International Journal of Forecasting 40 (2024) 1521–1538

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting UK inflation bottom up☆

Andreas Joseph [a,b,*,1], Galina Potjagailo [a], Chiranjit Chakraborty [a], George Kapetanios [c]

[a] Bank of England, United Kingdom
[b] DAFM, United Kingdom
[c] King's College London, United Kingdom

ARTICLE INFO

Keywords:
Inflation
Forecasting
Machine learning
State space models
CPI disaggregated data
Shapley values

ABSTRACT

We forecast CPI inflation indicators in the United Kingdom using a large set of monthly disaggregated CPI item series covering a sample period of twenty years, and employing a range of forecasting tools to deal with the high dimension of the set of predictors. Although an autoregressive model proofs hard to outperform overall, Ridge regression combined with CPI item series performs strongly in forecasting headline inflation. A range of shrinkage methods yields significant improvement over sub-periods where inflation was rising, falling or in the tails of its distribution. Once CPI item series are exploited, we find little additional forecast gain from including macroeconomic predictors. The forecast performance of non-parametric machine learning methods is relatively weak. Using Shapley values to decompose forecast signals exploited by a Random Forest, we show that the ability of non-parametric tools to flexibly switch between signals from groups of indicators may come at the cost of high variance and, as such, hurt forecast performance.

1. Introduction

Forecasting consumer price inflation accurately in the near and medium term has considerable implications for monetary policy, other policy choices, and business decisions in the wider economy. Inflation forecasts recently moved to the forefront of the policy debate. Accurate inflation forecasts are crucial for central banks to design appropriate and timely policy responses and to communicate the path at which inflation is expected to return to target—particularly following the recent rise above inflation targets in many advanced economies. However, forecast performance can vary with the state of the economy (Odendahl, Rossi, & Sekhposyan, 2022). Forecast mistakes can be large around turning points or periods of high inflation since the time series process of inflation and its relationship with macroeconomic predictors can become unstable. Drawing information from disaggregated price dynamics across different sectors might be particularly useful since this may help to detect broad-based increases across items and turning points early on. Non-linear and non-parametric models may help deal with large changes in the predictors and the macroeconomic variables of interest.

This paper explores the forecasting gains for aggregate inflation measures from this angle: we use a unique large set of disaggregated item index series comprising the consumer price index (CPI) and a range of forecasting approaches, including novel machine learning tools, to forecast aggregate inflation measures.

☆ Disclaimer: The views expressed in this paper are solely those of the authors and do not necessarily represent those of the Bank of England.

* Corresponding author.
E-mail addresses: andreas.joseph@bankofengland.co.uk (A. Joseph), galina.potjagailo@bankofengland.co.uk (G. Potjagailo), chiranjit.chakraborty@bankofengland.co.uk (C. Chakraborty), george.kapetanios@kcl.ac.uk (G. Kapetanios).
[1] Data Analytics for Finance and Macro (DAFM) Research Centre, King's College London, United Kingdom.

https://doi.org/10.1016/j.ijforecast.2024.01.001
0169-2070/

We forecast monthly CPI headline, core, and service inflation in the United Kingdom at 3–12 months ahead horizons. As predictors, we use a large set of monthly CPI items and, for comparison, a set of standard macroeconomic indicators. We evaluate a range of forecasting methods that exploit this large information set in different ways: dimensionality reduction techniques (Principal Component Analysis, Partial Least Squares), shrinkage methods (Ridge, Lasso and Elastic Net regressions), as well as non-linear machine learning tools (Support Vector Machines, Artificial Neural Networks, and Random Forests). We consider 2002m1-2021m11, evaluating the models using rolling window pseudo-out-of-sample forecasts against an autoregressive (AR) benchmark. The original sample of CPI items is unbalanced, with items entering and dropping from the sample in accordance with their presence in a representative household's consumption basket. We train our models and run forecasts over rolling sample periods of 7 years, which assures balanced panels of items, with, on average, more than 500 items entering the models for a given forecast, thereby also tracking the changing composition of consumption.

The contribution of the paper is three-fold. First, we assess the forecasting gains by considering disaggregated item-level information. Item-level prices (e.g. "cereal bar", "light bulb", "cinema admission") matter for aggregate inflation since they directly form the basis for the aggregate consumer price index. The dynamics and interdependencies 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 overrepresented in the aggregate (Chu, Huynh, Jacho-Chávez, Kryvtsov, et al., 2018; Petrella, Santoro, & de la Porte Simonsen, 2018; Stock & Watson, 2020). Thus, incorporating item indices directly into forecasts allows us to exploit a rich information set (Hendry & Hubrich, 2011).

Second, we compare forecast performance across a range of models that represent different approaches to tackle the large dimension and disaggregation of the data. We compare well-established linear approaches, such as principal component analysis and shrinkage methods, with machine learning tools that are potentially stronger in detecting turning points and complex dynamics in item data due to their flexibility to learn unknown functional forms. To assess potential non-linearities in forecasting performance, we evaluate forecasts over sub-periods for which the aggregate inflation measure displayed certain characteristics, such as rising, falling, high or low inflation.

Third, the use of disaggregated data can help to communicate adjustments to forecasts based on dynamics observed across sub-components. To enhance the intuition of our results, we measure the contribution of individual items to forecasts using Shapley values (Lundberg & Lee, 2017; Strumbelj & Kononenko, 2010) and re-aggregate those into contributions from groups of items according to tractable CPI categories.

Our findings are as follows. Over the entire sample period, it is hard to beat the AR benchmark significantly, even when exploiting large data, and uncertainty in the forecasts is relatively wide, with considerable variation across models and horizons. Shrinkage methods combined with disaggregated CPI items provide some improvement at 3–6 months horizons, particularly Ridge regression for headline inflation and LASSO for core inflation forecasts. When evaluating the forecasts over sub-periods during which the aggregate inflation measure we forecast is rising or falling or at its tails, a wider range of significant improvements against the benchmark is observed, mainly for shrinkage methods. Ridge regression, Elastic Net and Lasso achieve improvements at different horizons, subperiods, and for different targets and at longer horizons of 6 and 12 months. As such, it is advisable to consider a range of shrinkage methods with varying forms of penalty and potentially model averaging – the variation across models and horizons suggests considerable model uncertainty. Dimensionality reduction techniques that rely on co-movement in the data and non-linear machine learning tools are rarely able to beat the benchmark – even during sub-periods when inflation dynamics might be unstable. We also find that adding macroeconomic indicators yields little additional forecast gain once disaggregated item series are controlled.

To understand forecast results further, we unpack the forecast signals exploited by the Ridge regression, the best-performing linear model, and the Random Forest, a machine learning model. We rely on Shapley values representing the 'payoff' from including a specific predictor in the model, conditional on other predictors being present (Lundberg & Lee, 2017). We compute Shapley weights that reflect the relative contribution to a given forecast over time associated with a group of predictors—such as core goods and service items, more volatile item categories, and macroeconomic indicators. Results show that the linear model exploits mostly stable signals and largely reads the strongest signals from the largest groups of predictors while relying much less on smaller components. The Random Forest instead assigns volatile weights, shifts between signals from groups of predictors frequently, and during certain periods assigns strongly overproportional weights to volatile categories such as energy items. While the shifts in Shapley weights exploited by the Random Forest occur during periods when inflation dynamics were indeed changing, such as during periods of low energy prices or in the aftermath of the Covid-19 pandemic, their high volatility makes them hard to interpret, and the additional flexibility does not translate into a stronger forecast performance.

Over the period of strongly rising inflation at the end of the sample period, both models, but to a larger extent the machine learning model, shift towards overproportionally relying on macroeconomic indicators and away from core goods and food price items. These item groups showed strong fluctuations around that period related to supply shortages in the aftermath of the pandemic; however, the models seem to identify these fluctuations as noise and down-weigh them, having not observed inflation rises of such size and nature over the sample period.

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 & Watson, 1999, 2008), univariate unobserved component models (Stock & Watson, 2007, 2016a), aggregation of forecasts of subcomponents (Hubrich, 2005), Bayesian VARs (Domit, Monti, & Sokol, 2019; Koop, 2013), dimensionality reduction (Kim & Swanson, 2018) and medium-sized DSGE models (Carriero, Galvao, & Kapetanios, 2019). With regard to machine learning tools and non-parametric approaches, earlier studies find that forecasts of US inflation with neural networks outperform autoregressive or random walk benchmarks at different horizons (Almosova & Andresen, 2023; Chen, Racine, & Swanson, 2001; McAdam & McNelis, 2005; Nakamura, 2005). Closer to our approach, Garcia, Medeiros, and Vasconcelos (2017) and Medeiros, Vasconcelos, Veiga, and Zilberman (2021) forecast Brazilian and US CPI inflation, respectively, using large sets of macroeconomic predictors with various methods, where for the US, the Random Forest performs best. Clark, Huber, Koop, and Marcellino (2022) find that a non-parametric specification of the conditional mean and innovations in US inflation using Gaussian process regression and Dirichlet process mixture achieves gains for point and density forecasts, particularly during the volatile period of the Covid-19 pandemic and in predicting left-tail risks. Similarly, Hauzenberger, Huber, and Klieber (2023) provide evidence that non-linear dimension reduction techniques with shrinkage priors improve US inflation forecasts in real time, and non-linear models are particularly useful during recessionary episodes.

Our analysis also relates to studies that have used disaggregated data to forecast aggregate series. Hernández-Murillo and Owyang (2006) and Owyang, Piger, and Wall (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. Beck, Carstensen, Menz, Schnorrenberger, and Wieland (2022) use high-frequency scanner data for product-level prices and quantities to nowcast and map German inflation, showing that high-frequency disaggregated information provides timely signals for inflation early in the month. Most closely 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.

The remainder of the paper is organised as follows. Section 2 describes the data used in the forecasting exercise and introduces the CPI item series data set. Section 3 describes the forecasting set-up and gives a brief model overview. Section 4 presents the forecast results for the entire sample period and over sub-periods during which inflation displayed certain characteristics. Section 5 addresses the black-box critique of our high-dimensional forecasting setting through Shapley value-based inference. Section 6 concludes.

2. Data

We use the headline CPI index from the UK's Office for National Statistics (ONS), transformed to year-on-year inflation rates, as the main target variable in our forecasting exercise. Additionally, we consider CPI core inflation that corresponds to the CPI headline index, excluding the generally more volatile food and energy components, as well as CPI core service inflation based on CPI indices of twelve service categories, excluding goods and more seasonally volatile services.[2] These inflation measures represent the less volatile component of consumer prices and are typically considered to be more closely linked to underlying and domestically generated price pressures.

Our main interest lies in exploring the predictive gain from using a large set of CPI disaggregated item series published by the ONS, which we describe in more detail below, to forecast aggregate inflation. Additionally, we explore forecast gain from a set of 46 macroeconomic series, selected to represent broad categories of UK economic and financial activity: unemployment and hours, real measures for retail trade, manufacturing and sales, international trade, labour costs, house price indexes, interest rates, stock market indicators, exchange rates, and import prices. Several studies have shown the predictive power of such macroeconomic data sets in forecasting inflation (Stock & Watson, 2002a, 2002b). This data is also readily available over longer sample periods and is continuously monitored by central banks and professional economists. Before estimation, the series are transformed to year-on-year log differences to achieve stationarity and are standardised (see Table A1 in the online appendix).[3]

2.1. CPI item series

The 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, guide 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

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

[3] CPI aggregate and item series are not revised after first publication. Since this study uses CPI item series as predictors, we do not account for real-time data issues with macroeconomic data and use the final data release.

collection in case of national prices. That is, item prices connect the disaggregated indices and aggregate inflation, which we exploit in this paper. The ONS constructs the UK CPI from an evolving set of representative monthly item indices weighted according to household expenditure shares. Single-item prices, or price quotes, are aggregated at the lowest level into item-level indices.[4] 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.

We use monthly item series from January 2002 until November 2021. There are overall over 1400 item indices over the total sample. However, many item indices do not cover the full sample period since, for each month, the ONS publishes only the 630–710 items that enter the consumption basket and, thus, the aggregate CPI at that point in time. Particularly over the first years of the sample, there were substantial changes in the basket, with items entering and dropping out of the basket frequently. This highly unbalanced data structure is challenging since we require a balanced sample of items for our estimations. If we were to pick those items that cover the full sample, we would be left with 280 item indices that are not representative of the consumption basket, particularly towards the end of the sample. We, therefore, opt to approximately imitate the evolving nature of the composition of different goods in aggregate CPI by running estimations over rolling windows of item samples. We choose a window length of eight years. Hence, we start with an initial balanced sample of items from 2002 to 2009. We then iterate the sample forward, with items discontinued at the end of the rolling window dropping out and new items fully covered over the rolling eight-year window entering at each iteration step. Items that do not have coverage for those eight years are dropped from that sample. For each estimation window, we estimate our models on the first seven years, and we use the last 12 months as the test sample to run out-of-sample forecasts—as such, we make sure that we use the same sample of items for training and testing at each point in time. As we iterate forward, the composition of our predictors evolves, mimicking the change in the consumption basket. On average, more than 400 item indices are included in a window, suggesting good overall coverage. Since there are more frequent changes in the basket at the early part of the sample period with more discontinued item series, the rolling estimation sample starts with 386 items for the window 2002–2009 and then gradually becomes larger until reaching a more stable size of 540–570 items for the later windows. Figure A1 in the online appendix depicts the evolving sample size for the eight-year sample window and two alternative window sizes. We face a trade-off when fixing the window length: a smaller window size implies a closer representation of the consumption basket with more items covered in each window, but it also gives a shorter training sample.[5]

We chain-link the item indices, and we take year-on-year log differences. The latter should remove stochastic seasonality where it is present without imposing seasonal filtering on series that might not show seasonal patterns.[6] The log transform should also help smooth extreme observations. Still, we also checked the robustness of results when replacing outlier observations that lie beyond six times the interquartile range, and the results were very similar. Item series are mean–variance standardised in line with the expanding window approach of our forecasting setting described in Section 3.

2.2. Descriptive statistics

We provide descriptive statistics for our disaggregated data to better understand how the item series dynamics compare to the aggregate CPI. Table 1 assesses the representativeness of our sample of item series. It summarises statistics of year-on-year item-level index growth rates grouped by divisions, the twelve largest sub-categories of the CPI using the final release classification, with their weights depicted in column 3 (November 2021; see also ONS (2019)).[7] The middle panel of the table compares the average number of items in our balanced panel based on 8-year rolling window estimations (column 5) to the average number of items available in each category per year in the unbalanced panel (column 4).[8] The series included in the balanced panel covers, on average, 69% of each division's item indices. The right panel of the table shows the median and standard deviation of yearly changes in our chained-linked index series. The median across items for most CPI divisions is comparable to average aggregate year-on-year price inflation, with some deviations for the categories "Clothing & footwear" and "Education". However, the standard deviations across items are relatively large for most categories, pointing to heterogeneity in the disaggregated data.

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

[5] We ran estimations for the window length of six years with five years used for training. Results were similar, though somewhat less significant, due to the shorter training sample. Alternatively, we ran an expanding window estimation, starting with an initial training period of 2002 to 2008 and then expanding it gradually, such that the number of items covered decreased over time and became less representative of the consumption basket, covering 280 items for the longest training period. This resulted in weaker forecasting gains compared to the rolling window approach. This indicates that tracking the composition of the consumption basket in more detail benefits forecasting.

[6] Figure A2 in the online appendix plots a selection of the transformed item series. We've also looked at autocorrelation functions to understand the statistical properties of the transformed series, which indicate that most year-on-year log differenced item series follow autoregressive processes, whereas for month-on-month log differenced item series, the autocorrelation we observe strong seasonal components for some series but no seasonality for others.

[7] A set of zero-weight indices not in the CPI have been added to Housing & Fuel (440 249, 410 201, 410 701, 410 703, 410 801, 440 202, 610 307, 610 308).

[8] Average numbers of series by divisions are not integers due to series dropping in and out over time, and due to the number of items having full coverage over the rolling window increasing over time.

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

	Description	Weight (%)	#items, unbalanced	#items, balanced	Coverage (%)	Median	SD
1	Food & non-alc. bev.	12	155.41	111.6	72	1.74	7.31
2	Acl. bev. & tobacco	5	26.57	15.2	56	2.02	4.23
3	Clothing & footwear	7	77.63	54.6	70	−0.71	5.59
4	Housing & fuels	13	37.04	30.3	82	2.67	6.42
5	Furnishing & house maint.	6	72.85	53.1	73	1.34	5.02
6	Health	3	20.09	14.7	73	1.77	4.28
7	Transport	14	43.58	31.5	72	2.46	7.27
8	Communication	3	9.18	5.6	61	1.91	10.69
9	Recreation & culture	15	112.35	64.6	56	1.41	8.07
10	Education	2	3.05	2.1	69	6.71	5.44
11	Restaurants & hotels	9	48.36	31.8	66	2.85	1.88
12	Misc. goods & services	11	76.13	52.4	69	1.65	8.42
13	Total	100	682.2	467.5	69	2.15	6.22

Notes: Division-level summary statistics of year-on-year percentage changes of item series. CPI weights (%) are taken from COICOP weights for November 2021. The total number of items (#), unbalanced, refers to all item series available on average between January 2002 and November 2021 in that division in the unbalanced panel published by the ONS. Note that this number does not need to be an integer because items enter and exit the CPI basket over time. The number of balanced items refers to those included in our sample since they cover at least the eight-year rolling window length. Coverage (%) is the fraction of our included set of items to all items. Median and standard deviations (SD) are taken over all observations in the balanced panel in a division. Source: ONS & authors' calculation.

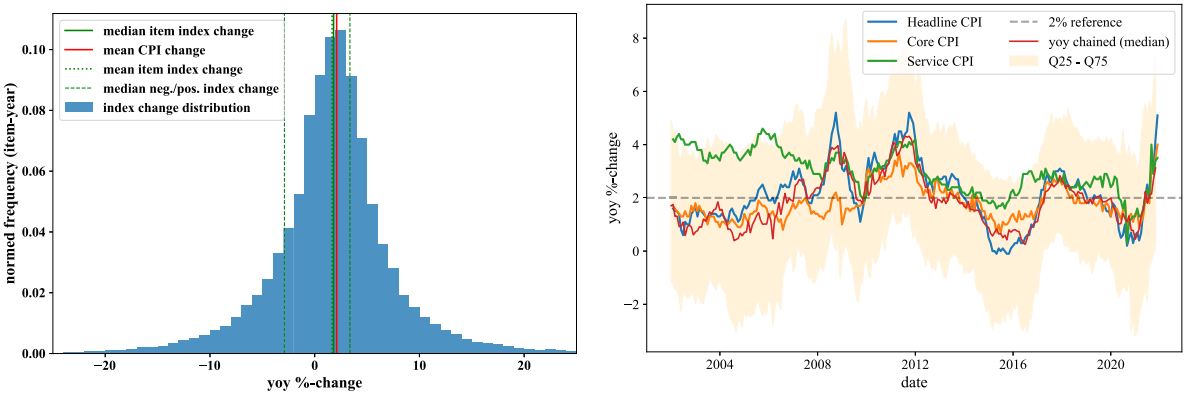

Fig. 1. Distribution and moments of year-on-year growth rates in item-level CPI indices.
Notes: Statistics are computed over chain-linked items included in our rolling window estimations. The left panel shows the histogram of CPI item growth rates over the entire sample period (blue bars), the overall mean and median over items (green solid and dotted lines), and the mean over negative and positive item growth rates (green dashed lines), the mean headline year-on-year CPI inflation for comparison (red solid line). The histogram bars are limited to ±25% for clearer presentation with a small number of changes beyond this range. The right panel shows the median (orange line) and interquartile range (orange swathe) of year-on-year growth rates of item indices over time, and for comparison, year-on-year changes of headline inflation (blue line, mean: 2.10%), core inflation (orange line, mean: 1.80%), and service inflation (green line, mean: 2.97%). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
Source: ONS and authors' calculation.

This also becomes evident in Fig. 1, which shows the distribution and moments of item series growth rates over the entire sample (left panel), as well as the evolution of the median and interquartile range over time compared to aggregate headline, core and service CPI inflation (right panel). As previously documented (Klenow & Kryvtsov, 2008; Ozmen & Sevinc, 2011), the distribution of disaggregated price changes has 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. In line with Table 1, the median of item index growth rates is close to average headline inflation, both on average over the entire sample as well as over time. There is, however, much heterogeneity across item dynamics as captured by the wide tails of the histogram (left panel) and the wide swathe representing the interquartile range across items over time (right panel).

The fit of the median across item indices to aggregate headline inflation improves over time (right panel, red vs. blue lines), in line with the improved coverage of item series through our rolling windows. During 2014–2016, core inflation lay above headline inflation related to a decline in energy prices; the median of the item series partially reflects that and lies closer to aggregate core inflation during these years. This suggests that the predictive power of item series might vary over time for the different aggregate inflation variables we forecast. Service inflation lay clearly above headline inflation in the first part of the sample and appears less associated with the dynamics of item series. However, during the recent rise in inflation in 2021, all inflation aggregates and the median and the entire distribution of item series jointly moved up substantially.

3. Methodology

3.1. Forecasting set-up and evaluation

We forecast monthly aggregate year-on-year UK CPI inflation (Headline, Core, or Services) y_{t+h} at horizons of $h = 1, \ldots, 12$ months. We start with an initial training sample for 2002m1-2008m12, over which we tune hyperparameters. We evaluate out-of-sample forecasts at the h-month horizon. We then iterate the training sample forward by one month in a pseudo-real-time setting, where we also adjust the composition of CPI items to ensure a balanced and representative sample in each estimation window, and we repeat the procedure of hyperparameter tuning and out-of-sample forecasting.

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

$$y^{AR}_{t+1} = \Sigma^{p}_{j=1}\gamma_j y_{t+1-j} + \epsilon_{t+1} \quad (1)$$

The forecasts at horizon h, y^{AR}_{t+h}, are constructed recursively once the model is fit. The number of lags is set to $p = 2$.

We evaluate the average precision of the forecasts against the $AR(2)$ benchmark based on relative root mean squared errors (RMSE), i.e. RMSE(m)/RMSE(AR). We test for statistical difference in forecast accuracy using the Diebold and Mariano (1995) test with Harvey's correction for short samples (Harvey & Newbold, 2000).[9] We run the forecast evaluations over the test period 2009m1 to 2021m11 and over sub-sets of months where aggregate outcomes (realised or predicted) fall into certain regimes at the point of the forecast, e.g. forecasts during periods of high or low headline inflation.

3.2. Separating non-linear signals from disaggregated data from lagged inflation dynamics

The forecasting methods we employ—outlined in detail in the following sub-section—can incorporate large data sets with a large number of predictors $x_t = (x_{1t}, \ldots, x_{Nt})'$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where typically $N > T$. However, lagged inflation dynamics will be assigned a prominent role in any model as inflation follows a persistent process. This can make picking up additional and non-linear information from disaggregated data challenging. Therefore, univariate models such as those proposed by Stock and Watson (2007) and Faust and Wright (2013) are typically difficult to beat. Multivariate linear models have achieved some success by separating a time-varying inflation trend or persistent component that reflects the slow-moving role of lagged dynamics and a cyclical component that can link inflation to a range of indicators (Bańbura & Bobeica, 2023; Stock & Watson, 2016b); in these models, while a smooth form of time-variation is assumed, at each point in time the link between inflation and both its trend and cycle is linear. By contrast, splitting out a trend and cycle within highly non-linear and non-parametric models is challenging, and controlling for lagged dynamics can blur the signals from the different components.[10]

Therefore, we opt for a simpler but flexible two-step approach that is ad-hoc but flexible enough to separate the persistent inflation component from the information derived from disaggregated data for a wide class of models. Specifically, we compute the forecast y^{AR}_{t+h} using the AR(2) model, and we compute the residual ϵ_{t+h} from that model. This strips off the part that is explained by the autoregressive component. Then, we use the residual as the target for the models outlined in the next section. In this, we are interested in forecasting only the component of inflation that is *not* already accounted for by autocorrelation:

$$\epsilon_{t+h} = g^{m}_{t+h}(\mathbf{x_t}, \beta) + u^{m}_{t+h}. \quad (2)$$

Model m includes N predictors $\mathbf{x_t}$, such as a set of CPI items and macroeconomic indicators. It processes them linearly or non-linearly via function $g(\cdot, \beta)$, with β being the respective set of parameters to be optimised using the training data. To retrieve the forecast of inflation at horizon h, we add the predicted component from model m to the predicted component from the AR(2), such that the full dynamics of inflation are reflected in our forecast,

$$y^{m}_{t+h} = y^{AR}_{t+h} + g^{m}_{t+h}. \quad (3)$$

This two-step approach helps to isolate the forecast gain from additional indicators beyond the part explained by lagged dynamics. This encourages the models to process both components prominently and to conduct dimension reduction of the large set of indicators more efficiently. In line with this, we find worse performance than an auto-regressive benchmark for all models when running the analysis directly on inflation with all components included jointly. The limitation of our approach, however, is that it does not account for non-linearities in the role of inflation persistence nor for non-linear interactions between the disaggregated data and lagged dynamics. Extending non-parametric methods in high-dimensional settings in this direction remains an important avenue for future research but is outside the scope of our empirical exercise.

3.3. Forecasting methods

Due to the large number of predictors, estimating (2) directly with each predictor included individually, with the dimension of β potentially much larger than T, would lead to over-parametrisation and high estimation uncertainty. Some form of dimensionality reduction is thus required (Bok, Caratelli, Giannone, Sbordone, & Tambalotti,

[9] Note that models AR and m are nested such that the validity of our test is guaranteed by the rolling window approach (Giacomini & White, 2006).

[10] Clark et al. (2022), Clark, Huber, Koop, Marcellino, and Pfarrhofer (2023) address this by employing Bayesian Additive Regression Trees (BART) that allow for non-linear relationships among macroeconomic variables as well as with their lags. However, for the machine learning models considered here, such additive treatments of various non-parametric components are less developed.

2018). 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).[11]

Dimensionality reduction techniques

We consider two forecast approaches that rely on dimensionality reduction techniques: Principal Component Analysis (PCA) and Partial Least Squares (PLS). These methods exploit the fact that economic series are often strongly correlated and thus can be summarised 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. The N indicator series $\mathbf{x_t}$ is summarised by K finite latent components f_k, and the latent components are included in the forecast regression instead of the indicators themselves. PCA summarises the joint variability of predictors x_t into a static factor, which is added into a prediction regression

$$\epsilon^{PCA}_{t+h} = \Sigma^{K}_{k=1}\beta_k f_{kt} + u^{PCA}_{t+h}\,. \tag{4}$$

On the other hand, PLS is a static model that combines predictors into a common component such that the covariance between the component and the target variable is maximised. The two models commonly use information densely, i.e., information from a wide range of available predictors is drawn upon by summarising them through common components.

Shrinkage methods

Shrinkage methods produce linear combinations of the regressors, where coefficients with little predictive power for the target variable are assumed to approach zero or are set equal to zero, according to a shrinkage parameter λ, which differs across models.

Ridge regression penalises the residual sum of squares with the sum of squared coefficients (L2-norm) (Hoerl & Kennard, 1970). This shrinks the coefficients of those predictors with a minor contribution in terms of the predictive ability of the model uniformly towards zero, albeit not exactly zero. The optimisation problem is:

$$\hat{\beta}^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{t}(\epsilon_{t+h} - \mathbf{x_t}\beta)^2 + \lambda \sum_{i}^{N} \beta_i^2 \right\} \tag{5}$$

for given values of $\lambda \geq 0$. In the case of no shrinkage, i.e., $\lambda = 0$, Ridge regression becomes equivalent to a linear OLS regression.

On the other hand, LASSO regression penalises the sum of squared residuals with the L1-norm, i.e., the sum of absolute coefficients (Tibshirani, 1996).[12] In this case, some coefficients are set exactly to zero. It is thus a sparse model which performs shrinkage through variable selection. The optimisation problem is

$$\hat{\beta}^{Lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{t}(\epsilon_{t+h} - \mathbf{x_t}\beta)^2 + \lambda \sum_{i}^{N} |\beta_i| \right\} \tag{6}$$

Finally, the Elastic Net is a hybrid approach which combines the previous L1 and L2 penalties (Zou & Hastie, 2005). In the first step, it finds Ridge regression coefficients, and in the second step, Lasso-type shrinkage, i.e., variable selection is applied. The "naïve" estimators of the Elastic Net, β^{n-EN} are computed by solving the problem:

$$\hat{\beta}^{n-EN} = \min_{\beta}\{\sum_{t}(\epsilon_{t+h}-\mathbf{x_t}\beta)^2+\lambda_1 \sum_{i}^{N} \beta_i^2+\lambda_2 \sum_{i}^{N} |\beta_i|\} \tag{7}$$

The correction factor $1+\lambda_2$ is applied and the estimators are given by $\hat{\beta}^{EN} = (1+\lambda_2)\hat{\beta}^{n-EN}$ to account for increased bias through double shrinkage.

Non-linear machine learning models

Whereas in the models presented so far, the function g that links predictors and the target variable was linear, machine learning models allow for a more complex non-linear function $g(\cdot)$. We use three types of machine learning models: Random Forests, Artificial Neural Networks, and Support Vector Machines.

A Random Forest is an ensemble of uncorrelated trees that are estimated separately (Breiman, 2001). The trees which are consecutively split the training data set into buckets (leafs) K_{leaf}, $Z = \{Z_1, \ldots Z_{N_{leaf}}\}$, until an assignment criterion is reached. The optimal estimate of the β is the average of the training target values within each tree leaf. The algorithm minimises the objective function within areas of the target space, i.e. these "buckets", conditioned on the input $\mathbf{x_t}$.

$$\epsilon^{RF}_{t+h} = \sum_{k=1}^{K_{leaf}} \hat{\beta}_k I(\mathbf{x_t} \in Z_k) + u^{RF}_{t+h}, \quad \text{with}$$
$$\hat{\beta}_k = 1/|Z_k| \sum_{\epsilon^{tr} \in Z_k} \epsilon^{tr}_{t+h},\ k \in \{1, \ldots, K_{leaf}\}\,. \tag{8}$$

The predictions of the individual trees are then averaged for a single prediction, reducing variance (bagging). The correlation between trees in a forest is reduced by building them from small enough random samples drawn with replacement from the training sample, alleviating overfitting. Tree models are mostly sparse as their hierarchical structure acts like a filter. That is, only variables which improve the fit are chosen during the construction of each tree during training.

Artificial Neural Networks (ANN) consist of an input layer, at least one hidden layer, and an output layer. An ANN becomes a linear function without a hidden layer, similar to solving the least squares problem. As part of our ANN architecture, we use multi-layer perceptron (MLP), a feed-forward network. The variables $\mathbf{x_t}$ in the input layer are multiplied by weight matrices W at each hidden layer,

[11] 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.

[12] The L1-Lasso-penalty makes the solutions non-linear 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 λ varies. For example, Least Angle Regression (LARs, Efron, Hastie, Johnstone, and Tibshirani (2004)) provides an efficient algorithm for computing the Lasso estimates.

where all internal hidden layers have size K_{hl}. The product of inputs from the previous layer is transformed by an activation function and passed on through the network until the linear output layer is reached, resulting in the final regression coefficients $\hat{\beta}$ and a prediction $\hat{\epsilon}_{t+h}$. Formally, this can be described as

$$\epsilon^{NN}_{t+h} = \sum_{k=0}^{K_{hl}} \hat{\beta}_k \tilde{g}_L(\tilde{g}_{L-1}(\tilde{g}_{L-2}(\dots(\tilde{g}_1(\mathbf{x_t}, W_1), \dots), W_{L-2}), \beta_{L-1}), W_L)_k + u^{NN}_{t+h} \quad (9)$$

As activation function $\tilde{g}(\cdot)$, we use the rectified unit function (ReLU); it acts as a gate for signals and introduces non-linearity into the model. The number of hidden layers, i.e., the depth of the network, the number of neurons in each layer and the weight penalisation are hyperparameters determined by cross-validation. 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.

Support Vector Machines (SVM) identify a set of training points, the support vectors, to either represent a boundary between classes (classification) or a line (regression) (Vapnik, 1998). It offers nice statistical properties and can capture non-linearities in the data through the use of kernels for the joint processing of test observations in conjunction with the support vectors (Wang, Wang, & Zhang, 2012; Xiang-rong, Long-ying, & Zhi-sheng, 2010). A support vector regression with a continuous target can be written as

$$\epsilon^{SVM}_{t+h} = \sum_{k=1}^{K} \beta_k \mathcal{K}\left(x^{tr}_k, \mathbf{x_t}\right) + u^{SVM}_{t+h}, \quad (10)$$

where the sum runs over the training sample. The K weights $\beta_k \geq 0$ mark the support vectors x^{tr}_k, jointly selected from the training data during optimisation. The kernel $\mathcal{K}(\cdot,\cdot)$ acts like an inner product and returns a scalar; we use a Gaussian kernel (radial basis function, RBF). Penalisation is achieved by imposing restrictions on β_k.

3.4. Tuning of hyperparameters

All of our models require some form of hyperparameter selection. For the shrinkage methods and machine learning tools, we use cross-validation procedures that depend on out-of-sample performance, differently from information criteria, which are "in-sample" statistics (Goulet Coulombe, Leroux, Stevanovic, and Surprenant, 2022). K-fold cross-validation assumes that samples are independent and identically distributed, resulting in an unreasonable correlation between training and testing instances in the time series context. Therefore, we opt for a variant of K-fold cross-validation where the model is evaluated on "future" observations least like those used to train the model. 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 the test set. We consider the average mean squared error over the test set a performance metric.

The hyper-parameters selected through cross-validation include the penalty imposed on shrinkage methods, the number of components for dimensionality reduction techniques, the maximum depth of trees for the Random Forest, the architecture of the ANN and the choice of the kernel function for SVM, as well as penalisation weights for the latter two models.[13] Given that the estimation is done over a rolling window, the selected hyperparameters can change over windows, as well as over the forecast horizon and specification. Overall, cross-validation favours similar parameters and model architectures across specifications that use different sets of predictors and horizons. However, they evolve somewhat over time, i.e. over rolling window test sets. 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 time series, model complexity increases, and the procedure selects deeper versions of the network and larger tree structures of the Random Forest.

4. Results

We present the results of the forecasting exercise focusing on relative RMSE against the $AR(2)$ benchmark and predicted value comparisons for the models with CPI items, either alone or in combination with macroeconomic series. We start with results over the entire sample period to gauge average performance. Then, we evaluate forecast performance over sub-periods with specific inflation characteristics, i.e., periods of rising and falling inflation or high and low inflation. The latter assesses whether disaggregated data and machine learning tools might be particularly useful when inflation experiences turning points or tail outcomes.

4.1. Forecast results over the entire sample period

Table 2 shows the absolute RMSE of the AR(2) for forecasts of the three different inflation measures for selected forecast horizons and the relative RMSE of other models, including different sets of predictors in the three panels, respectively. Overall, the AR(2) benchmark performs relatively well for all targets. The absolute RMSE for 3-month forecasts lies at around 0.6 for all targets; this corresponds to roughly half a standard deviation of headline inflation over the sample period and a full standard deviation for core and service inflation. Thus, forecasts are

[13] We choose the following grid sets: for Ridge, LASSO, Elastic Net, and ANN penalisation weights are $\lambda \in \{1e-05, 0.0001, 0.001, 0.01, 0.1, 1.0\}$, for Elastic Net L1-ratio $\in \{0.1, .5, .9, .95, 1\}$, for Forest max. depth $\in \{1, 2, \dots, 9, 10\}$, for ANN hidden layer dimension $\in \{(10, 2), (20, 2), (2, 3), (20, 3), (5, 5), (50, 50), (10, 10, 10)\}$ and activation function tanh or ReLU, for SVM $C \in \{10, 100, 1000\}$ and $\epsilon \in \{0.01, 0.1, 0.5, 0.9\}$, the kernel is chosen to be RBF. The range of the number of clusters for PLS is $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. It was set to $K = 6$ for the PCA based on the explained variance of about 2/3. Cross-validation suggested that this higher number leads to heavy over-fitting and under-performance, similar to factor models. The results of the cross-validation exercise on all chosen hyperparameters are available upon request.

Table 2
Forecasting exercise results.

	Target: headline CPI			Target: Core CPI			Target: Service CPI		
horizon	3	6	12	3	6	12	3	6	12
AR(2), absol. RMSE	*0.57*	*0.88*	*1.17*	*0.61*	*0.85*	*.98*	*0.59*	*0.85*	*1.19*
Models with CPI items (relative RMSE to AR(2))									
PCA	1.00	0.98	1.04	0.96	0.97	1.00	1.03	1.06	1.08
PLS	1.04	1.17	1.29*	1.11	1.05	1.09*	0.97	1.14*	1.10
Ridge	1.00	**0.84***	0.96	1.05	0.89	1.02	1.01	1.01	1.01
Lasso	0.98	1.01	0.96	**0.97***	1.05	0.98	1.04	1.09**	1.06
Elastic	1.00	1.00	0.98	0.99	0.99	1.00	0.99	1.01	1.03
SVM	1.00	1.02	1.14	1.02	1.03	1.08	1.04	1.05	1.03
Forest	0.99	1.07	1.21	1.01	1.09	1.19	1.06*	1.11**	1.14**
ANN	1.19**	1.24*	1.16	1.12	1.35	1.11	1.4	1.26	1.32**
Models with CPI items and macroeconomic indicators (relative RMSE to AR(2))									
PCA	1.01	1.00	0.97	0.95	0.96	1.01	0.96	1.06	1.07
PLS	1.20*	1.09	1.02	1.03	1.05	1.13**	0.98	1.15	1.15**
Ridge	0.90	0.99	0.97	1.05	0.91	0.97	0.95	1.02	1.01
Lasso	1.05	0.98	1.0	1.02	1.01	0.95	1.02	1.07	1.06
Elastic	0.95	1.00	1.00	0.98	0.98	0.97	0.99	1.00	1.03
SVM	1.01	0.99	1.01	1.01	1.02	1.08	1.05	1.05	1.03
Forest	1.07	0.96	0.99	1.01	1.10	1.15	1.06	1.09*	1.15**
ANN	1.25*	1.28**	1.41*	1.21**	1.19*	1.45***	1.31**	1.46**	1.14
Models with macroeconomic indicators (relative RMSE to AR(2))									
PCA	1.01	1.81	1.62	1.03	1.01	1.09***	1.17	1.42*	1.68
PLS	1.22**	1.72**	1.72***	1.13*	1.43*	1.31***	1.05	1.62*	1.58**
Ridge	1.18	1.52**	1.66***	1.16	1.15	1.49	0.98	1.26	1.44
Lasso	0.97	**0.99***	1.20	0.98	0.99	1.03	1.00	1.07	1.27
Elastic	0.98	1.02	0.99	0.98	0.99	1.01	0.95	0.99	1.02
SVM	0.99	1.10	1.26**	1.01	1.08	1.12*	1.08	1.11	1.12*
Forest	0.99	1.18	1.25	0.96	1.05	1.12	1.09*	1.22*	1.21**
ANN	1.16*	1.44**	1.53***	1.21*	1.22***	1.33	1.12	1.25**	1.42***

Notes: Root mean squared errors, absolute for AR(2) benchmark without additional predictors, and relative to benchmark for all other models. Forecasts of headline CPI inflation (left panel), Core inflation (middle) and Service inflation (right). Rolling window samples over 2002–2021 with seven years of training sample and out-of-sample forecasts at horizons of three, six, and 12 months. The significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. Significance levels: ***:1%, **:5%, *:10%. Relative RMSE for forecasts at the 1-month horizon were insignificant and not presented for space constraints.

relatively more precise for headline inflation, given the overall higher variation in the series. For the other models, it is difficult to significantly outperform the benchmark, with shrinkage methods performing better than machine learning tools or dimension reduction techniques. When including CPI items as predictors (upper panel), Ridge regression provides a substantial improvement by 16% against the AR at the 6-month horizon for headline, and LASSO a significant improvement at the 3-month horizon for core inflation. None of the models outperform the benchmark for core service inflation, which is a rather stable component.

The additional gain from macroeconomic indicators (middle and lower) is small, and the results are similar or worse than those of the models with CPI items only, including for the Ridge model. Thus, adding more predictors does not seem helpful and instead results in losing degrees of freedom.

Our result that the autoregressive model performs relatively well for forecasting inflation and that few model forecasts can outperform it on average over longer sample periods is in line with previous literature, see for instance (Domit et al., 2019; Faust & Wright, 2013).[14] Garcia et al. (2017) observe, for a comparable set of models and also using a large data set to forecast inflation in Brazil, that forecast performance varies over time and that the models that perform best on average are not among the ones that come first most frequently across a range of rolling window sub-periods and horizons.

Fig. 2 suggests model performance might vary over time; in our case, it shows the headline inflation predictions from models that use CPI item series predictors compared to the actual outturns (lagged by the number of months equal to the forecast horizon) and the AR(2) benchmark. Most model forecasts are quite close to the AR(2), but most have some excess volatility, explaining their weaker performance. For certain episodes throughout the evaluation sample, however, some models perform better than others. Ridge regression, for instance, appears to perform well at the 6-month horizon (middle column, middle panel) during the inflation rises in 2012 and 2017 or during the fall in 2014–2015. It also captures the rise of inflation at the end of the sample earlier than other models, albeit still with a sizeable forecast error.

Hence, whereas, on average, a simple AR(2) is hard to beat over the evaluation sample, this might vary over

[14] Faust and Wright (2013) in a comprehensive forecast comparison for US inflation find that no model forecast significantly improves upon an AR(1) forecast for CPI inflation, only institutional forecasts outperform it at very short horizons. They find other models, such as a PCA, which we also employ to perform up to 20% worse than the AR(1). For the UK, Domit et al. (2019) find that a BVAR model does not outperform the Bank of England forecasts.

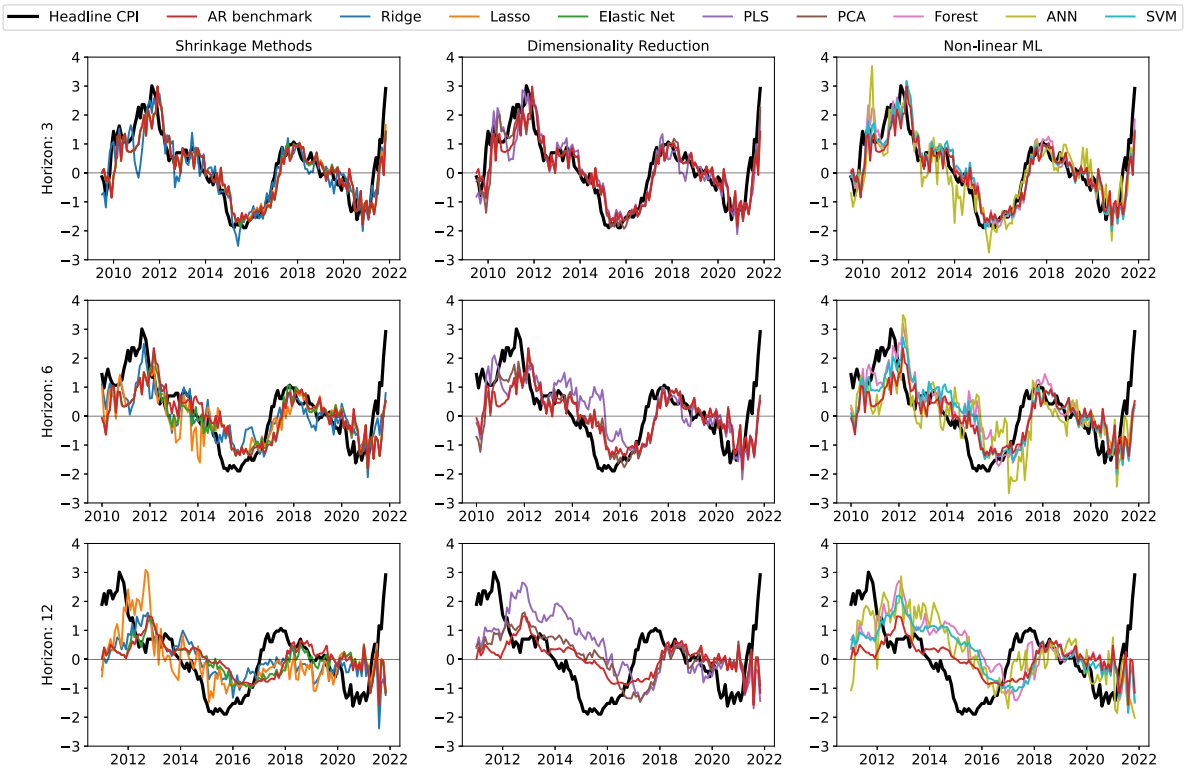

Fig. 2. Predicted values for headline CPI inflation, CPI items predictors.
Notes: Forecasts of CPI headline inflation (standardised) from different types of forecasting models (columns, coloured lines) using CPI items as predictors, over horizons $h = 3, 6, 12$ (rows). Out-of-sample predictions for rolling samples from 2009m2 to 2021m11, compared to the actual headline CPI inflation outcome (black lines) lagged by h months.

time and change over certain sub-periods. During periods where non-linearities or shifts in the inflationary process are occurring, non-linear models might have advantages and exploiting large data combined with effective shrinkage might be important to draw on a wide range of signals to learn about non-linearities (Goulet Coulombe, 2022; Hauzenberger et al., 2023). We evaluate this in the following sub-section.

4.2. Forecast evaluation over specific inflation outcomes

In the following, we evaluate the forecasting exercise differently over sub-periods where aggregate inflation, either predicted or realised, meets specific characteristics, e.g., when inflation has a certain momentum (rising/falling) or is in its tails (low/high). We adjust the evaluation exercise such that predictions are first assigned to a regime based on current knowledge. Then, the models that exploit large data are used during those regimes and evaluated against the AR benchmark. The regimes are defined ad-hoc based on similar definitions in the literature. They aim to assess statistical differences in the forecast performance during sub-periods that might reflect certain underlying regime shifts. However, regimes are not identified structurally given the non-structural nature of the forecasting exercise, and since identification is naturally challenging and subject to uncertainty over the relatively short sample.

4.2.1. Set up of forecast evaluation over sub-periods

At each point in time t of the evaluation sample, we first check whether a criterion is met according to a definition of regimes outlined below. If yes, we use the forecast of model m at horizon h at time t. If not, we use the AR(2) forecast. We then evaluate these regime-based forecasts where we use model m during sub-periods and the AR(2) otherwise, against the case of using the AR(2) model throughout.

With regard to the inflation series against which we evaluate a regime at time t, i.e. a reference value r_t, we pursue two options

- the *model forecast* of headline inflation (alternatively core or service inflation) from model m, at time t and horizon h, i.e. $r_t = y^m_{t+h}$, is evaluated against the regime characteristic
- the *outturn* for headline (or core or services) inflation at time t, i.e. $r_t = y_t$

Relying on the model forecast reflects that the practitioner uses the models to learn about regimes. While this option is more forward-looking about inflation over the forecast horizon, it implies that regimes might vary across models. Instead, relying on the outturn, which we explore for robustness and show in the online appendix, defines the regimes uniformly across models but is more backwards-looking since the regime is defined based on

inflation at the time of the forecast. Both approaches respect the information available at time t, while the choice of approach will depend on the goal of the modeller.

With regard to the characteristics that pin down regimes, our particular interest lies in periods where inflation dynamics might be unstable and subject to non-linearities or shifts. This might be around turning points where dynamics or the "momentum" in inflation are changing, or periods when inflation is particularly low or high (Forbes, Gagnon, & Collins, 2021), or at the tails of the distribution (Lopez-Salido & Loria, 2022). These are also periods of particular interest for policymakers who might need to adjust policy in response to changing inflation dynamics (Mann, 2021). Inflation dynamics might operate differently, e.g. agents might pay more attention to inflation when it is high. Pricing behaviour might change, and non-linear demand functions or kinked supply profiles might induce non-linearities and shifts in the Phillips curve (Harding, Lindé, & Trabandt, 2023). And spillovers across price categories might increase in periods of high inflation, such that disaggregated data might gain predictive power (Borio, Hofmann, & Zakrajšek, 2023). We approximate such regimes as periods where inflation outcomes are in the outer quartiles of their distribution over time or when inflation rates rise or fall consistently over subsequent periods. Regimes are defined, with respect to the reference value r_t, i.e. the current outturn $r_t = y_t$ or predicted value $y^m_{t_h}$ of aggregate (headline, core, or service) inflation, as follows:

1. *high*: r_t is above the 75th percentile of the distribution of aggregate inflation over time up until point t, $\mathcal{D}_t(y)$.[15]
2. *low*: r_t is below the 25th percentile of $\mathcal{D}_t(y)$.
3. r_t lies within its *interquartile range (IQR)* of $\mathcal{D}_t(y)$.
4. *rising*: the *the change* in r_t being positive for at least three consecutive months.[16]
5. *falling*: the *the change* in r_t being negative for at least three consecutive months.

All in all, the definition of regimes requires making ad-hoc choices about their nature and boundaries. Any such choice is somewhat arbitrary and naturally faces limitations since it does not account for the potential endogeneity of such regimes. This is reminiscent of the modelling choices and additional judgments practitioners must make. Using simple benchmark models in quieter times but looking at more complex models that exploit disaggregated data in periods that meet certain characteristics represents a tractable and easily applicable decision rule for practitioners. An alternative selection approach to pin down regimes could be to define a grid of options, e.g., different cut-off points for high/low inflation regimes, and then pin down the preferred regimes via cross-validation. This would determine regimes based on the strongest relative improvement of a model against the AR(2) during the regime. Another alternative could be running threshold models that determine regimes endogenously; this is challenging and subject to model and sampling uncertainty since the variation related to the regime has to be separated from other non-linear parts of the model. Given our data's relatively short time series dimension, which gives little information over time to help identify regimes, we leave this to future research.

4.2.2. Results over sub-periods

Table 3 shows the results for headline inflation forecasts over different horizons (columns) and using different sets of predictors (panels) for the case of using the respective model forecast as a basis to define regimes. For readability, the table only shows cases where improvements in a regime against the benchmark were statistically significant. Dimension reduction techniques are shown in blue, shrinkage methods in red, and machine learning tools in green. Overall results are similar when using the actual inflation outturns at time t to define regimes (see Tables A2 and A3 in the online appendix). We find a wide range of significant improvements against the AR(2) from using all three types of shrinkage models (Ridge, Lasso, Elastic Net) and a few improvements with principal component analysis (PCA) during the regime sub-periods. Similarly to the results evaluated over the entire sample period, we find little gain from adding macroeconomic indicators as predictors in addition to disaggregated CPI items.

Looking closer at the results over sub-periods, shrinkage methods reach relative forecast gains across horizons when the headline inflation forecast (or outturn) is falling and when inflation is high at the 6-month or 12-month horizons. Ridge regression continues to perform strongly across regimes, reaching improvements against the AR of up to 15%–25%; in various cases, it is outperformed by Lasso or the Elastic Net, particularly at the 12-month horizon.

The fact that there is no clear model winner and different shrinkage methods achieve gains over sub-periods and horizons is in line with findings by Medeiros et al. (2021) and suggests that a model combination or selection across many specification runs might be a robust choice to reduce model uncertainty. Fewer and, in tendency, smaller improvements are achieved for periods when inflation is rising, low, and in the IQR range. However, improvements are still substantial in a few cases, such as from the PCA that improves by 25% six months ahead against the AR(2) when inflation is rising. Inflation surges, such as the one observed at the end of the sample period, can be particularly difficult to predict, but exploiting disaggregated data can be helpful. Once inflation is high or falling, disaggregated data might carry signals about spillovers between item groups and the broad-based degree of inflation. On the other hand, when inflation is low and in the IQR range, the AR(2) model performs particularly well, and information from item series is likely less useful since spillovers between sectors tend to be smaller when inflation is low.

[15] We exploit the longest available time series to represent the distribution over time, starting in January 1997 for all three inflation measures. The definition captures both short-term stints of high inflation as well as more persistent high inflation periods.

[16] To smooth through volatile changes, we compute a one-sided three-month moving average of the change in r_t. Restricting the number of consecutive months to a higher value would capture more persistent momentum, but it would also reduce the number of episodes for which the criterion is met.

Table 3
Regime-dependent headline inflation forecasts (regimes based on forecasts).

Horizon	3		6		12	
Headline CPI inflation forecast — CPI items only						
Headline inflation forecast falling	Ridge	0.76**	Ridge	0.85**	Lasso	0.90**
	Lasso	0.99*			Elastic	0.97**
Headline inflation forecast rising	–		PCA	0.75***	Elastic	0.96*
Headline inflation forecast high	–		Elastic	0.98**	Elastic	0.81***
			Lasso	0.94***		
			Ridge	0.78***		
			PCA	0.75***		
Headline inflation forecast low	–		Ridge	0.80*	Lasso	0.87*
	–		–		Elastic	0.95***
Headline inflation forecast in IQR	–		–		Ridge	0.86***
Headline CPI forecast - CPI items & Macro indicators						
Headline inflation falling	Ridge	0.74**	Ridge	0.90*	Lasso	0.83***
	PLS	0.87**			Elastic	0.96**
Headline inflation forecast rising	–		Elastic	0.86***		
	–		PCA	0.77***		
Headline inflation high	–		Elastic	0.84***	Elastic	0.82***
			Ridge	0.86**		
			Lasso	0.98*		
			PCA	0.77***		
Headline inflation forecast low	–		–		Elastic	0.95**
					Lasso	0.84***
Headline inflation forecast in IQR	–		–		Ridge	0.86***
Headline CPI forecast - Macro indicators only						
Headline inflation forecast falling	Elastic	0.99**	Elastic	0.97**	–	
Headline inflation forecast low	–		Elastic	0.97*	–	

Notes: The different models are used over sub-periods during which the headline CPI inflation forecast serves as a reference value, i.e. the predicted value for a given model at a given horizon at the current point is falling/rising or high/low/IQR. Relative RMSE against the AR(2) benchmark are computed over respective regimes. A minimum of six months of improved forecast performance is required across the evaluation period. Only cases with significant relative forecast gains are listed (sub-periods where no models showed significant gains for a specification are not listed). Significance is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. Significance levels: ***:1%, **:5%, *:10%.

Even though we are focusing on periods where inflation dynamics might have been unstable, there are no occasions where machine learning tools beat the AR(2) benchmarks. The evidence supports using linear methods for forecasting headline UK inflation with disaggregated data. In Section 5, we will look into this result further by decomposing the forecast signals over time from a linear model, Ridge regression, compared to the Random Forest. These results indicate that while the Random Forest can flexibly move between signals from groups of indicators during periods when inflation dynamics are changing, the excess noise in signal extraction might be detrimental to forecast performance.

Table 4 shows the corresponding results over sub-periods for the core and service CPI inflation specifications. Results are overall comparable, again with the most significant improvements achieved by the different shrinkage methods and, in some cases, PCA. Now, there are also a few improvements from machine learning tools, mainly Support Vector Machines, but they are relatively weaker compared to what is achieved by linear models. For core inflation, improvements against the AR(2) are achieved across all the horizons, mainly when core inflation is falling, rising or high. However, improvements tend to be smaller than we have seen for headline inflation. Significant forecast gains are mainly achieved for service inflation for the 12-month horizon. Unlike the other inflation measures, the improvements occur mainly when service inflation is falling or low, although the PCA also achieves substantial gains when service inflation rises.

Overall, the forecast evaluation results over regimes confirm the finding seen for the overall sample: linear methods combined with disaggregated CPI items achieve improvements against an AR(2). During specific sub-periods, improvements can be more sizeable and observed for a wider range of models than the overall sample average. However, which model beats the AR(2) varies with the regime definition and, to a lesser extent, also with the horizon, such that model averaging across various shrinkage models might be preferred to mitigate model uncertainty. For headline and core inflation, exploiting large data improves performance mainly during periods when inflation is high or falling, possibly because spillovers between groups of item series gain relevance when inflation is high. Machine learning tools can rarely outperform the AR(2), even during periods when inflation dynamics are changing or at their tails; in the next section, we add some intuition to this result.

5. Forecast contributions over time-based on the shapley value framework

To interpret results, it is useful to understand which signals the models exploit over time and whether the contribution of sub-groups of CPI item series (e.g., core items, energy items, or goods vs. service items) are more relevant than others and when. This can be informative

Table 4
Regime-dependent core and service infl. forecasts (regimes based on forecasts).

Horizon	3		6		12	
Core CPI inflation forecast — CPI items only						
Core inflation forecast falling	–		SVM	0.99*	–	
Core infl. forecast rising	Elastic	0.97**	PCA	0.91***	Elastic	0.90***
	Lasso	0.96**	PCA	0.96*		
Core infl. forecast high	Elastic	0.98**	PCA	0.90***	Elastic	0.92**
	Lasso	0.97**				
	PCA	0.91**				
Core CPI forecast — CPI items & Macro indicators						
Core infl. forecast falling	–		SVM	0.97***	SVM	0.96***
					Ridge	0.85**
					Elastic	0.96*
Core infl. forecast rising	Elastic	0.97**	PCA	0.89***	Elastic	0.92***
					Lasso	0.90*
Core infl. forecast high	Elastic	0.97***	PCA	0.87***	Elastic	0.91**
	PCA	0.92**				
Core infl. forecast in IQR	–		–		Elastic	0.95***
Core CPI forecast - Macro indicators only						
Core infl. forecast falling	Lasso	0.98*	SVM	0.94**	–	
			PLS	0.92*		
			PCA	0.94**		
Core inflation rising	Elastic	0.99*	–		–	
Core infl. forecast in IQR	–		PCA	0.93***	–	
Service CPI forecast — CPI items only						
Service infl. forecast falling	Elastic	0.99*	–		Elastic	0.98**
					Ridge	0.92**
					Lasso	0.91*
Service infl. forecast rising	PCA	0.91**	PCA	0.79***	Elastic	0.94***
Service infl. forecast low	Elastic	0.99*	–		Elastic	0.97*
					Ridge	0.89***
					Lasso	0.86***
					NN	0.85***
Service infl. forecast in IQR	–		SVM	0.96**	Elastic	0.96***
Service CPI forecast - CPI items & Macro indicators						
Service infl. forecast falling	Ridge	0.78*	–		Ridge	0.92**
					Elastic	0.98*
Service infl. forecast rising	PCA	0.87***	PCA	0.77***	–	
	Elastic	0.97***	–		–	
Service infl. forecast low	–		–		Ridge	0.89***
					Lasso	0.90**
					NN	0.87**
Service infl. forecast in IQR	–		RF	0.92***	Elastic	0.97**
			SVM	0.96**		
Service CPI forecast - Macro indicators						
Service infl. forecast rising	–		Ridge	0.78*	–	
			Elastic	0.95***		
Service infl. forecast in IQR	Elastic	0.97**	–		–	

Notes: See notes to Table 3. Regimes are defined based on core (service) inflation forecasts.

for identifying relevant predictors for lower-dimensional forecasting frameworks or for communicating forecasts and informing policy decisions. It can help understand the results of the forecasting exercise and, in particular, why the performance of machine learning models is relatively weak, even when inflation dynamics might be unstable. This is not a straightforward task with non-parametric and non-linear models and with a large number of predictors. We assess which item groups provide forecast signals over time using a linearised model decomposition via Shapley values. We also examine CPI item groups' absolute and relative Shapley weights over time. We do this for the Random forest, a machine-learning model for which Shapley values can be derived straightforwardly, compared with the Ridge regression, the best-performing linear model in our forecast exercise. Shapley values will give us a linear representation of the forecast signals that are picked up by a model at different points in time, which helps communicate forecasts. It should be noted that Shapley decompositions do not have a causal or structural interpretation, which is out of the scope of any forecasting exercise.

5.1. Shapley values to explain statistical models

The first step is **model decomposition**. We employ the *Shapley additive explanations* framework (Lundberg & Lee, 2017; Strumbelj & Kononenko, 2010), which exploits an analogy between variables in a model and players in

a cooperative game and has a set of appealing analytical properties.[17] It consists of 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, or *Shapley values*, from the individual input variables.

Let the total number of predictors be N from price items and/or macroeconomic series. A predicted value $g^m_{t+h}(x_t)$, from our forecast exercise where model m predicts the residual part of inflation not explained by AR(2) at horizon h, can be decomposed into its Shapley components $\phi^{m,h}_{t,i}$ attributed to the forecast contribution from the ith variable at time t over horizon h.[18] That is,

$$g^m_{t+h}(x_t) = \sum_{i=1}^{N} \phi^{m,h}_{t,i} \,. \qquad \text{(decomposition)} \tag{11}$$

For a non-linear forecasting model, computation of (11) requires deriving the marginal contribution of predictor i by running sequential forecasts of all possible coalitions of predictors, with and without i. Thus, the Shapley value for predictor i (ignoring time subscript and forecast horizon superscript for the moment) is computed as

$$\phi_i = \sum_{\mathcal{S}\subseteq\mathcal{N}\setminus i} \frac{S!(N-S-1)!}{N!}[f(\mathcal{S}\cup\{i\}) - f(\mathcal{S})] \,. \tag{12}$$

Here, the payoff of a coalition $\mathcal{S}\subseteq\mathcal{N}$ is $f(\mathcal{S})$, the payoff of this coalition combined with predictor i is $f(\mathcal{S}\cup\{i\})$, and their difference measures the marginal contribution of i to that coalition. The intercept ϕ_0 corresponds to $f(\varnothing)$, i.e. with no variables in the model.[19] After summing these marginal contributions over all coalitions, we get an estimate of the contribution of variable i to a single model prediction. We therefore focus our analysis on a non-linear model where an exact solution exists, the Random Forest, and a linear model, the Ridge regression. For a linear model, the Shapley value of predictor i is simply the product of its regression coefficient w_i and the difference between the predictor value and its mean, i.e. $\phi_i = w_i(x_i - \mathbb{E}_t[x_i])$ with the expectation taken over the training data set. 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. (12) (see Lundberg, Erion, and Lee (2018) for details). For other models, coalitions could be sampled with a readjustment of the weights in (12), but comparing all possible combinations of predictors with $N \approx 400$ is computationally infeasible.[20]

Next, we **group and re-aggregate** the N Shapley values ϕ^h_{ti} into $K \ll N$ higher-level sub-groups denoted $\mathcal{G}_k$, which removes volatility in individual contributions and facilitates communication of results.

$$g^m_{t+h} = \sum_{k=0}^{K} \psi^{m,h}_{t,k} \quad \text{with}$$
$$\psi^{m,h}_{t,k} = \sum_{i\in\mathcal{G}_k} \phi^{m,h}_{t,i} \qquad \text{(sub-group aggregation)} \,, \tag{13}$$

with $\psi^h_{t0} = \phi^h_{t0}$ being the same intercept. We group CPI items into six groups corresponding to different CPI categories: core goods, food & beverages, energy, core services, volatile services, and macroeconomic indicators. This grouping reflects broad inflation categories that central banks typically monitor since they reflect groups that are more or less exposed to terms of trade shocks and volatile fluctuations and also intend to capture potential sector-level heterogeneity that might, for instance, lead to service items behaving differently from goods items (Delle Chiaie, Ferrara, & Giannone, 2022; Ha, Kose, & Ohnsorge, 2023).[21]

We then derive simple metrics for the relative contributions of sub-groups to the forecast based on **Shapley weights** of a group for a forecast at a particular horizon,

$$\bar{\Phi}^{m,h}_{t,k} = \frac{|\Phi|^{m,h}_{t,k}}{\sum_{k=1}^{K}|\Phi|^{m,h}_{t,k}} \quad \text{with} \quad |\Phi|^{m,h}_{t,k} = \sum_{i\in\mathcal{G}_k}|\phi^{m,h}_{t,i}|$$
(Shapley weight) (14)

$$\tilde{\Phi}^h_k = \frac{\bar{\Phi}^h_k}{N_k/N} \,.$$
(Shapley weight normalised by relative group size) (15)

Shapley weights are the shares of the absolute Shapley values of group k relative to the contributions from all groups. They sum to one over all K groups. The sub-groups of items that we consider are not of equal size, given that, for instance, there are more food items in the consumption basket than energy items. We would also expect larger item groups to carry a larger Shapley weight contribution. At the same time, item groups might be relatively more informative for a model at a given time if the variation in these items carries important signals that the model exploits for aggregate inflation despite a small group size.[22] To reflect this, we also compute normalised weights that divide the Shapley weights by the relative number of items within a group entering the model at a given time, with N_k being the number of items

[17] In particular, it is the only attribution framework that is local, linear, efficient, symmetric and respects null contributions and strong monotonicity of variables (see Lundberg and Lee (2017), Young (1985) for details).

[18] The AR forecast component can be added as an additional summand if one wishes to recover the combined model forecast.

[19] The $i = 0$ component is set to the mean predicted value in the training set and can be interpreted as an intercept.

[20] See Buckmann, Joseph, and Robertson (2022) for details and a discussion of different computational approaches.

[21] We tried an alternative grouping where we instead group items according to their persistence to see whether models pick up signals from slow-moving or volatile components more explicitly. Results confirmed the conclusions from the grouping by CPI categories since goods, energy, and food items tended to be categorised as having low or medium persistence. In contrast, service items mostly formed the highly persistent sub-group.

[22] Energy and volatile services comprise about 1.6% of items each, albeit energy with an over-proportional weight in the consumption basket. Food & beverages is a major sub-group, with 24% of items, but with an under-proportional consumption weight.

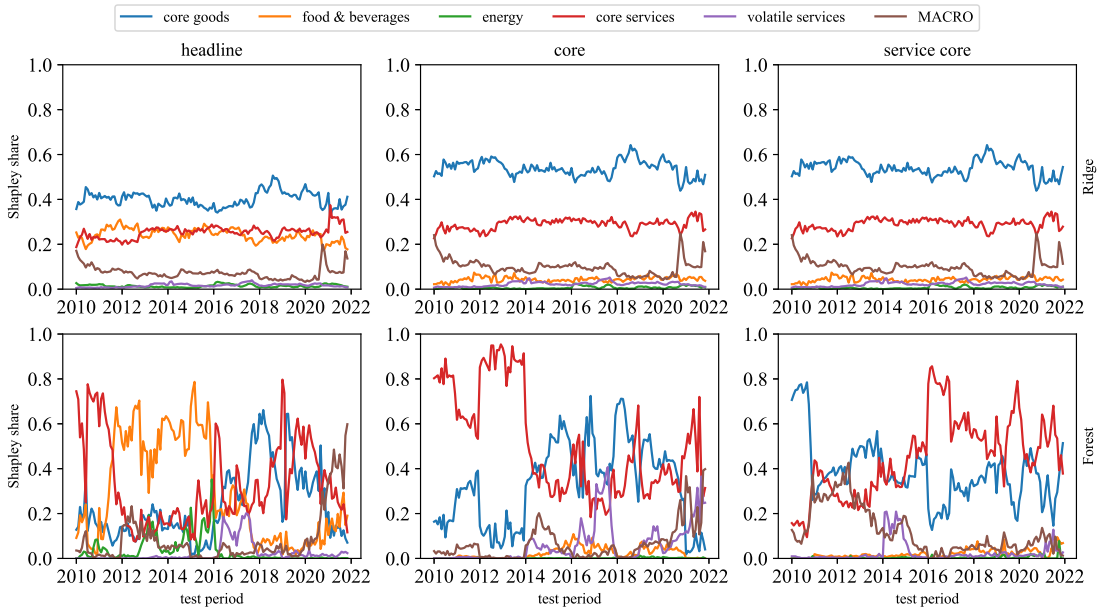

Fig. 3. Shapley weights over time, for forecasts at the 6-month horizon.
Notes: Shapley value contributions from a given group of predictors, relative to the total sum of Shapley value contributions, for forecasts at horizon $h = 6$ for different targets (columns), from Ridge regression (upper row) and the Random Forest (lower), over the evaluation period.

in group k. Normalised weights will be around one if a model relies about uniformly on items in each sub-group compared in terms of Shapley attributions but can take up lower or higher values if Shapley attributions for a group are under- or over-proportional to the number of items in a group. The normalised Shapley weights will show larger values for those groups that the model identifies as the relatively more relevant predictors at a given horizon based on what the model had learned until that point in time. Therefore, they capture the relative predictive ability of item groups according to the model, even if the actual predictive performance might not necessarily improve ex-post, such as during the recent inflation surge, which all models missed.

5.2. Shapley weights over time

Fig. 3 shows the Shapley weights over time for the two models (rows) and targets (columns) for forecasts at the 6-month horizon. This shows how the absolute contribution of groups of predictors (not normalised by group size) varies over time in the linear Ridge regression (upper row) compared to the Random Forest. We focus on the specifications, including CPI item series and macroeconomic variables. Figure A3 in the appendix shows Shapley weights over all horizons, averaged over time.

Core goods items—the largest item group in terms of number and weight in the CPI basket—are attributed the largest shares across models, targets and horizons in both models, although more so for the Ridge regression. The Random Forest generally assigns much more volatile weights and, apart from core goods, also reads strong signals from core service items. For headline inflation, food & beverage items also play a sizeable role for both models, although the weight dropped for the Random Forest in 2016 and only recovered slightly at the end of the sample. Interestingly, energy plays a small role for both models in line with its relatively small weight in the CPI basket. Apart from small fluctuations in the short run, the weights from Ridge regression are stable over time (and over horizons), in line with this being a linear model. On the other hand, Shapley weights for the Random Forest are very volatile and more difficult to interpret, with occasional strong short-lived drops, and the model at times switched between signals from different groups of indicators, such as between core goods and core services.

Next, we assess whether the models put over-proportional weights on certain groups of predictors in certain periods when the model ex-ante assigned a relatively stronger predictive ability to certain groups based on what it had learnt until then. Fig. 4 shows the normalised Shapley weights that penalise the number of predictors in a group. For the Ridge regression, those mostly fluctuate around one, suggesting that weights assigned are roughly proportional. The Random Forest very strongly over-weights certain components at times, suggesting that these components can have non-linear linkages with inflation that, within the model, raises their relative predictive power during certain periods characterised by larger shocks: energy items provide over-proportional signals for headline inflation between 2013 and 2016 when energy prices declined, volatile service items offer over-proportional signals to headline and core inflation forecasts between 2016 and 2018, in the aftermath of the Brexit referendum, and for core inflation again since 2020 when the Covid-19 pandemic led to a

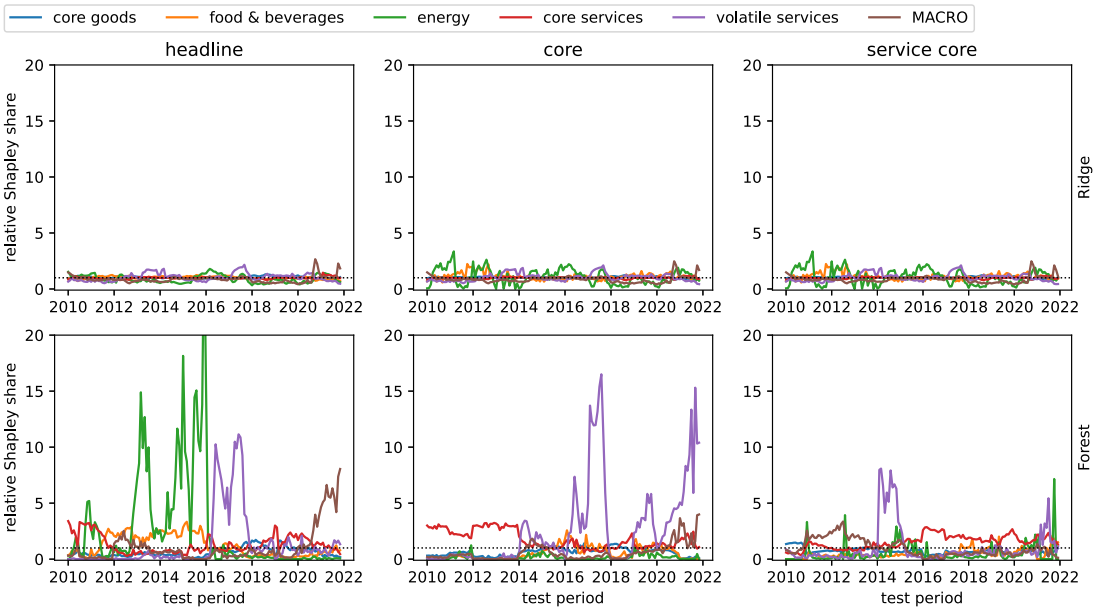

Fig. 4. Normalised Shapley weights over time.
Notes: Shapley value contributions from a given group of predictors, relative to the total sum of Shapley value contributions, normalised by the relative number of items in a group, for Ridge regression (upper row) and the Random Forest (lower), over the evaluation period. Shapley weights are from forecasts at a six-month horizon for different targets (columns).

rotation between service and goods prices. Consistently across both models, macroeconomic indicators provided over-proportional signals for headline inflation and rising absolute Shapley weights for this group of predictors since the end of 2020. The rise is more pronounced for the Random Forest, but even the linear Ridge regression shifts away from core goods and food items during that period towards macroeconomic data. Given that both models failed to forecast much of the rise in inflation in 2021, as seen in Fig. 2, the move away from relying on signals from these CPI item groups might have ultimately been detrimental for forecasting performance even if the models assigned stronger weights to them. Large fluctuations in core goods at that time might have reflected the effects of supply shortages in the aftermath of the pandemic rather than being mere outliers. For both models, this is challenging to pick up, having not seen a rise in inflation of this size and nature over the sample period.

Overall, the stable Shapley weight attribution of the linear Ridge regression suggests an advantage in terms of the robustness of model interpretations and performance. On the other hand, the Shapley weights from the Random forest are volatile. Still, some of the shifts over time appear intuitive and seem to reflect shifts in inflation dynamics occurring over time-related to macroeconomic shocks. While the highly non-linear random forest has the advantage of adapting to changes in the features of the data and exploiting signals from small groups of items over-proportionally, this does not necessarily lead to improved forecast performance. The machine learning model might have difficulty identifying the sub-groups of items with the strongest predictive content at a given point in time and distinguishing noise from meaningful shifts in volatility in the data.

6. Conclusion

We have conducted a forecasting exercise for UK inflation using a granular set of monthly CPI item series, as well as a set of macroeconomic indicators. 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. Shrinkage methods penalising the desegregated item series with varying forms of penalty result in the strongest forecast performance. Although it proves difficult to beat an autoregressive model over the entire sample period, the Ridge regression, a dense shrinkage method, yields significant forecast improvements when relying on disaggregated item series. When evaluated over periods during which the inflation variable of interest was rising, falling or at its tails, a wider range of shrinkage methods yields significant improvements, especially at longer forecast horizons of six and 12 months. Which shrinkage model performs best varies across horizons and regimes, suggesting some form of model averaging or “winner” selection across many specification runs might be preferable. Non-linear machine learning models perform less well overall, on average or over sub-periods. Evaluating forecast signals exploited over time using the Shapley value framework suggests that the Random Forest can flexibly shift between groups of predictors during periods when there were also shifts in inflation dynamics. Still, the non-linearity also adds much volatility to the signals it picks up, which looks

to be detrimental to forecast performance. The strongly performing Ridge regression instead assigns rather stable weights across time and forecast horizons.

The disaggregated CPI item series that we employ cover prices in a wide range of sectors and thus have the potential to help detect dynamics in the presence of shocks that affect prices across sectors differently or that might lead to spillovers across price categories, such as the shocks that have pushed up inflation in recently. However, models that have not seen such dynamics in the past, even models that are in principle highly flexible, will naturally have difficulty detecting the relevant signals in the data as long as there are few data points. Our analysis suggests that for the recent episode of rising inflation, both linear and non-linear models shifted away from reading signals from core goods in food price items, even though these item groups might have contained relevant signals related to supply shortages and food price shocks. These fluctuations were unseen over the sample period and thus identified as noise. An avenue of interest for future research is whether information from disaggregated item series can be exploited more effectively by separating lagged dynamics from the role of indicators in a more structured way within a range of machine learning models, potentially via enhancing such models with economically meaningful restrictions or scenario analysis. This could help regularise the high observed model variance in a way that aligns with economic intuition and the practitioners' judgement on the nature of economic shocks prevailing in the economy.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: George Kapetanios serves as editor to the International Journal of Forecasting.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2024.01.001.

References

Almosova, A., & Andresen, N. (2023). Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, *42*(2), 240–259.

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

Bańbura, M., & Bobeica, E. (2023). Does the Phillips curve help to forecast euro area inflation? *International Journal of Forecasting*, *39*(1), 364–390.

Beck, G., Carstensen, K., Menz, J.-O., Schnorrenberger, R., & Wieland, E. (2022). Using high-frequency scanner data to evaluate German food prices in real time. Unpublished manuscript.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., & Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, *10*, 615–643.

Borio, C., Hofmann, B., & Zakrajšek, E. (2023). Does money growth help explain the recent inflation surge? Bank for International Settlements Papers No 133.

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

Buckmann, M., Joseph, A., & Robertson, H. (2022). An interpretable machine learning workflow with an application to economic forecasting. Unpublished Manuscript.

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

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

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

Clark, T. E., Huber, F., Koop, G., & Marcellino, M. (2022). Forecasting US Inflation Using Bayesian Nonparametric Models. arXiv preprint arXiv:2202.13793.

Clark, T. E., Huber, F., Koop, G., Marcellino, M., & Pfarrhofer, M. (2023). Tail forecasting with multivariate Bayesian additive regression trees. *International Economic Review*, *64*(3), 979–1022.

Delle Chiaie, S., Ferrara, L., & Giannone, D. (2022). Common factors of commodity prices. *Journal of Applied Econometrics*, *37*(3), 461–476.

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

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

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

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

Forbes, K., Gagnon, J., & Collins, C. G. (2021). *Low inflation bends the Phillips curve around the world*: *NBER working paper no 29323*.

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

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica*, *74*, 1545–1578.

Goulet Coulombe, P. (2022). A neural Phillips curve and a deep output gap. Available at SSRN 4018079.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964.

Ha, J., Kose, M. A., & Ohnsorge, F. (2023). One-stop source: A global database of inflation. *Journal of International Money and Finance*, Article 102896.

Harding, M., Lindé, J., & Trabandt, M. (2023). Understanding post-covid inflation dynamics. *Journal of Monetary Economics*.

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

Hauzenberger, N., Huber, F., & Klieber, K. (2023). Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting*, *39*(2), 901–921.

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

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

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

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.

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

Klenow, P., & Kryvtsov, O. (2008). State-dependent or time-dependent pricing: Does it matter for recent U.S. inflation? *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.

Lopez-Salido, D., & Loria, F. (2022). Inflation at risk. Federal Reserve Board.
Lundberg, S., Erion, G., & Lee, S. (2018). Consistent individualized feature attribution for tree ensembles. ArXiv e-prints 1802.03888.
Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in neural information processing systems 30* (pp. 4765–4774).
Mann, C. L. (2021). On returning inflation back to target. Speech given at OMFIF, January 2021.
McAdam, P., & McNelis, P. (2005). Forecasting inflation with thick models and neural networks. *Economic Modelling*, *22*(5), 848–867.
Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119.
Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters*, *86*(3), 373–378.
Odendahl, F., Rossi, B., & Sekhposyan, T. (2022). Evaluating forecast performance with state dependence. *Journal of Econometrics*.
ONS (2019). Consumer prices indices technical manual. Weblinkhere.
Owyang, M. T., Piger, J., & Wall, H. J. (2015). Forecasting national recessions using state-level data. *Journal of Money, Credit and Banking*, *47*(5), 847–866.
Ozmen, U., & Sevinc, O. (2011). Price rigidity In Turkey : Evidence from micro data. Central Bank of the Republic of Turkey, Working Papers No. 1125.
Petrella, I., Santoro, E., & de la Porte Simonsen, L. (2018). Time-varying price flexibility and inflation dynamics. CEPR Discussion Paper No. DP13027.
Stock, J. H., & Watson, M. W. (1999). Forecasting inflation. *Journal of Monetary Economics*, *44*(2), 293–335.
Stock, J., & Watson, M. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, *97*, 1167–1179.
Stock, J., & Watson, M. (2002b). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162.
Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast? *Journal of Money, Credit and Banking*, *39*, 3–33.
Stock, J. H., & Watson, M. W. (2008). Phillips curve inflation forecasts. NBER Working Paper No 14322.
Stock, J. H., & Watson, M. W. (2016a). Core inflation and trend inflation. *The Review of Economics and Statistics*, *98*(4), 770–784.
Stock, J. H., & Watson, M. W. (2016b). Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In *Handbook of macroeconomics: vol. 2*, (pp. 415–525). Elsevier.
Stock, J. H., & Watson, M. W. (2020). Slack and cyclically sensitive inflation. *Journal of Money, Credit and Banking*, *52*(S2), 393–428.
Strumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, *11*, 1–18.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 267–288.
Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley&Sons Inc..
Wang, Y., Wang, B., & Zhang, X. (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., Long-ying, H., & Zhi-sheng, W. (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., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.