1.12 | 1.18 |
| Forest | 1.29** | 1.13 | 1.01 | 1.33* | 1.4*** | Forest | 1.27* | 1.13 | 1.1 | Forest | 1.2** | 1.14 | 1.21 |
| ANN | 1.38** | 0.96 | 0.91 | 0.72** | 0.96 | ANN | 1.3** | 0.96 | 1.06 | ANN | 1.06 | 0.87 | 1.12 |

**Benchmark: DFM**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 0.96 | 1.03 | 1.18 | 1.39** | 1.25* | AR | 0.92 | 1 | 1.22 | AR | 0.85* | 0.9 | 0.81 |
| PCA | 0.88 | 0.76* | 0.66** | 0.61*** | 0.51*** | PCA | 0.85** | 0.63*** | 0.69*** | PCA | 0.76** | 0.58*** | 0.47*** |
| PLS | 0.74*** | 0.59*** | 0.55*** | 0.57*** | 0.47*** | PLS | 0.82** | 0.56*** | 0.56*** | PLS | 0.71*** | 0.54*** | 0.43*** |
| Ridge | 0.74* | 0.53*** | 0.44*** | 0.41*** | 0.36*** | Ridge | 0.67*** | 0.52*** | 0.49*** | Ridge | 0.64*** | 0.42*** | 0.36*** |
| Lasso | 0.74** | 0.61** | 0.62** | 0.58*** | 0.46*** | Lasso | 0.82* | 0.56*** | 0.58*** | Lasso | 0.66*** | 0.54*** | 0.39*** |
| Elastic | 1.16 | 0.84 | 0.72* | 0.69** | 0.55*** | Elastic | 0.93 | 0.58*** | 0.81 | Elastic | 0.74** | 0.49*** | 0.41*** |
| SVM | 1.88*** | 1.25** | 1.01 | 1 | 0.86 | SVM | 1.12 | 0.68*** | 0.75* | SVM | 0.87 | 0.65*** | 0.55*** |
| Forest | 1.13 | 0.85 | 0.67** | 0.81 | 0.72*** | Forest | 1.08 | 0.7*** | 0.75* | Forest | 0.91 | 0.66*** | 0.56*** |
| ANN | 1.21 | 0.73* | 0.6** | 0.44*** | 0.49*** | ANN | 1.11 | 0.6*** | 0.73** | ANN | 0.81* | 0.5*** | 0.52*** |

Notes: Forecasts using 581 CPI item series as predictors. Sample period 2011-2019, out-of-sample predictions from 2016m5 until 2019m12. Root mean squared errors, relative to AR($p$) model (upper panel), PCA (middle panel), DFM (lower panel). Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $* * *\backslash ** \backslash *$ indicates significance at $10\%, 5\%, and 1\%$, respectively. Relative RMSE that are significant at a level of 10% or lower and taking values below 1 are marked in bold.

Table 4: Forecasting exercise results, CPI items and macroeconomic series predictors.

**Benchmark: AR(p)**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| PCA | 0.84 | **0.67**** | **0.53***** | **0.42***** | **0.4***** | PCA | 0.93 | **0.61**** | **0.5***** | PCA | 0.91 | **0.69**** | **0.58***** |
| DFM | **0.79**** | **0.6***** | **0.52***** | **0.44***** | **0.41***** | DFM | 0.87 | **0.52***** | **0.51***** | DFM | 0.98 | **0.72**** | **0.6***** |
| PLS | **0.72***** | **0.54***** | **0.47***** | **0.41***** | **0.38***** | PLS | 0.88 | **0.53***** | **0.46***** | PLS | 0.87 | **0.61***** | **0.53***** |
| Ridge | **0.75*** | **0.5***** | **0.37***** | **0.29***** | **0.28***** | Ridge | **0.68**** | **0.51***** | **0.39***** | Ridge | **0.76**** | **0.48***** | **0.46***** |
| Lasso | **0.73**** | **0.57***** | **0.51***** | **0.42***** | **0.35***** | Lasso | **0.71**** | **0.58***** | **0.41***** | Lasso | **0.78**** | **0.6***** | **0.47***** |
| Elastic | 1.11 | 0.79 | **0.6***** | **0.49***** | **0.45***** | Elastic | 0.98 | **0.58***** | **0.65***** | Elastic | 0.88 | **0.55***** | **0.51***** |
| SVM | 1.94*** | 1.2** | 0.85 | **0.71***** | **0.68***** | SVM | 1.2* | **0.67***** | **0.6***** | SVM | 1.02 | **0.72***** | **0.68***** |
| Forest | 1.08 | **0.78**** | **0.54***** | **0.58***** | **0.59***** | Forest | 1.06 | **0.6***** | **0.66***** | Forest | 1.08 | **0.74**** | **0.69***** |
| ANN | 0.89 | **0.68**** | 0.81 | **0.56**** | **0.61***** | ANN | 1.21** | **0.72**** | **0.55***** | ANN | 1.04 | **0.66***** | **0.56***** |

**Benchmark: PCA**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.19 | 1.5** | 1.88*** | 2.41*** | 2.5*** | AR | 1.07 | 1.63** | 2*** | AR | 1.1 | 1.44** | 1.73*** |
| DFM | **0.94*** | 0.9 | 0.98 | 1.07 | 1.02 | DFM | 0.93 | 0.84* | 1.01 | DFM | 1.08 | 1.04 | 1.03 |
| PLS | **0.85*** | 0.82* | 0.88 | 0.98 | 0.94 | PLS | 0.94 | 0.87 | 0.91 | PLS | 0.95 | 0.88 | 0.92 |
| Ridge | 0.89 | **0.76**** | **0.69***** | **0.7***** | **0.71***** | Ridge | **0.73***** | 0.84 | **0.77**** | Ridge | 0.83 | **0.69**** | **0.79*** |
| Lasso | 0.87 | 0.86 | 0.96 | 1 | 0.88 | Lasso | **0.76**** | 0.95 | 0.82 | Lasso | 0.86 | 0.86 | **0.81*** |
| Elastic | 1.32** | 1.18* | 1.13 | 1.17 | 1.12 | Elastic | 1.05 | 0.94 | 1.3 | Elastic | 0.97 | **0.79*** | 0.88 |
| SVM | 2.31*** | 1.81*** | 1.59*** | 1.7*** | 1.71*** | SVM | 1.29** | 1.1 | 1.21 | SVM | 1.12 | 1.03 | 1.17 |
| Forest | 1.29** | 1.17 | 1.01 | 1.41* | 1.48** | Forest | 1.14 | 0.99 | 1.31 | Forest | 1.19* | 1.07 | 1.2 |
| ANN | 1.05 | 1.02 | 1.53** | 1.34 | 1.52 | ANN | 1.3** | 1.17 | 1.09 | ANN | 1.14 | 0.95 | 0.97 |

**Benchmark: DFM**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.26** | 1.67*** | 1.91*** | 2.26*** | 2.46*** | AR | 1.15 | 1.94*** | 1.98*** | AR | 1.02 | 1.38** | 1.67*** |
| PCA | 1.07* | 1.11 | 1.02 | 0.94 | 0.98 | PCA | 1.07 | 1.19* | 0.99 | PCA | 0.92 | 0.96 | 0.97 |
| PLS | 0.91 | 0.91 | 0.9 | 0.92 | 0.92 | PLS | 1.01 | 1.04 | 0.91 | PLS | 0.88 | 0.84 | 0.89 |
| Ridge | 0.94 | **0.84*** | **0.7***** | **0.66***** | **0.7***** | Ridge | **0.79**** | 0.99 | **0.76**** | Ridge | **0.77*** | **0.66***** | **0.76**** |
| Lasso | 0.92 | 0.95 | 0.98 | 0.94 | 0.86 | Lasso | 1.13 | 1.12 | 1.29 | Lasso | **0.79*** | 0.83 | **0.78**** |
| Elastic | 1.41*** | 1.31** | 1.15 | 1.1 | 1.1 | Elastic | 1.38*** | 1.31* | 1.2 | Elastic | 0.89 | **0.76*** | 0.85 |
| SVM | 2.46*** | 2.01*** | 1.62*** | 1.59*** | 1.68*** | SVM | 1.38*** | 1.31* | 1.2 | SVM | 1.04 | 0.99 | 1.13 |
| Forest | 1.37*** | 1.3** | 1.03 | 1.32 | 1.46** | Forest | 1.22 | 1.17 | 1.3 | Forest | 1.1 | 1.02 | 1.16 |
| ANN | 1.12 | 1.13 | 1.56** | 1.26 | 1.5 | ANN | 1.4*** | 1.39** | 1.08 | ANN | 1.06 | 0.91 | 0.94 |

Notes: Forecasts using 581 CPI item series and 45 macroeconomic series as predictors. Sample period 2011-2019, out-of-sample predictions from 2015m5 until 2019m12. Root mean squared errors, relative to AR($p$) model (upper panel), PCA (middle panel), DFM (lower panel). Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $* * *\backslash ** \backslash *$ indicates significance at $10\%, 5\%, and 1\%$, respectively. Relative RMSE that are significant at a level of 10% or lower and taking values below 1 are marked in bold.

stronger improvements against the AR benchmark in Table 3 compared to Table B2). And once CPI items are included, the additional improvements from adding macroeconomic series are small for our sample period (i.e. little improvement in Table 4 compared to Table 3). The only exception is the DFM model, which performs rather poorly in the specification with CPI items only; its performance strongly improves once macroeconomic series are added (Table 4).

Second, CPI item information is particularly useful at longer horizons: the relative performance gains of all models against the AR model increase with the horizon. At horizons of 9-12 months, all models strongly improve against the AR benchmark for all targets. This said, a few models even beat the AR model at the 1-month horizon, with significant and substantial improvements in RMSE of up to 25%. This is the case for the PLS, Ridge and Lasso in the specification with CPI items only, and also for the DFM when CPI items and macroeconomic predictors are used.

Third, Ridge regression is the strongest model for the 2011-2019 sample period. This holds throughout horizons and for all three targets. At higher horizons, Ridge reaches strong improvements in RMSE against the AR benchmark of up to 70% for headline inflation, 60% for core inflation, and 55% for service inflation. But also at short horizons, it robustly beats the AR benchmark for all three targets. The Ridge regression is also the only one that yields substantial forecast improvements (up to 30% lower RMSE) against the PCA regression as well as against the DFM which is a strong benchmark in the specification with CPI items and macroeconomic series. At the same time, the Lasso and PLS methods also show high forecast accuracy in most specifications

Forth, machine learning methods, in particular the ANN, perform reasonably well at higher horizons. The models clearly beat the AR benchmark. But in most cases they are outperformed by the linear shrinkage and dimensionality reduction methods. The non-parametric and non-linear nature of machine learning models does not appear to give them a significant advantage compared to the simpler, linear models. This is likely due to a combination of two factors. First, these more complex models have larger data needs, such that their full potential cannot be fully utilised within our rather short sample of CPI item series. Second, non-linear dependencies are likely to play a lesser role over shorter forecasting horizons or shorter sample periods. However, as we show in the next sub-section, over a longer sample period that includes the 2008 financial crisis, for which we have only macroeconomic predictors available, machine learning methods are among the strongest forecasting tools at higher horizons.

The above findings are reflected in the model predictions of headline inflation, shown in Figure 3 for the specification with CPI items only. The figure shows out-of-sample predicted values from the different models (coloured lines) for the period 2015m5-2019m12 over varying horizons (rows of sub-plots). For comparison, black lines indicate the actual outcome of standardised headline inflation, lagged by the number of months corresponding to the horizon. Figure 3 illustrates that most models predict the actual outcome of headline inflation quite closely, even at longer horizons. All models capture the increase in inflation that occurred in 2017 in the aftermath of the Brexit referendum, even at horizons of 9 to 12 months ahead. By contrast, the simple AR benchmark performs well 1-3 months ahead, but fails to capture the rise in inflation at longer horizons. In line with the RMSE results shown above, the forecasts from the PCA and Ridge models show the best fit for the actual inflation outcome. The Lasso, ANN, and Random Forest predictions are somewhat more volatile, whereas Elastic Net and the SVM slightly under-predict the variation in inflation over the out-of-sample period. The DFM forecasts capture the general tendency in inflation movements, but exhibit large spikes. This relatively poor performance of the model likely results from the fact that the model captures joint fluctuations among some of the volatile CPI items, that are not present at the aggregate level.

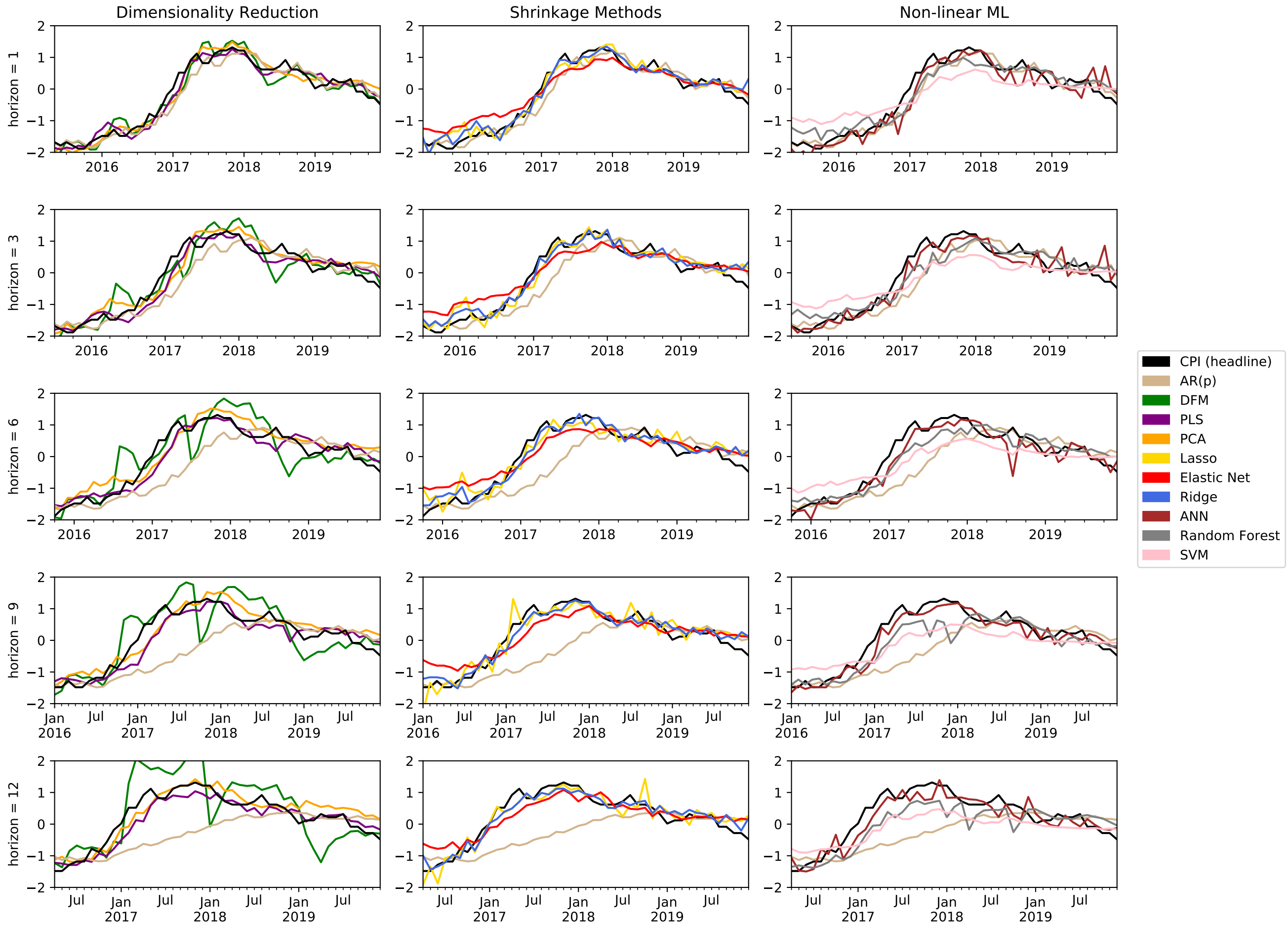

Figure 3: Predicted values for headline CPI inflation, models with CPI item series only. Notes: Forecasts of CPI headline inflation (standardised), from different types of forecasting models (columns, coloured lines) using CPI items and macroeconomic series as predictors, for different horizons $h$ (rows). Out-of-sample predictions for 2015m5 to 2019m12. Compared to the actual headline CPI inflation outcome (black lines), lagged by $h$ months.

Figures B2 and B3 in the appendix present the same results for CPI items combined with macroeconomic predictors and for macroeconomic predictors only, respectively. Indeed, the DFM forecasts are much smoother and closer to the actual inflation outcome once macroeconomic predictors are added to the specification. For most other models, however, predicted values based on specifications including CPI items are somewhat less volatile compared to those using macroeconomic series only. This relates to the fact that macroeconomic series include financial series and production and service indicators which provide information on CPI inflation dynamics, but tend to be more volatile than the latter. By contrast, CPI items show very similar first moment dynamics to aggregate CPI, albeit with a rather high dispersion across items, as depicted by the descriptive statistics in Figure 2. Thus, as long as CPI items are combined with a model that can filter out this general tendency from disaggregated CPI items correctly while disregarding some of the CPI items that show more volatility or discrete jumps, forecasting performance can be improved against more traditional macroeconomic predictors.

## 4.2 Longer sample period with macroeconomic data

Since CPI item series are only available from 2011 onward, we deal with a relatively short sample. Thus, we are limited in our capacity to assess the predictive ability of our models over a longer sample period, and we might underestimate the uncertainty around our forecasts due to the short sample. In the following, we run an alternative forecasting exercise using only macroeconomic series as predictors, but over a longer sample period. The macroeconomic series are available over the period 2000m1 to 2019m12. We define the tuning sample to be of a length comparable to the baseline specification, but assess forecast accuracy over a much longer evaluation period, running recursive out-of-sample forecasts over the period 2004m5 to 2019m12. This includes the Great Recession, during which the fall in inflation turned out to be weaker than the contraction of the real side of the economy would have suggested ("missing disinflation puzzle") (Coibion and Gorodnichenko, 2015; Lindé and Trabandt, 2019), as well as the period of subdued inflationary dynamics during the years 2013-2016 that followed a fall in oil prices but has also been related to a possible flattening of the Phillips curve (Carney, 2017; Forbes, 2019). The longer sample period thus covers more variability in inflation, as well as potential turning points and trend shifts in inflation dynamics. This means that simpler, linear models that perform well over a shorter period might face more difficulty forecasting inflation precisely over this longer period, while more flexible models might have an advantage.

Table 5 shows relative RMSE for this specification. Again, we forecast headline, core, and service inflation over horizons of one to twelve months against three benchmarks. A few findings stand out. First, similarly to the shorter sample, all models clearly outperform the AR benchmark at horizons of 3-12 months for all targets. And half of the models (DFM, PLS, Ridge, Lasso, Forest) also achieve significant improvements against the AR in forecasting headline inflation at the very short one-month horizon.

Table 5: Forecasting exercise results, macroeconomic series predictors, 2000-2019 sample.

**Benchmark: AR(p)**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| PCA | 1.05 | **0.83*** ** | **0.86** ** | **0.9* ** | 0.94 | PCA | 1.44*** | 1.02 | **0.83*** ** | PCA | 1.96*** | **0.88*** ** | **0.74*** ** |
| DFM | **0.9** ** | **0.86*** ** | **0.87** ** | **0.89** ** | 0.93 | DFM | 1.05 | 0.94 | **0.81*** ** | DFM | 1.03 | **0.81*** ** | **0.78*** ** |
| PLS | **0.84* ** | **0.75*** ** | **0.78*** ** | **0.75*** ** | **0.79*** ** | PLS | 1.14** | **0.88** ** | **0.71*** ** | PLS | 1.25*** | **0.74*** ** | **0.65*** ** |
| Ridge | **0.64*** ** | **0.62*** ** | **0.58*** ** | **0.64*** ** | **0.65*** ** | Ridge | 0.98 | **0.72*** ** | **0.54*** ** | Ridge | 1.07 | **0.59*** ** | **0.51*** ** |
| Lasso | **0.69*** ** | **0.79*** ** | **0.78*** ** | **0.88** ** | **0.89*** ** | Lasso | 1.06 | 1.03 | **0.69*** ** | Lasso | 1.01 | **0.71*** ** | **0.58*** ** |
| Elastic | 1.13** | 0.95 | **0.91*** ** | **0.93*** ** | **0.93*** ** | Elastic | 1.29*** | 0.97 | **0.85*** ** | Elastic | 1.57*** | **0.88** ** | **0.74*** ** |
| SVM | 1.27*** | **0.82*** ** | **0.64*** ** | **0.59*** ** | **0.57*** ** | SVM | 1.13** | **0.72*** ** | **0.63*** ** | SVM | 1.61*** | **0.86*** ** | **0.67*** ** |
| Forest | **0.86*** ** | **0.85*** ** | **0.68*** ** | **0.76*** ** | **0.79*** ** | Forest | 1.21*** | 0.98 | **0.87** ** | Forest | 1.21*** | 0.96 | **0.74*** ** |
| ANN | 1 | **0.74*** ** | **0.62*** ** | **0.52*** ** | **0.56*** ** | ANN | 1.16** | **0.68*** ** | **0.62*** ** | ANN | 1.32*** | **0.71*** ** | **0.56*** ** |

**Benchmark: PCA**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 0.95 | 1.2*** | 1.17** | 1.11* | 1.07 | AR | **0.7*** ** | 0.98 | 1.21*** | AR | **0.51*** ** | **1.13*** ** | 1.35*** |
| DFM | **0.86*** ** | 1.03 | 1.02 | 0.99 | 0.99 | DFM | **0.73*** ** | **0.92*** ** | 0.98 | DFM | **0.53*** ** | **0.91** ** | 1.06 |
| PLS | **0.79*** ** | **0.89* ** | **0.92* ** | **0.84*** ** | **0.85*** ** | PLS | **0.79*** ** | **0.86*** ** | **0.86*** ** | PLS | **0.64*** ** | **0.83*** ** | **0.89** ** |
| Ridge | **0.61*** ** | **0.74*** ** | **0.67*** ** | **0.71*** ** | **0.69*** ** | Ridge | **0.68*** ** | **0.71*** ** | **0.65*** ** | Ridge | **0.55*** ** | **0.67*** ** | **0.69*** ** |
| Lasso | **0.66*** ** | 0.94 | 0.91 | 0.98 | 0.94 | Lasso | **0.74*** ** | 1.01 | **0.83*** ** | Lasso | **0.52*** ** | **0.8*** ** | **0.79*** ** |
| Elastic | 1.07 | 1.14** | 1.07 | 1.03 | 0.99 | Elastic | **0.89** ** | 0.96 | 1.02 | Elastic | **0.8*** ** | 1 | 1.01 |
| SVM | 1.21*** | 0.99 | **0.74*** ** | **0.66*** ** | **0.6*** ** | SVM | **0.79*** ** | **0.71*** ** | **0.76*** ** | SVM | **0.82*** ** | 0.98 | 0.91 |
| Forest | **0.82*** ** | 1.01 | **0.8*** ** | **0.85** ** | **0.84** ** | Forest | **0.84*** ** | 0.96 | 1.05 | Forest | **0.62*** ** | 1.09 | 1 |
| ANN | 0.95 | **0.89* ** | **0.72*** ** | **0.58*** ** | **0.59*** ** | ANN | **0.81*** ** | **0.67*** ** | **0.75*** ** | ANN | **0.67*** ** | **0.8*** ** | **0.75*** ** |

**Benchmark: DFM**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.11** | 1.17*** | 1.15** | 1.12** | 1.08 | AR | 0.95 | 1.06 | 1.23*** | AR | 0.97 | 1.24*** | 1.28*** |
| PCA | 1.17*** | 0.98 | 0.98 | 1.01 | 1.01 | PCA | 1.37*** | 1.08*** | 1.02 | PCA | 1.9*** | 1.09** | 0.94 |
| PLS | 0.93 | **0.87** ** | **0.9** ** | **0.84*** ** | **0.86*** ** | PLS | 1.08* | **0.94* ** | **0.87*** ** | PLS | 1.21*** | **0.91* ** | **0.84*** ** |
| Ridge | **0.72*** ** | **0.72*** ** | **0.66*** ** | **0.71*** ** | **0.7*** ** | Ridge | 0.93 | **0.77*** ** | **0.66*** ** | Ridge | 1.04 | **0.73*** ** | **0.65*** ** |
| Lasso | **0.77*** ** | 0.92 | **0.9* ** | 0.99 | 0.96 | Lasso | 1.01 | 1.09* | **0.85*** ** | Lasso | 0.98 | **0.88** ** | **0.75*** ** |
| Elastic | 1.25*** | 1.11 | 1.05 | 1.04 | 1 | Elastic | 1.22*** | 1.04 | 1.04 | Elastic | 1.52*** | 1.09* | 0.95 |
| SVM | 1.42*** | 0.96 | **0.73*** ** | **0.67*** ** | **0.61*** ** | SVM | 1.08 | **0.77*** ** | **0.77*** ** | SVM | 1.56*** | 1.07 | **0.86** ** |
| Forest | 0.95 | 0.99 | **0.79*** ** | **0.85*** ** | **0.85** ** | Forest | 1.15* | 1.04 | 1.07 | Forest | 1.17** | 1.19* | 0.95 |
| ANN | 1.11 | **0.87* ** | **0.71*** ** | **0.59*** ** | **0.6*** ** | ANN | 1.1 | **0.73*** ** | **0.76*** ** | ANN | 1.28*** | **0.88** ** | **0.71*** ** |

Notes: Forecasts using 45 macroeconomic series as predictors. Sample period 2010-2019, out-of-sample predictions from 2004m5. Root mean squared errors, relative to AR(p) model (upper panel), PCA (middle panel), DFM (lower panel). Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $* * *\backslash{*}{*}\backslash{*}$ indicates significance at $10\%, 5\%, and 1\%$, respectively. Relative RMSE that are significant at a level of 10% or lower and taking values below 1 are marked in bold.

Second, Ridge regression is among the strongest models also for the longer sample, thus confirming our results above. It achieves forecasting gains of more than 35% against the AR benchmark for headline inflation and, remarkably, this improvement is achieved throughout all forecasting horizons considered. For core and service inflation, it reaches improvements of 30% to 40% at the 6-month horizon and up to 50% at the 12-month horizon. This indicates the robustness of our results from the shorter sample period over the longer sample. On the other hand, the other two shrinkage methods, Lasso and Elastic Net, perform worse for headline inflation forecasts at higher horizon compared to their performance with CPI items. Although the Lasso does show high forecast accuracy at short horizons for headline inflation as well as for service inflation.

Third, machine learning tools, in particular the SVM and the ANN show a high forecast accuracy, particularly at horizons of 6-12 months. For predictions of headline inflation 9-12 months ahead, these two models show the strongest performance with improvements against the AR benchmark of more than 40%, even outperforming Ridge regression. Hence, these models prove particularly useful for inflation forecasts at higher horizons over a sample period which includes periods of higher inflation volatility or potential trend shifts. They are particularly effective for forecasting the more volatile headline inflation component, while they lose part of their advantage when it comes to forecasting the more stable core and service inflation rates.

Figure 4 shows results for the predicted values of headline inflation from the different models, along with the actual outcome. For horizons of 6-12 months, clear differences across models are visible. Looking at the 2008-2009 financial crisis episode, all models somewhat underestimate the increase in inflation in 2008, but then differ in their ability to capture the 2009 decline. The dimensionality reduction techniques (DFM, PCA, PLS) lag behind and strongly overestimate the fall in inflation in 2009 in line with the missing disinflation puzzle. On the other hand, Ridge regression, SVM and ANN capture the 2009 decline in inflation remarkably well both in terms of size and timeliness, even at higher horizons. Next, looking at the protracted weakness in inflation during the years 2013-2016, the 12-month ahead forecasts from dimensionality reduction techniques completely miss this until 2015, when predicted values start to go down only slowly. By contrast, the 12-month ahead forecasts from Ridge, SVM and ANN capture the protracted decline in inflation very closely, albeit with some excess volatility at higher frequencies. For the years 2016-2017, when inflation remained low initially and then rose in the second half of 2016 following the Brexit vote, the fit of Ridge regression's 12-month ahead forecasts deteriorates somewhat, first underestimating the protracted low level of inflation and then over-estimating the increase. By contrast, the SVM and ANN forecasts track inflation outcomes very well during this time at all horizons. These two models, and in particular the ANN, also capture the return of headline inflation to lower levels during 2018-2019 exceptionally well, apart from few higher-frequency spikes. These two machine learning models appear particularly apt in capturing turning points early on even at higher horizons. The Random Forest, on the other hand, performs well for lower horizons, but shows discrete jumps and misses turning points at higher horizons. Finally, the Lasso and Elastic Net seem to suffer from excess sparsity at higher horizons, not capturing inflation dynamics at all during some periods—these models seem to pick up signals from the policy rate which was restricted by the effective lower bound.

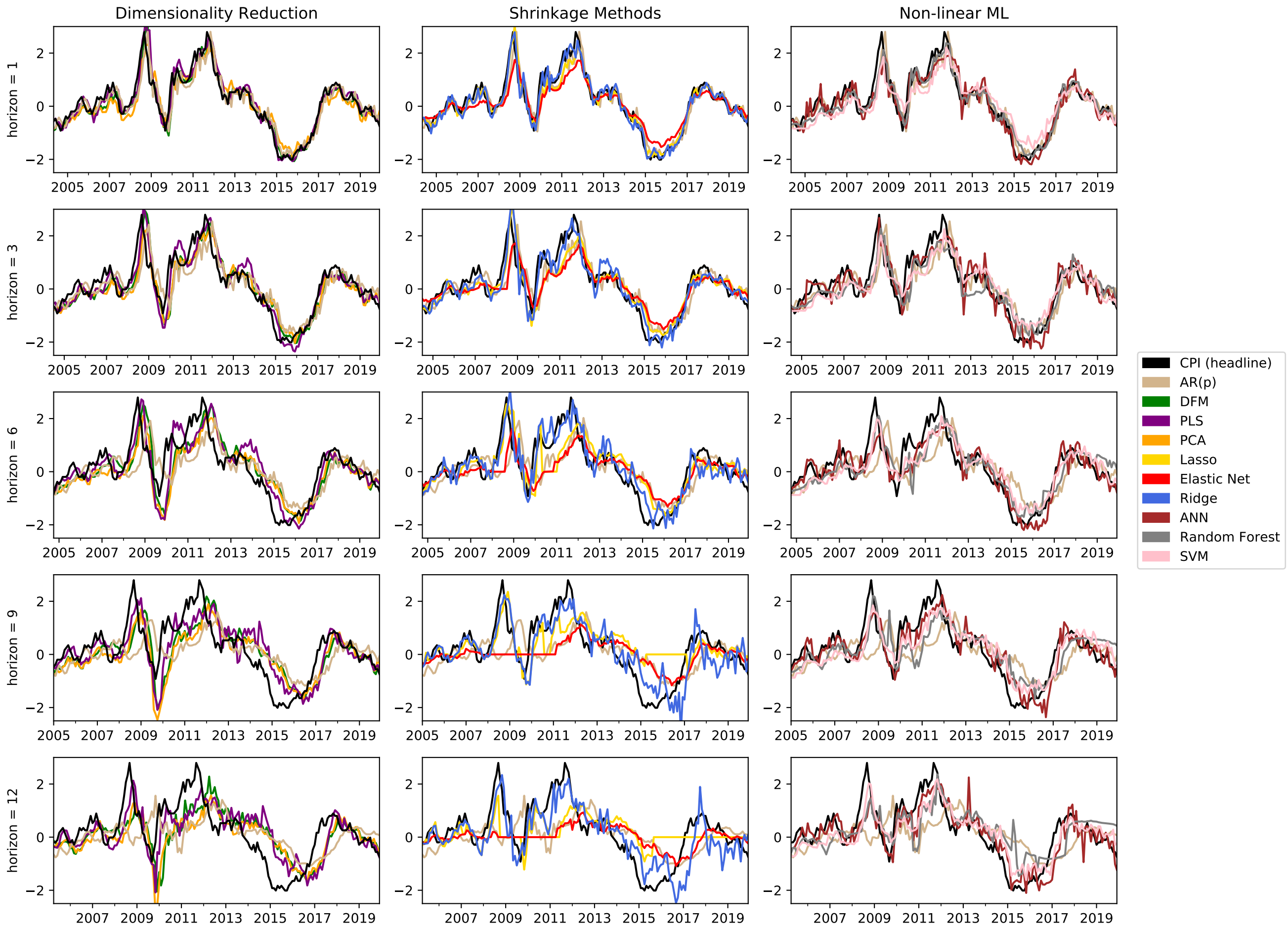

Figure 4: Predicted values for headline CPI inflation, models with macroeconomic series, 2000-2019 sample.

Notes: Forecasts of CPI headline inflation (standardised), from different types of forecasting models (columns, coloured lines) using CPI items and macroeconomic series as predictors, for different horizons $h$ (rows). Out-of-sample predictions from 2004m5 to 2019m12. Compared to the actual headline CPI inflation outcome (black lines), lagged by $h$ months.

## 4.3 Discussion of forecasting "horse race" results

Our main findings from the two previous sub-sections can be summarised as follows. First, using large datasets for forecasting aggregate inflation in the UK strongly improves forecast accuracy at horizons of 3-12 months. The use of disaggregated CPI item series yields stronger gains than the use of macroeconomic predictors for the sample period 2011-2019 for all models except the DFM. Second, Ridge regression performs very well across horizons, targets and specifications. It is particularly strong when combined with CPI items and over the 2011-2019 sample period. Third, SVM and the ANN show the best forecast accuracy when forecasting headline inflation over the longer sample period 2000-2019, for which only macroeconomic predictors are available. The two models capture most turning points and episodes of changes in volatility in aggregate inflation dynamics particularly well throughout the extended sample. Why do some models perform better when combined with CPI items rather than macro data or over different sample? We can shed additional light on these results by comparing forecasts across specifications and by relating the findings to the existing literature.

Figure 5 compares absolute RMSE across specifications with different sets of predictors and the two sample periods (coloured bars) for each model. The figure illustrates that, while the SVM and ANN perform worse than Ridge regression and PCA for the shorter sample period and with CPI item predictors, the forecasting accuracy of the two machine learning models is remarkably stable across specifications. The two models do not show any deterioration in RMSE for the longer sample period (light blue bars) compared to the other specifications, in contrast to all other models. The ANN performs particularly well: it almost reaches the forecast accuracy of Ridge regression and PCA for the short sample and when combined with CPI items, while being the strongest model for the longer sample. This result points to the potential of machine learning models to perform better in the future also when combined with CPI item series, as more data become available yielding an increased $T/N$ ratio.

Our findings contribute to the debate in the forecasting literature on the use of sparse versus dense models. Whereas we do not discuss formal definitions of sparsity here, which are not straight-forward in the non-linear setting, we can derive considerations from our results in terms of comparing results from different models. We understand a model to be rather sparse if it selects a small set of explanatory variables with the highest predictive power and rather dense if it uses the full range of available economic predictors while allowing for the impact of some of them to be small through shrinkage or regularisation. Giannone et al. (2017) assess the predictive accuracy of sparse and dense modelling techniques combined with various macroeconomic and financial data sets for the United States. They conclude that, despite the potential advantage of sparse models of being easier to interpret, they are rarely preferred to dense models in economics. Our results confirm their findings for the United Kingdom with respect to macroeconomic data: in the specifications where only macroeconomic predictors are included, sparse models such as the Lasso, Elastic Net and Random Forest perform worse than their dense counterparts at higher horizons, picking up dynamics from individual predictors such as the policy rate restricted by the Effective Lower Bound.

However, when using disaggregated CPI items as predictors, sparse methods do appear among the best performing models. Although the strongly performing Ridge regression

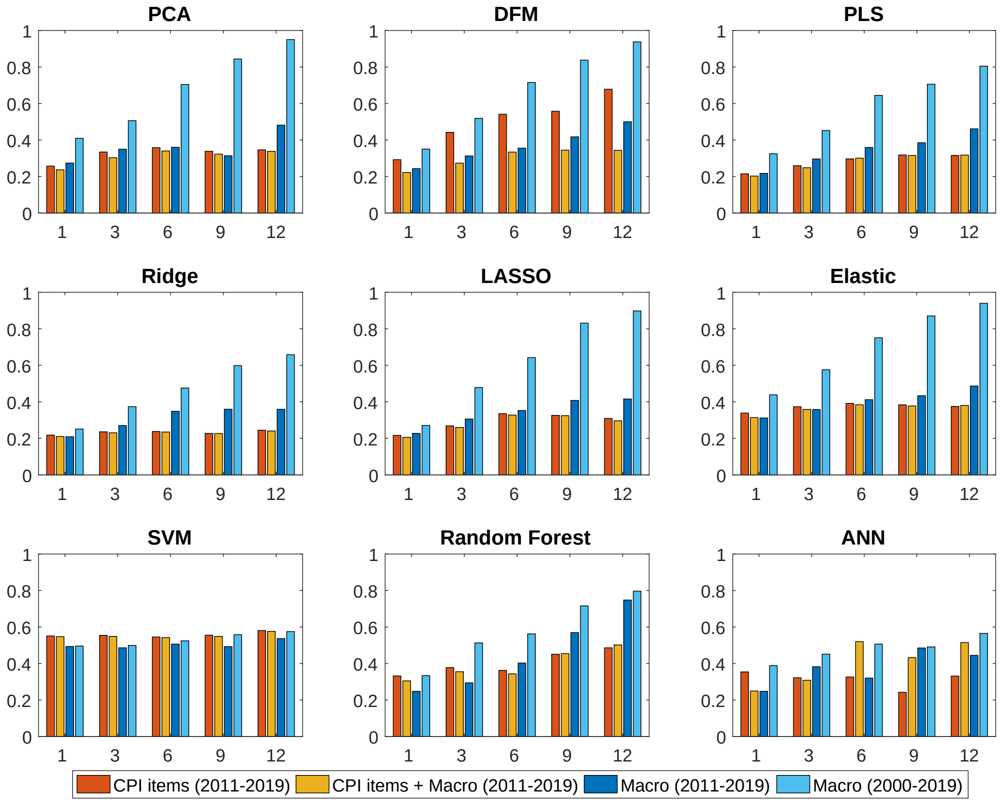

Figure 5: RMSE for headline CPI inflation, varying horizons and sets of predictors.
Notes: Average RMSE from forecasts of CPI headline inflation, from different forecasting models (subplots) over varying horizons (x-axis). Bars indicate specifications with different sets of predictors from left to right: CPI items (orange), CPI items and macroeconomic series (yellow), macroeconomic series only (dark blue), macroeconomic series over extended sample period 2000-2019 (light blue).

and PCA can both be considered dense models, the sparse Lasso performs almost as well as Ridge, and also Elastic Net and Random Forest are among the models with relatively high forecast accuracy.[16] Hence, what seems to work well with disaggregated CPI items is efficient shrinkage and penalisation of the large information set, while it is of secondary importance whether the shrinkage method is a sparse or a dense one. On the other hand, exploiting dynamic co-movement in the data appears to work less well with CPI item data compared to the macroeconomic context. As such, the DFM performs poorly when fed with item series only. This relates to the fact that many CPI items show discrete jumps and non-standard dynamics and some of those co-move, without their common component being reflected in aggregate inflation dynamics. With such data at hand, it helps to downgrade this information directly through shrinkage instead of modelling co-movement dynamically. This said, our results based on CPI items consider a sample

[16] In section Shapley 5, we provide evidence based on Shapley values and regressions indicating that the Random Forest uses information in a less dense manner compared to Ridge regression, in terms of the distribution of predictive shares across CPI item predictors. We therefore discuss it as one of the rather sparse models here.

period during which aggregate CPI inflation was relatively stable —the difference between sparse and dense models might be more pronounced over longer sample periods.

Finally, our findings also contribute to the understanding of the forecasting performance of linear models compared to non-linear, non-parametric machine learning tools. In our analysis, machine learning models have the highest forecast accuracy for headline inflation over the longer sample period where aggregate inflation exhibits different volatility regimes. Machine learning models prove to be particularly strong in capturing turning points and changes in dynamics such as the "missing disinflation" after 2009 as early as one year ahead. These models therefore also appear promising for capturing extraordinary events such as the Covid-19 shock, which we leave for future research. At the same time, according to our results, Ridge regression yields robust and accurate forecasts of inflation using both CPI items and macroeconomic data. Over the longer sample period, it performs almost as well as the machine learning tools, while it beats them over the shorter period. Linearity makes Ridge regression estimation easy, and interpretation is facilitated through the estimated shrinkage weights that are assigned to individual predictors. This makes the model a relevant candidate for central banks and institutions that aim to forecast inflation with large datasets.

# 5 Opening the forecasting "black boxes"

We have shown in the previous section that the use of disaggregate CPI item data combined with a wide range of models improves aggregate inflation forecasts. However, this large set of predictors combined with potentially complex models comes with the drawback of challenges in interpreting forecast outcomes, i.e. our results are subject to the "black box critique". Dimensional reduction and shrinkage methods do provide tractable measures of contributions from individual predictors (e.g. factor loadings or regression parameters after shrinkage). Yet, finding a meaningful way to re-aggregate signals from individual items to wider classes or sectors to help interpretation remains a challenge. The non-parametric and non-linear machine learning tools additionally come with the difficulty to pin down which variables drive model predictions.

For the interpretation of results, three questions are of interest. First, what is the contribution of a predictor to the forecast? Second, are certain groups of predictors, e.g. CPI items that belong to a COICOP class, more relevant than others for forecasts of a specific target? And third, conditional on how much a component explains in a model, is there a clear association between the component and actual inflation, i.e. does it statistically co-move with the target value? Answers to these questions can be informative for identifying relevant predictors for lower-dimensional forecasting frameworks, or for communicating forecasts and to inform policy decisions. We address them through a model-agnostic approach that proceeds in three steps: model decomposition, partial re-aggregation and statistical testing. We describe the approach in more detail in the following before presenting results.

## 5.1 Shapley values and regressions

The first step is **model decomposition**. We employ the *Shapley additive explanations* framework (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017) which exploits an analogy between variables in a model and players in a cooperative game. The Shapley value framework has a set of appealing analytical properties while being applicable to any model.[17] It consists in calculating the 'payoff' for including a specific predictor in the model, conditional on other predictors being present. Each prediction (i.e. a predictive value at time $t$ and horizon $h$) from a model is decomposed into the sum of contributions from individual predictors, the so-called *Shapley values*.

Let the total number of predictors be $M = N + p$, with $N$ price item and macro series, and $p$ lags of the target variable, as described in section 3. The predicted value, $\hat{y}_{t+h}$, of a model at time $t$ for the forecast horizon $h$ can be decomposed into its Shapley components $\phi^h_{tj}$ for the $j^{th}$ variable. That is,

$$\hat{y}_{t+h} = \sum_{j=0}^{M} \phi^h_{tj} \equiv \Phi^h_t, \qquad \text{(decomposition)} \tag{4}$$

The $j = 0$ component is set to the mean predicted value in the training set and can be interpreted as an intercept. For a non-linear forecasting model, computation of (4) requires deriving the marginal contribution of predictor $j$ by running sequential forecasts of all possible coalitions of predictors, with and without $j$. Thus, the Shapley value for predictor $j$ is computed as

$$\phi_j = \sum_{S \subseteq M \setminus j} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \left[ f(S \cup \{j\}) - f(S) \right]. \tag{5}$$

Here, the payoff of a coalition $S$ is $f(S)$, the payoff of this coalition combined with predictor $j$ is $f(S \cup \{j\})$, and their difference measures the marginal contribution of $j$ to that coalition. The intercept $\phi_0$ corresponds to $f(\varnothing)$, i.e. with no variables in the model (see Appendix A for more details). After summing these marginal contributions over all coalitions, we get an estimate of the contribution of variable $j$ to a single model prediction. Comparing all possible combinations of predictors with $M \approx 600$ quickly becomes computationally infeasible. We therefore focus our analysis on models where an exact solution exists, namely linear models and the random forest. For a linear regression model, the Shapley value of predictor $j$ is simply the product of its regression coefficient

[17] In particular, it is the only attribution framework that is local, linear, efficient, symmetric and respects null contributions and strong monotonicity of variables. A more detailed description of Shapley values in the context of model decomposition as well as Shapley regression is given in Appendix A.

$w_j$ and the difference between the predictor value and its mean, i.e. $\phi_{tj} = w_j(z_{tj} - \mathbb{E}_t[z_{tj}])$ with the expectation taken over the training dataset. For the random forest, or tree-based models more generally, variable coalitions correspond to paths down the branches of the model where these variables lie on the same branch. These can generally be enumerated easily, reducing the complexity of the sum in Eq. 5 (see Lundberg et al. (2018) for details). For other models, coalitions can be sampled with a readjustment of the weights in (5).

The second step is **context-specific re-aggregation**. Predictors in our case are mostly micro price indices. Their values have a clear interpretation as the relative price of a narrowly defined product. However, single item series can be volatile and difficult to keep track of, as is the case with the corresponding Shapley values. We therefore aggregate the $N$ Shapley components $\Phi_t^h$ of a model into $K << N$ higher-level meso-components denoted $\Psi_t^h$,

$$\hat{y}_{t+h} = \sum_{k=0}^{K} \psi_{tk}^h \equiv \Psi_t^h \tag{6}$$

$$\text{with} \quad \psi_{tk}^h = \sum_{j \in \mathcal{C}} \phi_{tj}^h \qquad \text{(meso-aggregation)}\,, \tag{7}$$

with $\psi_{t0}^h = \phi_{t0}^h$ being the same intercept. The grouping of meso components is given by $\mathcal{C}$. For headline inflation, we allow for $K = 13$ components representing the contributions to predictions from the twelve largest COICOP2 classes (called divisions), respectively, and one component summarising contributions from the $p$ lags of the dependent variable.[18] Note that Eq. 6 & 7 can take more general forms for what follows, such weighted averages.

The final step consists in **statistical inference**. We regress the test target values, i.e. the values we aim to forecast, on our $K$ Shapley meso-components using the following *Shapley regression* (see Joseph, 2019).

$$y_{t+h} = \hat{\alpha}' + \sum_{k=1}^{K} \hat{\beta}_k \psi_{tk}^h + \epsilon_i \qquad \text{(statistical inference)} \tag{8}$$

Equation 8 tests if there is a non-zero alignment between the Shapley components $\psi_{tk}^h$ and $y_{t+h}$. When running the Shapley regression, we pool predictions over forecasting periods and horizons in order to increase statistical power in view of the relatively short test period. However, to check if the the importance of different components changes across horizons, we distinguish between short horizons (1-6 months) and longer horizons (7-12 months). We control for possibly correlated errors across horizons.

[18] The included COICOP2 classes are Food & non-alcoholic beverages, Alcoholic beverages & tobacco, Clothing & footwear, Housing & fuels, Furnishing & house maintenance, Health, Transport, Communication, Recreation & culture, Education, Restaurants & hotels, Miscellaneous goods & services (see also (ONS, 2019)).

## 5.2 Results based on Shapley values

Our analysis focuses on Ridge regression and the Random Forest, and mainly on the specification with CPI items and inflation lags as predictors. Apart from the relative ease of computing Shapley values for these models, their comparison is also of more general interest. The former is a dense linear model, while the latter is an often sparse non-linear model. As discussed previously, the forecasting performance of Ridge regression is consistently better than that of the Random Forest. The Shapley analysis sheds some light on why this is the case.

The results derived from forecasts using CPI item predictors are summarised in Table 6. Results for the decomposition of forecasts from Ridge regression and Random Forest into the contributions from CPI divisions are shown on the left and right hand side, and short and longer forecasting horizon in the upper and lower part of the table, respectively. The number below each coefficient is the *predictive share* from the Shapley decomposition of each prediction allocated to that division across horizons. It is an indicator for the overall importance of an aggregated component for that model. Note that almost all coefficients in Table 6 are positive. This is so by construction as Shapley values absorb the sign of a component, and negative coefficients, especially when highly significant, would point to a poor model fit, e.g. through insufficient convergence.[19] However, the underlying assumption behind this is that training and test data are drawn from the same distribution. This need not be the case in out-of-sample forecasting situations like ours, where there likely is some unknown drift in the data generating process. This means that negative and significant coefficients *can* point to persistent model surprises.

We do not expect models and specifications to be strictly comparable in all cases, due to the different nature of the two models, their varying forecast performance, and the small time dimension relative to the input space. All these factors increase the likelihood of models picking up different signals. Nonetheless, we observe some consistency across horizons, targets and models, while discrepancies are in line with expected differences in the model structure and forecasting performance.

Results in Table 6 indicate that the contribution of CPI classes to the forecast in terms of predictive shares and the significance of Shapley coefficients remains rather similar between low and high forecast horizons, for both models and all three targets. Only the contribution of lagged coefficients decreases with higher horizons, as we would expect. The stable role of CPI items across horizons is in line with the strong forecasting gains they provide relative to the AR benchmark at higher horizons, as documented in section 4.1. Interestingly, Food & non-alcoholic beverages (short: Food) receives a high contribution to the Ridge forecast of headline inflation both at lower and higher horizons, and the

[19] This can be easily seen for Ridge regression, where the Shapley values of a variable are the product of the corresponding coefficients and the input values, such that the sign of the coefficient is already accounted for.

Table 6: Comparison of Shapley-value-based model inference using only item indices.

| | Ridge Regression | | | Random Forest | | |
|---|---|---|---|---|---|---|
| | *headline* | *core* | *service core* | *headline* | *core* | *service core* |
| division | | | 1 – 6 months horizon | | | |
| LAG | **0.09**** | **0.08**** | **0.13**** | **0.04*** | **0.16***** | 0.04 |
| | 0.01 | 0.02 | 0.02 | 0.01 | 0.05 | 0.03 |
| Food & non-alc. bev. | **0.11**** | – | – | **0.38***** | – | – |
| | 0.23 | – | – | 0.38 | – | – |
| Acl. bev. & tobacco | 0.02 | **0.05**** | 0.03 | 0.01 | 0.05 | 0.08 |
| | 0.04 | 0.06 | 0.07 | 0.01 | 0.02 | 0.01 |
| Clothing & footwear | **0.10***** | **0.12***** | **0.20***** | **0.04**** | **0.05*** | 0.01 |
| | 0.08 | 0.15 | 0.14 | 0.01 | 0.04 | 0.07 |
| Housing & fuels | **0.10***** | **0.10***** | 0.06 | **0.06***** | -0.00 | **0.11**** |
| | 0.08 | 0.05 | 0.06 | 0.05 | 0.08 | 0.06 |
| Furnishing & house maint. | **0.10***** | **0.10***** | **0.12***** | **0.04**** | -0.01 | **0.10**** |
| | 0.11 | 0.13 | 0.13 | 0.02 | 0.06 | 0.04 |
| Health | **0.04**** | **0.10***** | 0.06 | 0.01 | 0.03 | **0.15***** |
| | 0.04 | 0.05 | 0.05 | 0.03 | 0.03 | 0.02 |
| Transport | **0.21***** | **0.21***** | **0.15***** | **0.22***** | **0.24***** | **0.22***** |
| | 0.08 | 0.10 | 0.10 | 0.12 | 0.13 | 0.12 |
| Communication | -0.02 | 0.01 | 0.03 | **0.13***** | 0.05 | **0.09**** |
| | 0.03 | 0.04 | 0.04 | 0.03 | 0.05 | 0.02 |
| Recreation & culture | **0.18***** | **0.29***** | **0.30***** | **0.12***** | **0.25***** | **0.21***** |
| | 0.11 | 0.16 | 0.17 | 0.07 | 0.14 | 0.19 |
| Education | 0.0 | 0.02 | **0.11**** | **0.03*** | -0.03 | 0.05 |
| | 0.01 | 0.01 | 0.01 | 0.02 | 0.00 | 0.15 |
| Restaurants & hotels | **0.15***** | **0.13***** | -0.01 | **0.17***** | **0.15***** | 0.05 |
| | 0.09 | 0.06 | 0.04 | 0.14 | 0.12 | 0.06 |
| Misc. goods & services | **0.08**** | **0.15***** | **0.17***** | **0.15***** | **0.39***** | **0.14***** |
| | 0.10 | 0.16 | 0.17 | 0.11 | 0.28 | 0.23 |
| | | | 7 – 12 months horizon | | | |
| LAG | -0.01 | **-0.06*** | 0.01 | **0.2***** | **0.16***** | 0.08 |
| | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 |
| Food & non-alc. bev. | **0.23***** | – | – | **0.08**** | – | – |
| | 0.21 | – | – | 0.18 | – | – |
| Acl. bev. & tobacco | **0.06***** | **0.09***** | **0.17***** | -0.03 | 0.0 | 0.10 |
| | 0.04 | 0.06 | 0.07 | 0.01 | 0.01 | 0.05 |
| Clothing & footwear | **0.07***** | **0.11***** | **0.17***** | **0.1***** | **0.06*** | **0.14**** |
| | 0.08 | 0.14 | 0.16 | 0.02 | 0.06 | 0.04 |
| Housing & fuels | 0.03 | 0.06 | -0.01 | **0.30***** | **0.13***** | 0.05 |
| | 0.07 | 0.07 | 0.06 | 0.07 | 0.05 | 0.09 |
| Furnishing & house maint. | **0.09***** | **0.08***** | **0.12**** | **0.12***** | **0.12***** | 0.00 |
| | 0.07 | 0.10 | 0.10 | 0.04 | 0.05 | 0.10 |
| Health | -0.02 | **0.11***** | 0.02 | -0.02 | 0.02 | **0.1*** |
| | 0.05 | 0.07 | 0.05 | 0.02 | 0.01 | 0.05 |
| Transport | **0.14***** | **0.16***** | **0.13**** | **0.14***** | **0.21***** | **0.18**** |
| | 0.07 | 0.10 | 0.09 | 0.07 | 0.10 | 0.06 |
| Communication | **-0.06**** | -0.04 | **0.08*** | **0.06**** | -0.01 | 0.04 |
| | 0.03 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| Recreation & culture | **0.13***** | **0.29***** | **0.29***** | **0.12***** | **0.23***** | **0.18***** |
| | 0.10 | 0.16 | 0.18 | 0.06 | 0.15 | 0.31 |
| Education | 0.03 | -0.01 | -0.03 | 0.03 | -0.0 | 0.04 |
| | 0.01 | 0.01 | 0.01 | 0.07 | 0.05 | 0.02 |
| Restaurants & hotels | **0.3***** | **0.26***** | **0.08*** | **0.18***** | **0.25***** | **0.2**** |
| | 0.11 | 0.06 | 0.05 | 0.22 | 0.21 | 0.11 |
| Misc. goods & services | **0.15***** | **0.21***** | **0.13**** | **0.22***** | **0.26***** | **0.14**** |
| | 0.14 | 0.17 | 0.16 | 0.22 | 0.30 | 0.14 |

Notes: Summary of Shapley regression results (8) for Ridge regression (LHS) and Random Forest (RHS) for headline, core and service core inflation. CPI models components for different horizons are grouped. The share of each aggregate component (6) is given below each coefficient. Core and service core targets do not contain item components from food and non-alcoholic beverages. Significance levels: ***:1%, **:5%, *:10%. Panel-HAC standard errors grouped by forecast horizon have been used. Source: ONS and authors calculations.

size of the corresponding Shapley coefficient even increases at higher horizons. This is surprising, since Food is typically considered to be relevant at short horizons only. The strong role of this class for Ridge forecasts might in part explain the strong performance of this model. On the other hand, for the Random Forest, which shows a lower forecast accuracy at higher horizons, the contribution of Food declines considerably for higher horizons. Other CPI classes that show high predictive shares and significant Shapley coefficients across models and horizons are Furnishing & House maintenance, Transport, Recreation & culture, and Miscellaneous goods & service. The latter two classes, which include services and less volatile goods items, show higher contributions to the predictions of core and service core inflation than headline inflation.

Generally, the Shapley decomposition results reflect that Ridge regression extracts information from a wider range of CPI item classes than the Random Forest. This is reflected in the more evenly distributed component shares and the larger number of significant components in the case of Ridge regression. The shrinkage for Ridge regression rarely reduces coefficients to zero, whereas the Random Forest often ignores variables altogether in a high-dimensional setting.[20] If there are shared trends for items series in the same division relevant for the optimisation problem, the forest is more likely to lean towards this group of predictors while ignoring others, and it is, in this sense, more similar to the Lasso.

We repeat this analysis including the 43 macroeconomic predictors, which we group into real activity, house prices and financial market indicators.[21] Analogous decomposition and inference results are shown in Table B3 in the Appendix. The results are in line with Table 6. The rather sparse nature of the Random Forest compared to Ridge regression becomes clear in this specification as well, particularly at high horizons. The Random Forest, putting more weight on a smaller number of variables, learns less from macroeconomic information at longer forecasting horizons, while it extracts much signal from them on shorter horizons.

Overall, the Shapley decomposition results suggest that the better performance of Ridge regression can be explained, at least in part, by its better handling of the large dimension of the problem, whereas the importance of non-linearities that the Random Forest can account for is secondary in this setting. Given the nature of the problem we address, it is not possible to assess whether any of the presented models reflects the ground truth. Rather, the presented approach aims at addressing the black box nature of both the high dimensionality of the problem and the models built on top of it. Given that many models delivered sizeable forecasting gains, the above approach presents a standardised framework within which to discuss results and differences between models.

[20] This means they are never selected at split points during optimisation within individual trees of the forest.

[21] Other price indices have been removed as those are captured by the lag of the target.

# 6 Conclusion

We have conducted a forecasting exercise with the goal to predict UK inflation using a unique and granular set of monthly CPI item series. We have considered out-of-sample forecasting using a wide range of models that deal with the high dimensionality of the data set in different ways: dimension reduction techniques including a dynamic factor model, shrinkage methods, and non-linear machine learning tools. We also compared our results to the case of including or only relying on more standard macroeconomic time series to evaluate when more granular price data sources are valuable and in combination with which models these are best suited for our task.

We have shown that the disaggregate CPI item series have predictive power for multiple aggregate CPI measures, and independently of the model used. At the same time, we have documented the dominance of Ridge regression when exploiting CPI item predictors. This model achieves substantial and robust gains in forecast accuracy against an AR benchmark and compared to the other models considered. But also the Lasso, Partial Least Squares and Principal Component Analysis perform well when combined with CPI item predictors. On the other hand, the Dynamic Factor Model only performs strongly when macroeconomic series are added to its information set. Non-linear machine learning models beat the AR benchmark in all specifications, but when combined with CPI items, they do not reach up to the best shrinkage methods. They do, however, perform well for a longer sample period when combined with macroeconomic series, providing precise 1-year ahead forecasts around critical turning points, such as the global financial crisis in 2008-2009. This suggest that machine learning models may have potential going forward as more data become available, e.g. for analysing the impact of the Covid-19 shock and how this is reflected in our micro price data.

Generally, our analysis shows the strong potential of using disaggregate item-level price data to forecast aggregate inflation. Item-level series connect individual prices at the product level with the macroeconomic CPI inflation concepts policy makers and economists are ultimately interested in. Apart from the observed improvements in forecast accuracy, forecasts derived from item series can help researchers and policy makers to interpret and communicate adjustments to forecasts based on dynamics observed across sub-groups of items and economic sectors. But the large dimension of the input space, the volatility of individual CPI items, and the opacity of some of the models that can deal with such large data also pose challenges for the interpretation of forecasting results. We have addressed this challenge through the Shapley value and regression framework. The method derives linear contributions to the forecast from groups of items along CPI divisions. It thus represents a universal way to understand and communicate forecast results within economic policy settings.

# References

Almosova, A. and N. Andresen (2019). Nonlinear inflation forecasting with recurrent neural networks. Unpublished manuscript.

Aparicio, D. and M. I. Bertolotto (2020). Forecasting inflation with online prices. *International Journal of Forecasting 36*(2), 232–247.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bańbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013). Now-casting and the real-time data flow. In *Handbook of economic forecasting*, Volume 2, pp. 195–237. Elsevier.

Bluwstein, K., M. Buckmann, A. Joseph, M. Kang, S. Kapadia, and Ö. Simsek (2020). Credit growth, the yield curve and financial crisis prediction: evidence from a machine learning approach. Bank of England Staff Working Paper No 848.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Carney, M. (2017). [De]Globalisation and inflation. Speech at the 2017 IMF Michel Camdessus Central Banking Lecture.

Carriero, A., A. B. Galvao, and G. Kapetanios (2019). A comprehensive evaluation of macroeconomic forecasting methods. *International Journal of Forecasting 35*(4), 1226–1239.

Chen, X., J. Racine, and N. R. Swanson (2001). Semiparametric ARX neural-network models with an application to forecasting inflation. *IEEE Transactions on neural networks 12*(4), 674–683.

Chu, B., K. Huynh, D. Jacho-Chávez, O. Kryvtsov, et al. (2018). On the evolution of the united kingdom price distributions. *The Annals of Applied Statistics 12*(4), 2618–2646.

Coibion, O. and Y. Gorodnichenko (2015). Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation. *American Economic Journal: Macroeconomics 7*(1), 197–232.

Coulombe, P. G., M. Leroux, D. Stevanovic, S. Surprenant, et al. (2019). How is machine learning useful for macroeconomic forecasting? Unpublished manuscript.

Diebold, F. M. and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business & economic statistics 20*(1).

Domit, S., F. Monti, and A. Sokol (2019). Forecasting the UK economy with a medium-scale Bayesian VAR. *International Journal of Forecasting 35*(4), 1669–1678.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Faust, J. and J. H. Wright (2013). Forecasting inflation. In *Handbook of economic forecasting*, Volume 2, pp. 2–56. Elsevier.

Fisher, A., C. Rudin, and F. Dominici (2018). All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. *arXiv preprint 1801.01489*.

Forbes, K. (2019). Inflation Dynamics: Dead, Dormant, or Determined Abroad? NBER Working Paper No. 26496.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The elements of statistical learning*, Volume 1. Springer series in statistics New York.

Garcia, M. G., M. C. Medeiros, and G. F. Vasconcelos (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting 33*(3), 679–693.

Giannone, D., M. Lenza, and G. E. Primiceri (2017). Economic predictions with big data: The illusion of sparsity. CEPR Discussion Paper No. DP12256.

Giannone, D., L. Reichlin, and D. Small (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics 55*(4), 665–676.

Groen, J. J. and G. Kapetanios (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis 100*, 221–239.

Harvey, D. and P. Newbold (2000). Tests for multiple forecast encompassing. *Journal of Applied Econometrics 15*(5), 471–482.

Hendry, D. F. and K. Hubrich (2011). Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate. *Journal of business & economic statistics 29*(2), 216–227.

Hernández-Murillo, R. and M. T. Owyang (2006). The information content of regional employment data for forecasting aggregate conditions. *Economics Letters 90*(3), 335–339.

Hubrich, K. (2005). Forecasting euro area inflation: Does aggregating forecasts by hicp component improve forecast accuracy? *International Journal of Forecasting 21*(1), 119–136.

Ibarra, R. (2012). Do disaggregated cpi data improve the accuracy of inflation forecasts? *Economic Modelling 29*(4), 1305–1313.

Ishwaran, H. and M. Lu (2019). Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. *Statistics in Medicine 38*(4), 558–582.

Joseph, A. (2019). Shapley regressions: A framework for statistical inference on machine learning models. *arXiv preprint 1903.04209*.

Kapetanios, G. (2004). A note on modelling core inflation for the uk using a new dynamic factor estimation method and a large disaggregated price index dataset. *Economics Letters 85*(1), 63–69.

Kapetanios, G. and M. Marcellino (2009). A parametric estimation method for dynamic factor models of large dimensions. *Journal of Time Series Analysis 30*(2), 208–238.

Kim, H. H. and N. Swanson (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting 34*(2), 339–354.

Klenow, P. and O. Kryvtsov (2008). State-Dependent or Time-Dependent Pricing: Does it Matter for Recent U.S. Inflation? *The Quarterly Journal of Economics 123*(3), 863–904.

Koop, G. M. (2013). Forecasting with medium and large bayesian vars. *Journal of Applied Econometrics 28*(2), 177–203.

Lindé, J. and M. Trabandt (2019). Resolving the missing deflation puzzle.

Lundberg, S., G. Erion, and S. Lee (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv e-prints 1802.03888*.

Lundberg, S. and S.-I. Lee (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, pp. 4765–4774.

McAdam, P. and P. McNelis (2005). Forecasting inflation with thick models and neural networks. *Economic Modelling 22*(5), 848–867.

Medeiros, M. C., G. F. Vasconcelos, Á. Veiga, and E. Zilberman (2019). Forecasting inflation in a data-rich environment: the benefits of machine learning methods. *Journal of Business & Economic Statistics*, 1–22.

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters 86*(3), 373–378.

ONS (2019). Consumer Prices Indices Technical Manual. Web link here.

Owyang, M. T., J. Piger, and H. J. Wall (2015). Forecasting national recessions using state-level data. *Journal of Money, Credit and Banking 47*(5), 847–866.

Ozmen, U. and O. Sevinc (2011). Price Rigidity In Turkey : Evidence From Micro Data. Central Bank of the Republic of Turkey, Working Papers No. 1125.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review 25*(1), 221–47.

Petrella, I., E. Santoro, and L. de la Porte Simonsen (2019). Time-varying price flexibility and inflation dynamics. Unpublished Manuscript.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games 2*(28), 307–317.

Shrikumar, A., P. Greenside, and A. Kundaje (2017). Learning important features through propagating activation differences. *arXiv preprint 1704.02685*.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97*, 1167–1179.

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

Stock, J. H. and M. W. Watson (1999). Forecasting inflation. *Journal of Monetary Economics 44*(2), 293–335.

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

Stock, J. H. and M. W. Watson (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking 39*, 3–33.

Stock, J. H. and M. W. Watson (2008). Phillips curve inflation forecasts. NBER Working Paper No 14322.

Stock, J. H. and M. W. Watson (2016). Core inflation and trend inflation. *Review of Economics and Statistics 98*(4), 770–784.

Stock, J. H. and M. W. Watson (2019). Slack and cyclically sensitive inflation.

Stone, C. (1977). Consistent nonparametric regression. *The Annals of Statistics 5*(4), 595–620.

Strumbelj, E. and I. Kononenko (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research 11*, 1–18.

Thorsrud, L. A. (2018). Words are the new numbers: A newsy coincident index of the business cycle. *Journal of Business & Economic Statistics*, 1–17.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Vapnik, V. (1998). *Statistical learning theory.* John Wiley&Sons Inc., New York.

Wang, Y., B. Wang, and X. Zhang (2012). A new application of the support vector regression on the construction of financial conditions index to cpi prediction. *Procedia Computer Science 9*, 1263–1272.

Xiang-rong, Z., H. Long-ying, and W. Zhi-sheng (2010). Multiple kernel support vector regression for economic forecasting. In *2010 International Conference on Management Science & Engineering 17th Annual Conference Proceedings*, pp. 129–134. IEEE.

Young, P. (1985). Monotonic solutions of cooperative games. *International Journal of Game Theory 14*, 65–72.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology) 67*(2), 301–320.

# A Forecasting Models

### Dimensionality reduction techniques

**State space model.** We estimate a dynamic factor model in line with Kapetanios (2004) and Kapetanios and Marcellino (2009). The $N$ stationary monthly indicator variables $x_t$, i.e. the CPI item series and optionally a set of macroeconomic series described in section 2, admit a common factor structure.

The model can be written as

$$x_t = \Lambda f_t + \zeta_t \tag{9}$$

$$f_t = \Phi f_{t-1} + \eta_t. \tag{10}$$

The measurement equation (9) decomposes the observables $x_t$ into $r$ common factors $f_t$, with $r << N$, and $N$ idiosyncratic components $\zeta_t$. The factors relate to the observables via the $N \times r$ loadings matrix $\Lambda$. The idiosyncratic components follow a cross-sectionally independent Gaussian white noise process, with $\zeta_t \sim$ i.i.d$(0, \Sigma_d)$, $\Sigma_d$ being diagonal, as well as $E(\zeta_t, \eta_t) = 0$. As such, all comovement is captured by the factors, whereas the idiosyncratic components are uncorrelated with the factors and among each other for all $t$. The state equation (10) models the dynamics of the $(r \times 1)$ factors $f_t$ as an AR(p) process, with $\Phi$ being a polynomial of order $p$, and with $\eta_t \sim N(0, \Sigma_{\eta_t})$. The number of lags $p$ is set to be maximum twelve and is selected using the BIC, which suggests $p = 2$ for most specifications. We set the number of factors to $r = 5$ in the specifications where we include CPI item series predictors and to $r = 3$ when we use the smaller number of macroeconomic series predictors.[22]

The unobserved factor is estimated via the Kalman filter. For this, we specify starting values for the factors via principal components. Regarding the variance of the factors, we choose an approximate diffuse distribution of the initial state. As a by-product of the filtering recursions, the likelihood is evaluated and maximum likelihood estimation is used to estimate the parameters. The state-space representation allows the use of the Kalman filter to obtain projections of the state variables. In a second step, the factors obtained from the Kalman filter are included in forecasting regression as follows

$$\hat{y}_{t+h} = \hat{\alpha} + \Sigma_{j=1}^{p} \hat{\beta} f_t + \Sigma_{j=1}^{p} \hat{\gamma}_j y_{t-j+1}. \tag{11}$$

where the forecast of the target variable, $\hat{y}_{t+h}$ is obtained from the factors and the lags of the target variables.

[22] The Bai and Ng (2002) selection criteria suggested a high numbers of factors with a very high explained variance share. Since this does not correspond to the goal of dimension reduction, we instead select the number of factors equal to the lowest number of factors which explains 50% of the variance in the data.

**Principal Component Analysis (PCA) and regression.** is a simpler version of the model above, essentially representing a static factor model. It is widely used in the forecasting literature, having been introduced by Stock and Watson (2002c). The indicator series $x_t$ are summarized by a static factor $f_t$ so that, differently from the state space model, the dynamics of which are not modeled explicitly. It is estimated in a straight-forward manner as the first principal component from the set of indicator series. In a second step, the factor is included in forecasting regression similar to (11). The key idea is, similarly to the model above, that a small number of principal components suffices to explain most of the variability in the data, and that these components also hold the bulk of predictive power for the target variable $y_{t+h}$. We set the number of principal components, similarly to the number of factors in the state space model, to $r = 5$ or to $r = 3$ depending on the data set used.

**Partial Least Squares (PLS).** is a dimensionality reduction technique that estimates multiple regressions under a large but finite number of regressors. PLS is similar to PCR in the sense that orthogonal linear combinations of $k$ series $x_t$ are estimated and then used for prediction of $y_{t+h}$. However, instead of maximizing the share of variability in the indicator series by common components, the linear combinations are chosen such that the covariance between these linear combinations and the target variable $y_{t+h}$ is maximized. PLS is less prone to the problem of irrelevance of estimated factors to predict the target, and can outperform PCA particularly when the factor structure among the indicator variables is weak (Groen and Kapetanios, 2016). We treat the number of linear combinations as a hyperparameter and select $k = 6$ it using cross-validation from a pre-specified grid.

### Shrinkage methods

**Ridge Regression.** is a shrinkage method that penalises the residual sum of squares with the sum of squared coefficients (L2-norm). This shrinks the coefficients of those predictors with a minor contribution in terms of predictive ability of the model towards zero, albeit they never become exactly zero. As such, the Ridge regression is a dense modelling technique—it uses the full range of predictors, although assuming that the contribution of many of them might be small. Under our framework, the optimisation problem can be written as:

$$\hat{\beta}^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i}^{T} (y_i - \alpha - \sum_{j}^{N} \beta z_{ij})^2 + \lambda \sum_{j}^{N} \beta_j^2 \right\} \tag{12}$$

for given values of $\alpha$ and $\lambda \geq 0$. It is common practice to centre the values of predictors around the mean first, and not to include the constant term.[23] The parameter $\lambda$ stands for the penalty imposed on coefficients and controls its overall magnitude. We have $\hat{\beta}^{Ridge} \to \hat{\beta}^{OLS}$ as $\lambda \to 0$ which is the no penalty case, and $\hat{\beta}^{Ridge} \to 0$ as $\lambda \to \infty$. Selecting a good value for the tuning parameter $\lambda$ is crucial and is done via cross-validation.

**Least Absolute Shrinkage and Selection Operator (Lasso).** The fact that the Ridge regression includes all the $N$ parameters in the model can be a disadvantage, particularly in short sample periods and thus little degrees of freedom. The Lasso is an alternative to Ridge regression that overcomes this obstacle (Tibshirani, 1996). Lasso regressions penalise the sum of squared residuals with the L1-norm, i.e. the sum of absolute coefficients. In this case, some of the coefficients are set exactly to 0. The Lasso estimators $\hat{\beta}^{Lasso}$ are computed by solving the following optimisation problem :

$$\hat{\beta}^{Lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i}^{T} (y_i - a - \sum_{j}^{N} \beta z_{ij})^2 + \lambda \sum_{j}^{N} |\beta_j| \right\} \tag{13}$$

As such, the Lasso is a sparse modelling technique which performs shrinkage in terms of variable selection; it, thus, tends to give more parsimonious models compared to the Ridge. Again, the values of the parameters are centred, the constant term is excluded, and cross-validation is employed for the selection of the tuning parameter $\lambda$.[24]

**Elastic Net.** is a hybrid approach which combines the previous L1 and L2 penalties (Zou and Hastie, 2005) . The "naïve" estimators of the Elastic Net, $\beta^{n-EN}$ are computed by solving the problem:

$$\hat{\beta}^{n-EN} = \min_{\beta} \{ \sum_{i}^{T} (y_i - a - \sum_{j}^{N} z_{ij}\beta)^2 + \lambda_1 \sum_{j}^{N} \beta_j^2 + \lambda_2 \sum_{j}^{N} |\beta_j| \} \tag{14}$$

The naïve version of Elastic Net method finds an estimator in a two-stage procedure: First, for each fixed $\lambda_2$ it finds the ridge regression coefficients, and then a Lasso-type shrinkage is applied. This kind of estimation incurs a double amount of shrinkage which leads to increased bias and poor predictions. However, using the correction factor $1 + \lambda_2$ the prediction performance is improved and the elastic net estimators are given by $\hat{\beta}^{EN} =$

[23] The reason for this is that the Ridge regression coefficients estimates can substantially change when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the objective function.

[24] The L1-Lasso-penalty makes the solutions nonlinear in the $y_i$'s, and there is no closed form, unlike for the Ridge regression. However, there are efficient algorithms for computing the entire path of solutions as $\lambda$ varies. For example, Least Angle Regression (LARs, Efron et al. (2004)) provides an efficient algorithm for computing the Lasso estimates.

$(1 + \lambda_2)\hat{\beta}^{n-EN}$.

### Non-Linear Machine Learning Models

**Tree Models and Random Forests.** Tree-based models are a non-parametric methods for both regression and classification problems. The idea behind them is to consecutively split the training dataset until an assignment or stopping criterion with respect to the target variable into a "data bucket" or leaf is reached. Splitting the vector of predictors $z_t$ (predictors and lags of dependent variable) into $N_{leaf}$, $Z = \{Z_1, \dots Z_{N_{leaf}}\}$, the optimal estimates of the $\beta$ "coefficients" is just the average of the training target values $y^{tr}_{t+h}$ within each leaf of a tree. The regression function is

$$y_{t+h} = \sum_{m=1}^{N_{leaf}} \hat{\beta}_m I(z_t \in Z_m) + \varepsilon_t, \quad \text{with} \quad \hat{\beta}_m = 1/|Z_m| \sum_{y^{tr} \in Z_m} y^{tr}_{t+h},\, m \in \{1, \dots, N_{leaf}\}\,. \tag{15}$$

A disadvantage of regression trees is that they are not identically distributed: they are built adaptively to reduce the bias. This may lead to severe over-fitting. Ensemble approaches such as a "Random Forest" (Breiman, 2001) are routinely used to overcome this problem. A Random Forest is an ensemble of uncorrelated trees which are estimated separately. The correlation between trees in a forests is (partially) broken by building them from small-enough random samples drawn with replacement (bootstraps) from the full training sample.

The predictions of the individual trees are then averaged for a single prediction reducing variance (bagging). A general drawback of random forests, as compared to single trees, is that they are hard to interpret due to the built-in randomness with causes the differences between individual trees.

**Artificial neural networks (ANN)** are similar to linear and non-linear least squares regressions and can be viewed as an alternative statistical approach to solving the least squares problem. A standard architecture of ANNs are multilayer perceptrons (MLP), a form of feed-forward network, which we use in our analysis. The variables $z_t$ in the input layer are multiplied by weight matrices $W_i, i \in \{1, \dots, L\}$. These are the parameters of the model symbolically connecting the nodes of different layers of the network. The number of rows in each such coefficient matrix determines the number of neurons in that layer where all internal hidden layers have size $N_h$. By passing through a hidden layer, the product of inputs from the previous layer, the input layer or a hidden layer, are transformed by an activation function and passed on to the next hidden or the output layer. The output layer is linear in our case represented by the final regression coefficients $\hat{\beta}$ together with the $N_h$ output from the last hidden layer resulting in the prediction $\hat{y}_{t+h}$. The number of hidden layers $L$ determines the depth of the network, with deeper networks being generally more accurate but also needing more data for training. Formally, this can be described as

$$y_{t+h} = g(z_t, W) + \varepsilon_t = \sum_{k=0}^{N_h} \hat{\beta}_k \, g_L(g_{L-1}(g_{L-2}(\dots g_1(z_t, W_1), \dots, W_{L-2}), \beta_{L-1}), W_L)_k + \varepsilon_t \tag{16}$$

The activation function $g(\cdot)$ acts as a gate for signals and introduces non-linearity into the model. Common choices are the hyperbolic tangent, the rectified unit function (ReLU) or the logistic function. The precise form is often subject to hyperparameter tuning.

**Support Vector Machines (SVM)** Were originally introduced as a classification method based on the idea of identifying a small set of input points, the support vectors, to represent class boundaries in the classification problems (Vapnik, 1998). The model has recently gained attention among the economics and finance communities as it offers nice statistical properties and can handle and capture non-linearities in the data (Xiang-rong et al., 2010; Wang et al., 2012). A support vector regression, with a continuous target as in our case, can be written as

$$y_{t+h} = \hat{\alpha}_0 + \sum_{i=1}^{m} \hat{\alpha}_i \mathcal{K}\big(z_i^{tr}, z_t\big) + \varepsilon_t \,, \tag{17}$$

where the sum runs over the training sample. If strictly bigger than zero, the weights $\hat{\alpha}_i \geq 0$ mark the support vectors $z_i^{tr}$ jointly selected from the training data during optimisation. The Kernel $\mathcal{K}(\cdot, \cdot)$ acts like an inner product and returns a scalar. It allows the incorporation of non-linearities into the model where we use a Gaussian kernel (radial basis function, RBF). Penalisation is achieved by imposing restrictions on the $\hat{\alpha}_i$.

## Model Shapley values

The machine learning models described above are non-parametric and error consistent (Stone, 1977; Joseph, 2019), which means that they approximate any sufficiently well-behaved function arbitrarily well when provided with enough training data. But their high flexibility typically makes them difficult to interpret. In particular, it is hard to ascertain which specific variables drive model predictions and through what functional relationship they are important.

We address this issue by adopting the *Shapley additive explanations* framework (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017). It uses the concept of Shapley values (Shapley, 1953; Young, 1985) from cooperative game theory. In that context, Shapley values are used to calculate the payoff distribution across a group of players. Analogously, we use them to calculate the 'payoff' for including different predictors in our models. More precisely, each predictive value at a given horizon $h$ and using a model $f(z_t)$ is decomposed into a sum of contributions from each predictor $j = 1, ..., M$, namely its *Shapley values*. This enables us to understand which variables have large predictive value for each predictions and, when averaged, for the model as a whole.

The Shapley value framework has a set of appealing analytical properties while being applicable to any model (Lundberg and Lee, 2017). In particular, it is the only attribution framework that is local, linear, efficient, symmetric and respects null contributions and strong monotonicity of variables.[25]

As before, let $M = N + p$ be the set of $N$ predictors and $p$ lags of the dependent variable. Then, the we can define the Shapley value matrix as $\Phi_{T\times M}(z)$ corresponding to the predictor matrix $x_{T\times M}$, and $\phi_j(z_t)$ as the Shapley value of observation $t$ and predictor $j$ the individual elements. The predicted value of observation $t$ is decomposed into the sum of the Shapley values $\hat{y}_{t+h} = \sum_{j=0}^{M} \phi_j(z_t)$, where $\phi_0$ an intercept. It is taken to be the base value that is set to the mean predicted value in the training set.

For a linear regression model, the Shapley value of predictor $j$ is simply the product of its regression coefficient $w_i$ and the difference between the predictor value $z_{tj}$ and its mean, i.e. $\phi_{tj} = w_j(z_{tj} - \mathbb{E}_t[z_j])$. Computing Shapley values for a more general machine learning model is computationally more complex and is based on Shapley's work in game theory. In a cooperative game, the individual contribution within a coalition of players is not directly observable but the payoff generated by the the group as a whole is. To determine the contribution of player $j$, coalitions can be formed sequentially and $j$'s contribution can be measured by her marginal contribution when entering a coalition, which also depends on the other players in that group. Imagine player $j$ joins a coalition in which player $i$ has similar skills. In this case, $j$'s contribution is smaller than if she had joined the group when $i$ was absent. Therefore, all possible coalitions of players need to be evaluated to make a precise statement of $j$'s contribution to the joint payoff.

More formally ,with $M = N + p$ being the set of all players in the game, and $f(S)$ be the payoff of a coalition $S$. Then the Shapley value for player $j$ is computed by:

$$\phi_j = \sum_{S \subseteq M \setminus j} \frac{|S|!(|M| - |S| - 1)!}{|M|!} \left[ f(S \cup \{j\}) - f(S) \right]. \tag{18}$$

In our case, we make the analogy between the payoff and the predicted value estimated by the model for a particular observation $t$, i.e. $f(z_t) = \sum_{j=0}^{M} \phi_{tj}(z_t)$ (Strumbelj and Kononenko, 2010). The set of players $M$ correspond to the predictors used in the model.

[25] Other approaches to make variable attributions include local methods LIME (Ribeiro et al., 2016) and DeepLIFT (Shrikumar et al., 2017) and global metrics like permutation importance (Breiman, 2001; Fisher et al., 2018). But these do in general not fulfil the Shapley value properties, making them less faithful attribution methods. For example, permutation importance only measures the relative importance of the individual variables across the whole dataset and thus cannot be used to identify functional relationships learned by the models, which is particularly important for non-linear models.

It follows that the computation of the Shapley values has to be done for each individual observation for which we want to explain the predicted value. To compute the exact Shapley value of variable $j$ for observation $t$, one has to compute how much variable $j$ adds to the predictive value $(f_t(S \cup \{j\}) - f_t(S))$ in all possible subsets of the other variables $(S \subseteq M \setminus j)$. As an example, take three regressors in a linear model and the prediction $y_{t+h}$ as the payoff. We then compute all regressions with one, two and the three regressors and examine the marginal contribution of each regressor in each case. Next, we take the weighted average of marginal contributions accounting for the number of permutations of groups of one, two and three variables. The weights are given by the combinatorial factor in (18).

Contrary to a cooperative game (or linear model), predictors not in $S$ cannot be left out as this would not allow the general model to produce predictions. Instead, these predictors are integrated out using all observed values in the training set, the so-called background. That is, for each $z_t$ and $S$, values of variable components including lags not in $S$ are replaced by those in the background over which is averaged.

The evaluation of (18) can not be performed exhaustively for non-linear models given the large number of terms to be evaluated for only a modest number of predictors. We instead rely on either efficient model-specific implementations, e.g. Lundberg et al. (2018) for tree-based models, or sampling from the coalition space.

## Shapley regressions

Shapley values measure how much individual variables drive predictions of a model, independent of the overall accuracy of the model. In other words, taken in isolation, Shapley values do not show how reliably the variables actually predict the true outcome, which is a question of statistical inference.

To judge the economic and statistical significance of predictors, we use *Shapley regressions* (Joseph, 2019). To the best of our knowledge, there is no other consistent statistical framework that allows for joint testing of significance of individual predictors on non-parametric models.[26] In our context, the Shapley regression framework achieves this by regressing the target value $y_{t+h}$ on the Shapley values $\phi_j(z_t) = \phi_{tj}$ using a linear regression,

$$y_{t+h} = \hat{\alpha} + \sum_{j=1}^{M} \phi_{tj}\,\hat{\beta}_j + \epsilon_t\,, \tag{19}$$

where $\hat{\alpha}$ is the intercept. The non-linear and unobservable function of the predictors in a black box model is transformed via Shapley values into an additive, i.e. linear, parametric space which makes the estimation of p-values a simple regression exercise.

[26] Ishwaran and Lu (2019) introduce testing on variable importance in tree-based models. However, this measure does not possess all properties of Shapley values and may thus be an unreliable metric.

The coefficients $\beta$ measure the alignment between the forecast inflation and divisional Shapley components. Eq. 19 is a case of inference using generated regressors (Pagan, 1984). Valid inference requires the independence of the estimation of $\Phi$ and $\hat{\beta}$ and fast enough convergence of $\Phi$ (Joseph, 2019). This point is addressed via unbalanced sample splitting between training and test sets as is standard in machine learning applications. Potentially slow convergence of machine learning models happens on the training set, while faster convergence during the regression stage is evaluated on the test set. This ratio is taken to the conservative extreme in our case with only a single test observation at each horizon $t + h$ and training set.

# B Additional Tables and Figures

Table B1: Macroeconomic Series used as Predictors

| Code | Variable Name | Source | Transf | cat. |
|---|---|---|---|---|
| 1 | IoS: Services, Index | ONS | LD | real |
| 2 | PNDS: Private Non-Distribution Services: Index | ONS | LD | real |
| 3 | IoS: G: Wholesales, Retail and Motor Trade: Index | ONS | LD | real |
| 4 | IoS: 47: Retail trade except of motor vehicles and motorcycles: Index | ONS | LD | real |
| 5 | IoS: 46: Wholesale trade except of motor vehicles and motorcycles: Idx | ONS | LD | real |
| 6 | IoS: 45: Wholesale & Retail Trade & Repair Motor V. & M'cycles: Idx | ONS | LD | real |
| 7 | IoS: O-Q: PAD, Education and Health Index | ONS | LD | real |
| 8 | IoP:Production | ONS | LD | real |
| 9 | IoP:Manufacturing | ONS | LD | real |
| 10 | Energy output (utilities plus extraction) Pound Sterling (Index | ONS | LD | real |
| 11 | IoP: SIC07 O. Idx D-E: Utilities: El., Gas, Water Supply, Waste Mngnm. | ONS | LD | real |
| 12 | IOP: B:MINING AND QUARRYING: | ONS | LD | real |
| 13 | RSI:VolumeAll Retailers inc fuel:All Business Index | ONS | LD | real |
| 14 | Construction Output: Seasonally Adjusted: Volume: All Work | ONS | LD | real |
| 15 | BOP Total Exports (Goods) | ONS | LD | real |
| 16 | BOP Total Imports (Goods) | ONS | LD | real |
| 17 | PPI Output | ONS | LD | real |
| 18 | PPI Input | ONS | LD | real |
| 19 | Nationwide House Price MoM | BoE database | D | hp |
| 20 | RICS House Price Balance | BoE database | D | hp |
| 21 | M4 Money Supply | BoE database | LD | real |
| 22 | New Mortgage Approvals | BoE database | LD | real |
| 23 | Bank of England UK Mortgage Approvals | BoE database | LD | real |
| 24 | Average Weekly Earnings | ONS | LD | real |
| 25 | LFS Unemployment Rate | ONS | D | real |
| 26 | LFS Number of Employees (Total) | ONS | LD | real |
| 27 | Claimant Count Rate | ONS | D | real |
| 28 | New Cars Registrations | BoE database | LD | real |
| 29 | Oil Brent | BoE database | LD | fin |
| 30 | UK base rate | BoE database | L | fin |
| 31 | 3m LIBOR | BoE database | L | fin |
| 32 | FTSE all share | BoE database | LD | fin |
| 33 | Sterling exchange rate index | BoE database | LD | fin |
| 34 | GBP EUR spot | BoE database | LD | fin |
| 35 | GBP USD spot | BoE database | LD | fin |
| 36 | FTSE 250 INDEX | BoE database | LD | fin |
| 37 | FTSE All Share | BoE database | LD | fin |
| 38 | UK focused | BoE database | LD | fin |
| 39 | S&P 500 | BoE database | LD | fin |
| 40 | Euro Stoxx | BoE database | LD | fin |
| 41 | Sterling ERI | BoE database | LD | fin |
| 42 | VIX | BoE database | LD | fin |
| 43 | UK VIX - FTSE 100 volatility index | BoE database | LD | fin |

*Notes:* Sources are the Office for National Statistics (ONS), the Bank of England database (BOE), IHS Markit/CIPS, the Confederation of British Industries (CBI), LLoyds Bank, the European Commission. Transformation codes: LD = log year-on-year difference, L = levels, D = year-on-year difference. Category (Cat) codes: real = real activity, hp = house prices, fin = financial.

Table B2: Forecasting exercise results, macroeconomic series predictors, 2011-2019 sample.

**Benchmark: AR(p)**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| PCA | 0.97 | **0.77*** | **0.56***** | **0.4***** | **0.57***** | PCA | 1.17 | 0.61*** | **0.54***** | PCA | 1.17 | **0.69***** | **0.76**** |
| DFM | 0.87 | **0.69**** | **0.56***** | **0.54***** | **0.59***** | DFM | 1.05 | **0.65***** | **0.47***** | DFM | 1.02 | **0.73**** | **0.72**** |
| PLS | **0.77**** | **0.65***** | **0.56***** | **0.5***** | **0.55***** | PLS | 1.03 | **0.58***** | **0.51***** | PLS | 0.98 | **0.69***** | **0.76**** |
| Ridge | **0.74**** | **0.59***** | **0.54***** | **0.46***** | **0.42***** | Ridge | 0.88 | **0.55***** | **0.44***** | Ridge | 0.99 | **0.65***** | **0.6***** |
| Lasso | 0.81 | **0.67**** | **0.55***** | **0.52***** | **0.49***** | Lasso | 0.93 | **0.67**** | **0.47***** | Lasso | **0.9**** | **0.64***** | **0.65***** |
| Elastic | 1.11 | 0.78 | **0.64**** | **0.56***** | **0.58***** | Elastic | 1.01 | **0.71**** | **0.54***** | Elastic | 0.97 | **0.66***** | **0.8**** |
| SVM | 1.75*** | 1.07 | **0.79*** | **0.63***** | **0.63***** | SVM | 1.04 | **0.64***** | **0.58***** | SVM | 0.96 | **0.72***** | **0.66***** |
| Forest | 0.88 | **0.65***** | **0.63***** | **0.73**** | 0.88 | Forest | 1.01 | **0.62***** | **0.79*** | Forest | 1.04 | 0.9 | **0.68***** |
| ANN | 0.88 | 0.84 | **0.5***** | **0.62**** | **0.53***** | ANN | 0.98 | **0.73*** | **0.47***** | ANN | 0.99 | 0.81 | **0.59***** |

**Benchmark: PCA**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.03 | 1.3* | 1.77*** | 2.48*** | 1.76*** | AR | 0.86 | 1.63*** | 1.84*** | AR | 0.86 | 1.46*** | 1.32** |
| DFM | 0.89 | **0.89**** | 0.99 | 1.33** | 1.04 | DFM | **0.9*** | 1.06 | 0.86** | DFM | 0.88 | 1.07 | 0.95 |
| PLS | **0.79***** | **0.85**** | 1 | 1.23** | 0.96 | PLS | **0.88**** | 0.95 | 0.93 | PLS | **0.84**** | 1.01 | 1.01 |
| Ridge | **0.76***** | **0.77**** | 0.97 | 1.14 | **0.75**** | Ridge | **0.76***** | 0.89 | **0.81*** | Ridge | 0.85 | 0.94 | **0.8*** |
| Lasso | **0.83*** | 0.87 | 0.98 | 1.3** | 0.86 | Lasso | **0.79***** | 1.08 | 0.86 | Lasso | **0.78**** | 0.93 | 0.86 |
| Elastic | 1.14 | 1.02 | 1.14 | 1.38*** | 1.01 | Elastic | 0.87 | 1.16* | 0.99 | Elastic | **0.83**** | 0.95 | 1.05 |
| SVM | 1.8*** | 1.39*** | 1.41*** | 1.57*** | 1.11 | SVM | 0.89 | 1.04 | 1.06 | SVM | **0.83**** | 1.04 | 0.88 |
| Forest | 0.91 | 0.84* | 1.12 | 1.81*** | 1.55** | Forest | **0.86*** | 1.01 | 1.45* | Forest | **0.89*** | 1.31*** | 0.9 |
| ANN | 0.91 | 1.09 | 0.89 | 1.55*** | 0.92 | ANN | 0.84 | 1.19* | 0.86 | ANN | **0.85*** | 1.18 | **0.78**** |

**Benchmark: DFM**

| | *Target: headline CPI* | | | | | | *Target: Core CPI* | | | | *Target: Service CPI* | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| horizon | 1 | 3 | 6 | 9 | 12 | horizon | 1 | 6 | 12 | horizon | 1 | 6 | 12 |
| AR | 1.16 | 1.46** | 1.8*** | 1.86*** | 1.69*** | AR | 0.95 | 1.53*** | 2.13*** | AR | 0.98 | 1.36** | 1.39** |
| PCA | 1.12 | 1.12** | 1.01 | **0.75**** | 0.96 | PCA | 1.11* | 0.94 | 1.16** | PCA | 1.14 | 0.94 | 1.05 |
| PLS | 0.89 | 0.95 | 1.01 | 0.92 | 0.92 | PLS | 0.98 | 0.9 | 1.08 | PLS | 0.95 | 0.94 | 1.06 |
| Ridge | 0.86 | 0.86 | 0.98 | 0.86 | **0.72**** | Ridge | **0.84*** | **0.84*** | 0.94 | Ridge | 0.97 | 0.88 | 0.83 |
| Lasso | 0.93 | 0.98 | 0.99 | 0.98 | **0.83*** | Lasso | 0.96 | 1.09 | 1.14 | Lasso | 0.95 | 0.89 | 1.1 |
| Elastic | 1.28* | 1.14 | 1.16* | 1.04 | 0.97 | Elastic | 0.88 | 1.02 | 1 | Elastic | **0.88**** | 0.87 | 0.9 |
| SVM | 2.03*** | 1.55*** | 1.43*** | 1.18* | 1.07 | SVM | 0.99 | 0.98 | 1.23 | SVM | 0.94 | 0.98 | 0.92 |
| Forest | 1.02 | 0.94 | 1.13 | 1.36*** | 1.49** | Forest | 0.96 | 0.95 | 1.67** | Forest | 1.02 | 1.23* | 0.94 |
| ANN | 1.02 | 1.22 | 0.9 | 1.16 | 0.89 | ANN | 0.93 | 1.12 | 0.99 | ANN | 0.97 | 1.11 | **0.81*** |

Notes: Forecasts using 43 macroeconomic series as predictors. Sample period 2011-2019, out-of-sample predictions from 2015m5. Root mean squared errors, relative to AR(p) model (upper panel), PCA (middle panel), DFM (lower panel). Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. $* * *\backslash ** \backslash *$ indicates significance at $10\%, 5\%, and 1\%$, respectively. Relative RMSE that are significant at a level of 10% or lower and taking values below 1 are marked in bold.

Table B3: Comparison of Shapley-value-based model inference including macroeconomic indicators.

| | Ridge Regression | | | Random Forest | | |
|---|---|---|---|---|---|---|
| | *headline* | *core* | *service core* | *headline* | *core* | *service core* |
| division | 1 – 6 months horizon | | | | | |
| LAG | **0.14***** | **0.12***** | **0.13**** | **0.03*** | **0.08**** | 0.0 |
| | 0.01 | 0.02 | 0.01 | 0.01 | 0.03 | 0.03 |
| Food & non-alc. bev. | **0.13***** | – | – | **0.39***** | – | – |
| | 0.21 | – | – | 0.28 | – | – |
| Acl. bev. & tobacco | **0.03**** | **0.05**** | -0.01 | -0.01 | 0.04 | 0.06 |
| | 0.04 | 0.06 | 0.06 | 0.01 | 0.02 | 0.01 |
| Clothing & footwear | **0.08***** | **0.09***** | **0.2***** | **0.03*** | **0.06*** | 0.01 |
| | 0.08 | 0.14 | 0.13 | 0.01 | 0.04 | 0.06 |
| Housing & fuels | **0.13***** | **0.12***** | 0.08 | 0.02 | -0.03 | 0.06 |
| | 0.07 | 0.05 | 0.05 | 0.04 | 0.06 | 0.05 |
| Furnishing & house maint. | **0.1***** | **0.11***** | **0.12***** | **0.05***** | -0.03 | **0.1**** |
| | 0.10 | 0.12 | 0.11 | 0.02 | 0.05 | 0.03 |
| Health | -0.02 | **0.08***** | 0.06 | 0.02 | 0.03 | 0.04 |
| | 0.03 | 0.05 | 0.04 | 0.03 | 0.02 | 0.02 |
| Transport | **0.18***** | **0.13***** | **0.11**** | **0.16***** | **0.15***** | **0.13**** |
| | 0.07 | 0.09 | 0.10 | 0.07 | 0.07 | 0.11 |
| Communication | -0.01 | 0.02 | 0.03 | **0.08***** | 0.04 | **0.07*** |
| | 0.03 | 0.04 | 0.04 | 0.02 | 0.04 | 0.02 |
| Recreation & culture | **0.13***** | **0.24***** | **0.28***** | **0.07***** | **0.21***** | **0.22***** |
| | 0.10 | 0.15 | 0.15 | 0.06 | 0.12 | 0.16 |
| Education | -0.0 | 0.04 | **0.11**** | **0.04**** | -0.02 | 0.03 |
| | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.14 |
| Restaurants & hotels | **0.15***** | **0.13***** | 0.0 | **0.11***** | **0.1***** | **0.08*** |
| | 0.08 | 0.05 | 0.03 | 0.10 | 0.08 | 0.06 |
| Misc. goods & services | **0.08***** | **0.13***** | **0.14***** | **0.13***** | **0.36***** | **0.12**** |
| | 0.09 | 0.14 | 0.15 | 0.10 | 0.22 | 0.23 |
| REAL | 0.03 | **0.09***** | 0.04 | **0.12***** | **0.24***** | **0.34***** |
| | 0.03 | 0.04 | 0.04 | 0.16 | 0.20 | 0.03 |
| HP | **-0.07**** | -0.03 | 0.04 | **0.07***** | **0.18***** | **-0.11**** |
| | 0.01 | 0.01 | 0.01 | 0.06 | 0.03 | 0.01 |
| FINANCIAL | **0.11***** | **0.12***** | **0.14***** | **0.13***** | 0.04 | -0.0 |
| | 0.05 | 0.05 | 0.06 | 0.01 | 0.02 | 0.03 |
| | 7 – 12 months horizon | | | | | |
| LAG | **0.12**** | 0.02 | -0.01 | **0.16***** | **0.11***** | **0.11*** |
| | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.00 |
| Food & non-alc. bev. | **0.22***** | – | – | **0.18***** | – | – |
| | 0.19 | – | – | 0.14 | – | – |
| Acl. bev. & tobacco | **0.06***** | **0.05**** | **0.14**** | 0.0 | -0.05 | 0.07 |
| | 0.04 | 0.05 | 0.07 | 0.01 | 0.01 | 0.04 |
| Clothing & footwear | **0.04*** | **0.12***** | **0.13**** | **0.1***** | **0.08**** | **0.15**** |
| | 0.07 | 0.12 | 0.14 | 0.02 | 0.05 | 0.04 |
| Housing & fuels | **0.08**** | **0.07*** | -0.02 | **0.29***** | **0.1**** | 0.08 |
| | 0.06 | 0.06 | 0.05 | 0.06 | 0.04 | 0.07 |
| Furnishing & house maint. | **0.1***** | **0.08**** | **0.15***** | **0.13***** | **0.18***** | -0.01 |
| | 0.07 | 0.09 | 0.10 | 0.04 | 0.05 | 0.09 |
| Health | -0.02 | **0.08**** | 0.02 | 0.02 | 0.0 | **0.11*** |
| | 0.04 | 0.05 | 0.05 | 0.02 | 0.01 | 0.04 |
| Transport | **0.14***** | **0.12***** | **0.12**** | **0.14***** | **0.23***** | **0.18**** |
| | 0.06 | 0.08 | 0.08 | 0.07 | 0.07 | 0.05 |
| Communication | -0.04 | **-0.05*** | **0.09*** | 0.01 | -0.03 | 0.03 |
| | 0.03 | 0.04 | 0.04 | 0.00 | 0.01 | 0.03 |
| Recreation & culture | **0.11***** | **0.2***** | **0.27***** | **0.12***** | **0.27***** | **0.12**** |
| | 0.09 | 0.14 | 0.16 | 0.06 | 0.14 | 0.28 |
| Education | 0.02 | -0.01 | -0.06 | **0.07**** | -0.01 | 0.02 |
| | 0.01 | 0.01 | 0.01 | 0.07 | 0.04 | 0.03 |
| Restaurants & hotels | **0.23***** | **0.2***** | 0.07 | **0.15***** | **0.2***** | **0.22***** |
| | 0.10 | 0.05 | 0.05 | 0.19 | 0.17 | 0.10 |
| Misc. goods & services | **0.13***** | **0.19***** | **0.14**** | **0.2***** | **0.26***** | **0.17***** |
| | 0.12 | 0.14 | 0.14 | 0.20 | 0.26 | 0.13 |
| REAL | 0.0 | **0.1***** | **0.09*** | **0.07**** | **0.2***** | -0.01 |
| | 0.05 | 0.06 | 0.04 | 0.11 | 0.10 | 0.05 |
| HP | **-0.12***** | -0.06 | **-0.11**** | 0.01 | -0.03 | 0.02 |
| | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.04 |
| FINANCIAL | **0.09***** | **0.19***** | 0.06 | 0.01 | -0.0 | -0.02 |
| | 0.05 | 0.08 | 0.05 | 0.01 | 0.02 | 0.02 |

Notes: Summary of Shapley regression (8) for Ridge Regression (LHS) and Random Forest (RHS) models for headline, core and service core inflation. CPI models components for different horizons are grouped. The share of each aggregate component (6) is given below each coefficient. Core and service core targets do not contain item components from food and non-alcoholic beverages. Macroeconomic aggregate components (REAL, HP FINANCE) correspond to Table B1. Significance levels: ***:1%, **:5%, *:10%. Panel-HAC standard errors grouped by forecast horizon have been used. Source: ONS and authors calculations.

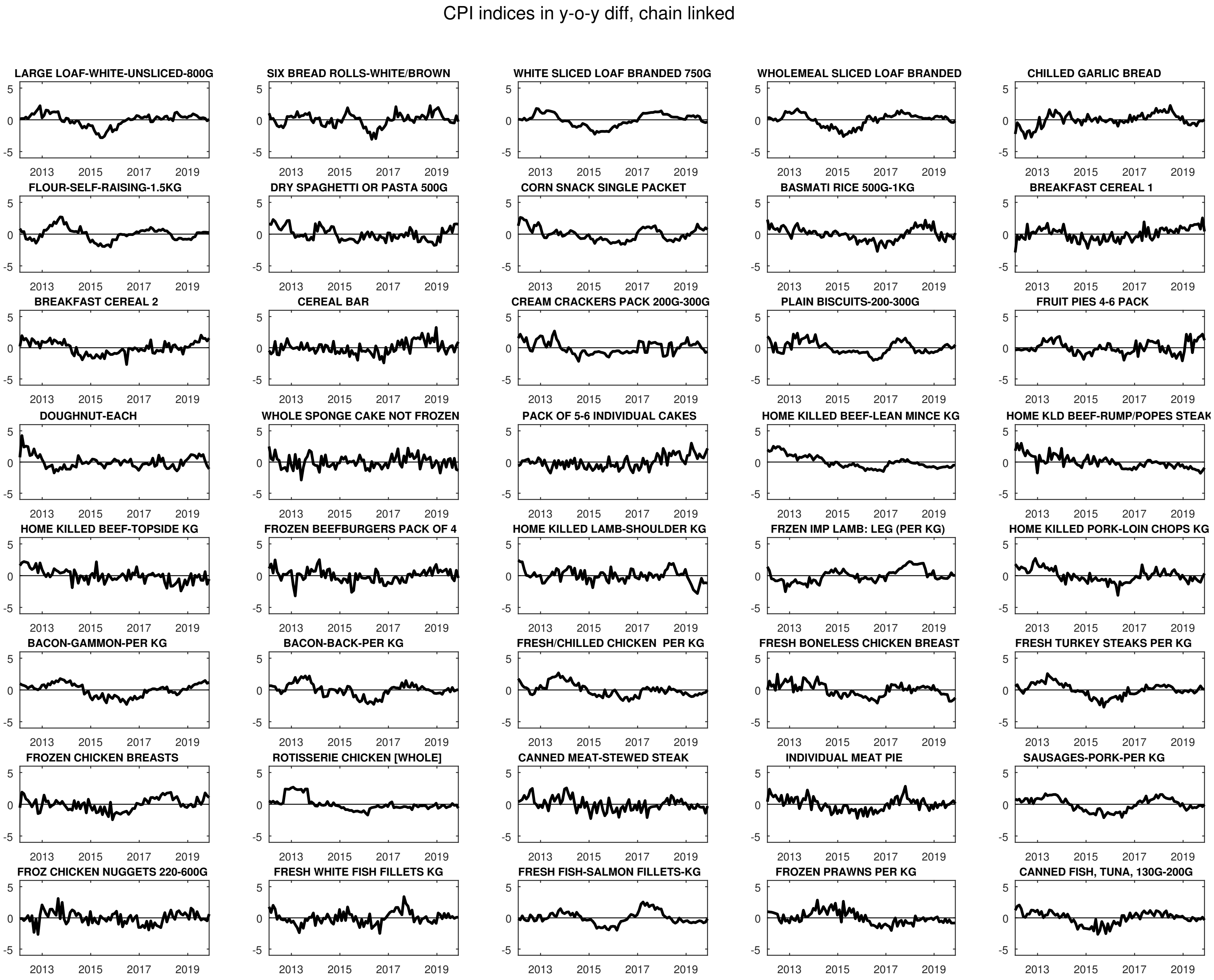

Figure B1: Selected item series. Notes: Data in year-on-year growth rates, standardised. Item identifiers No. 210102 to No. 211207.

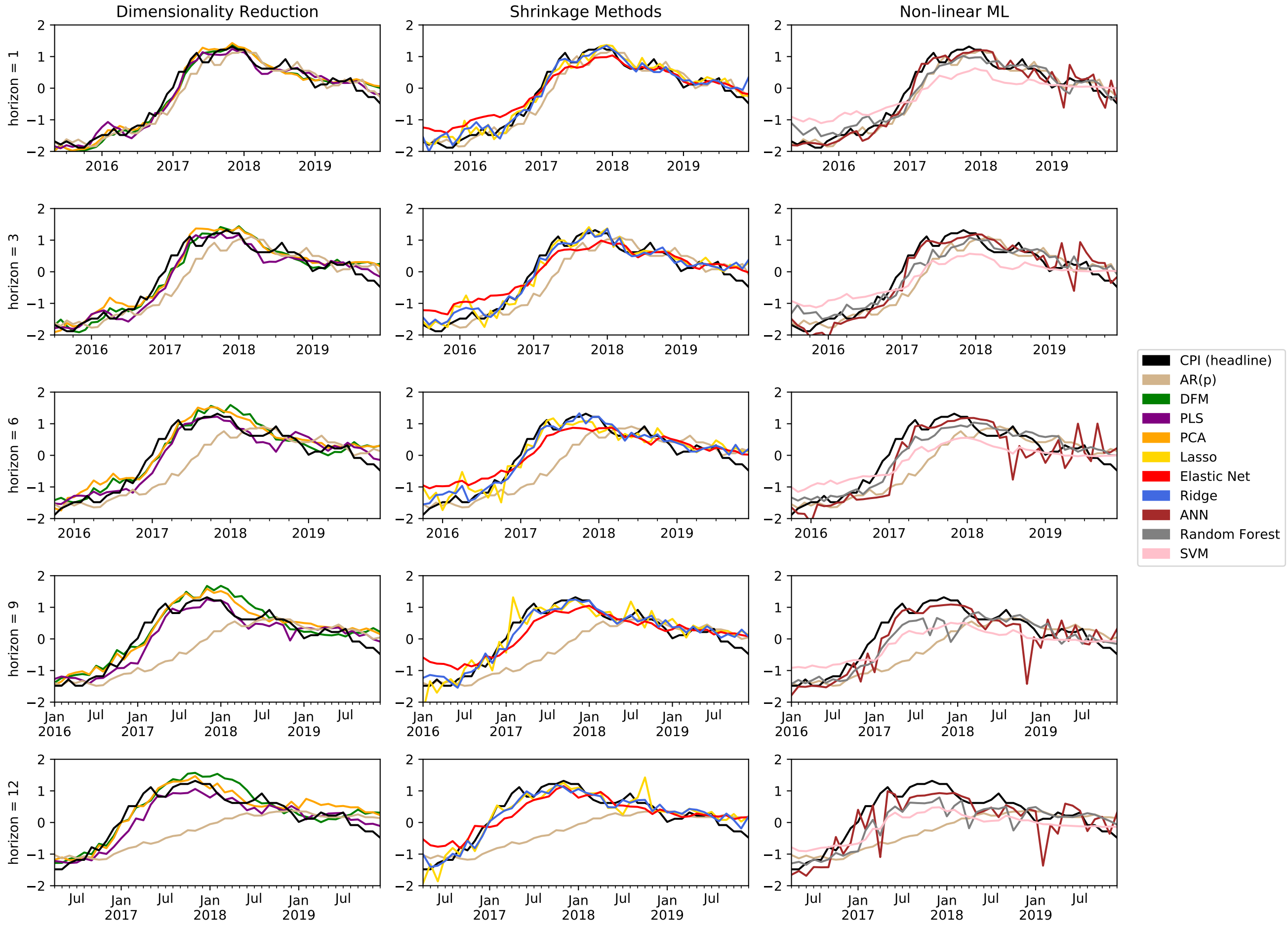

Figure B2: Predicted values for headline CPI inflation, models with CPI items and macroeconomic series.

Notes: Forecasts of CPI headline inflation (standardised), from different types of forecasting models (columns, coloured lines) using CPI items and macroeconomic series as predictors, for different horizons $h$ (rows). Out-of-sample predictions for 2015m5 to 2019m12. Compared to the actual headline CPI inflation outcome (black lines), lagged by $h$ months.

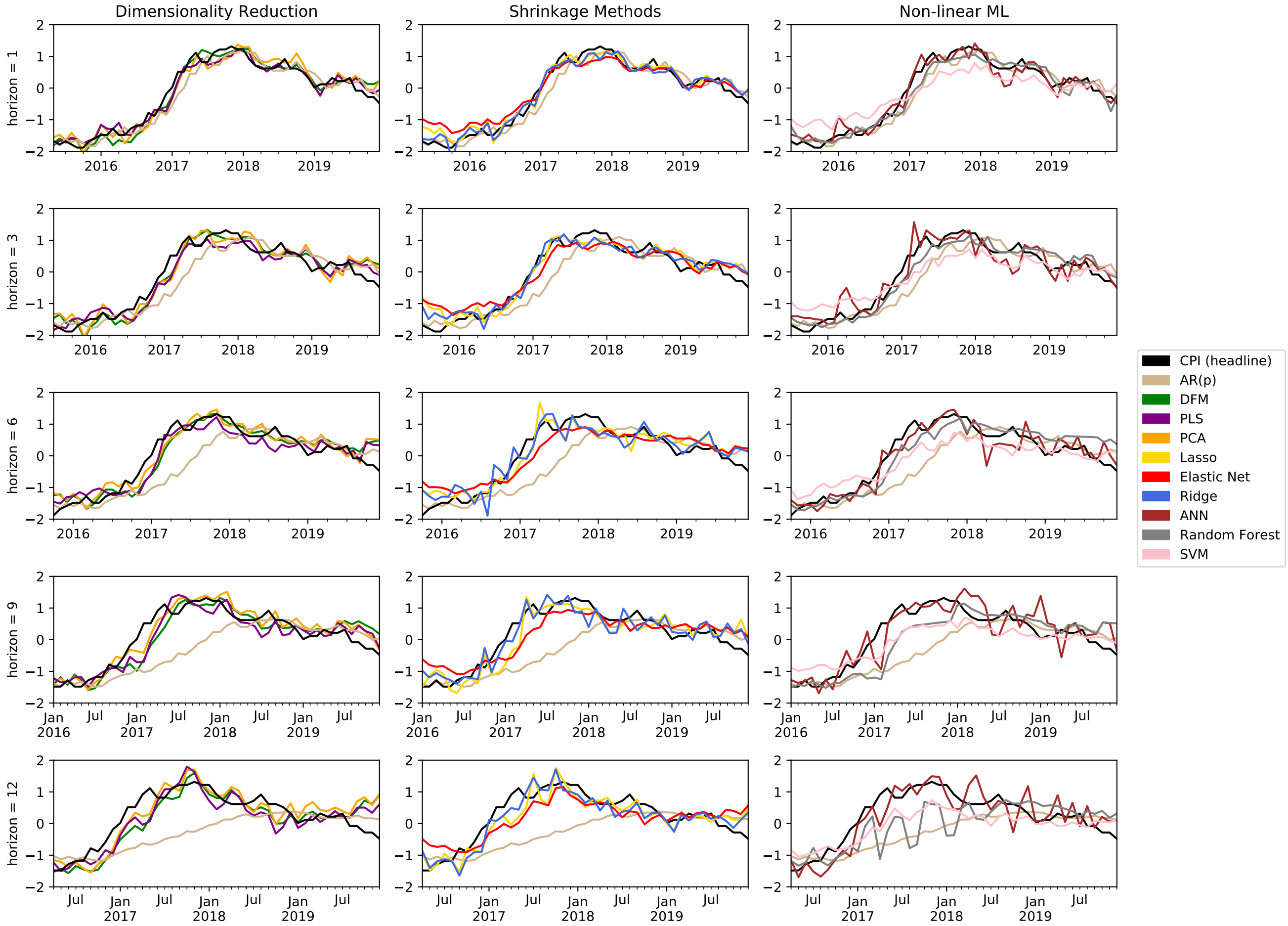

Figure B3: Predicted values for headline CPI inflation, models with macroeconomic series, 2011-2019.

Notes: Forecasts of CPI headline inflation (standardised), from different types of forecasting models (columns, coloured lines) using macroeconomic series as predictors, for different horizons $h$ (rows). Out-of-sample predictions for 2015m5 to 2019m12. Compared to the actual headline CPI inflation outcome (black lines), lagged by $h$ months.