BANK OF ENGLAND

Staff Working Paper No. 915

Forecasting UK inflation bottom up

Andreas Joseph, Galina Potjagailo, Eleni Kalamara, Chiranjit Chakraborty and George Kapetanios

September 2022

This is an updated version of the Staff Working Paper originally published on 26 March 2021

BANK OF ENGLAND

Staff Working Paper No. 915
Forecasting UK inflation bottom up

Andreas Joseph,[1] Galina Potjagailo,[2] Eleni Kalamara,[3] Chiranjit Chakraborty[4] and George Kapetanios[5]

Abstract

We forecast CPI inflation in the United Kingdom up to one year ahead using a large set of monthly disaggregated CPI item series and a wide set of forecasting tools, including dimensionality reduction techniques, shrinkage methods, and non-linear machine learning models. We find that over the full sample period 2002–21, the Ridge regression combined with CPI item series yields substantial improvement against an autoregressive benchmark at the six-month horizon, whereas the benchmark is hard to beat with other models and for other horizons. However, when considering periods of time where aggregate CPI inflation measures exhibit changes in momentum (rising or falling) or tail values, a wide range of models leads to substantial significant relative forecast gains. Exploiting CPI items through shrinkage methods yields strongest gains at horizons of 6–12 months when headline and core inflation measures are rising or falling. At shorter horizons and when inflation is rising, machine learning tools combined with CPI items and macroeconomic indicators are more useful. We also provide a model-agnostic approach based on model Shapley value decompositions to interpret and communicate signals from groups of items according to interpretable CPI categories.

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

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

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

The views expressed in this paper are those of the authors, and not necessarily those of the Bank of England or its committees. We are grateful to the participants of the 2020 Conference on Real-Time Data Analysis, the 2020 Conference on Methods, and Applications, Modelling with Big Data & Machine Learning, the 2020 Banca d'Italia and Federal Reserve Board Joint Conference on Nontraditional Data & Statistical Learning, the Office for National Statistics, and the participants of internal seminars for highly valuable comments. Special thanks to Joshua Heming who pointed us to some inconsistencies in our analysis. Because of this and since we substantially extended our data set towards a longer sample period, some of our main results differ from those published in March 2021.

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

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

ISSN 1749-9135 (on-line)

1 Introduction

Forecasting consumer price inflation accurately in the near and medium term has large implications for monetary policy and other policy choices as well as business decisions in the wider economy. In particular during periods of changing momentum in inflation—such as the ongoing rise in inflation above inflation targets in many advanced economies in the aftermath of the Covid-19 shock—inflation forecasts move to the forefront of the policy debate. Accurate inflation forecasts are crucial for central banks for the design of appropriate and timely policy responses and for communicating the path at which inflation is expected to return to target. However, forecast performance can vary with the state of the economy (Odendahl et al., 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. During such periods, 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. At the same time, non-linear and non-parameteric models may be well suited to deal with large changes in both the predictors and the macroeconomic variables of interest.

In this paper, we explore 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.

In particular, we forecast monthly CPI headline, core, and service inflation in the United Kingdom at horizons of 1-12 months ahead. As predictors we use a large set monthly CPI items and, for comparison, a set of standard macroeconomic indicators. We evaluate a wide range of forecasting methods that exploit this large information set in different ways: dimensionality reduction techniques (Principal Component Analysis (PCA), Partial Least Squares (PLS)), shrinkage methods (Ridge, Lasso and Elastic Net regressions), as well as non-linear machine learning tools (Support Vector Machines (SVM), Artificial Neural Networks (ANN), Random Forests). We consider the period 2002m1-2021m11, evaluating the models using rolling window pseudo out-of-sample forecasts against an autoregressive 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 8 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 from considering disaggregated item level information. Item-level prices (e.g. "cereal bar",

"light bulb", "cinema admission") matter for aggregate inflation since they directly relate to the aggregate consumer price index. The ONS constructs aggregate CPI inflation from the item indices, that are themselves aggregations of price quotes collected in shops and centrally collected prices. The dynamics and inter-dependencies of disaggregated price items are complex and the distributional moments of item indices do not necessarily translate linearly to the aggregate level. As such, prices of different items or sectors can behave asynchronously, the frequency and dispersion of price adjustments can vary across items and over time, and the characteristics of certain groups of items can be over-represented in the aggregate (Chu et al., 2018; Petrella et al., 2019; Stock and Watson, 2019). This suggests that by incorporating item indices directly into a flexible model the forecaster is able to exploit a rich set of information (Hendry and Hubrich, 2011). The use of disaggregated information can also help to communicate adjustments to forecasts based on dynamics observed in different sectors.

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

Third, we provide a model-agnostic and flexible approach to address the "black box critique" of machine learning models that compares the signals from a diverse set of models in a uniform manner. We measure the contribution of individual items to forecasts using Shapley values (Strumbelj and Kononenko, 2010; Lundberg and Lee, 2017), and re-aggregate those into contributions from groups of items according to interpretable CPI categories.

Our findings are as follows. First, over the entire sample period, it is hard to significantly beat the AR benchmark. Only the Ridge regression using disaggregated item series achieves a significant and substantial improvement at the 6-month horizon for headline inflation, while LASSO slightly improves the core inflation forecast compared to the benchmark at the 3-month horizon. Second, the picture changes when evaluating the forecasts over sub-periods during which aggregate inflation is rising, falling, high or low, and a wide range of significant improvements against the benchmark is observed. This indicates that it is important to consider non-linearities over time when forecasting UK inflation. Exploiting a large set of predictors substantially helps to forecast inflation during turning points when inflation dynamics are changing or inflation outturns fall into tails. Third, there is not one single model that performs best across sub-periods and

horizons, and so it is advisable to consider a wide range of models. When inflation is rising or falling, shrinkage methods perform best at horizons of 6-12 months when combined with CPI items, whereas machine learning methods tend to be stronger when also fed with macroeconomic indicators and for shorter horizons of 1-3 months when inflation is rising. PCA tends to outperform the benchmark at different horizons during periods when inflation is high, low, or falling. Finally, for the current environment of rising and high inflation, the results over sub-periods imply that our approach can be particularly useful for forecasting headline and core CPI inflation with shrinkage or machine learning methods. Service inflation, on the other hand, can be difficult to forecast in the current environment given that our approach typically yields forecast gains when service inflation is falling, low or stable. However, useful signals may still be derived for turning points towards falling inflation levels.

We look at the Ridge regression and Random Forest to analyse model interpretability. This reveals intriguing model differences. The Ridge allocates mostly stable importances to different sub-groups across horizons and targets, as measured by group-aggregated Shapley value shares. The Random Forest, by contrast, shows variance across both dimensions. This is due to the Forest being the more flexible higher-variance model potentially returning a richer information set. Both models are also seen to give comparatively more weight, relative to the share of the corresponding inputs, to item sub-groups which are known to be more volatile, like energy or food & beverages. While we do not find a general association between model performance and sub-group importance, this analysis can offer a starting point into interpreting idiosyncratic drivers of predictions.

Our analysis relates to various strands of the forecasting literature. A vast literature focuses on forecasting inflation using a wide range of approaches such as Philips curve-based models (Stock and Watson, 1999, 2008), univariate unobserved component models (Stock and Watson, 2007, 2016), aggregation of forecasts of sub-components (Hubrich, 2005), Bayesian VARs (Koop, 2013; Domit et al., 2019), dimensionality reduction (Kim and Swanson, 2018) and medium-sized DSGE models (Carriero et al., 2019). 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 (Chen et al., 2001; McAdam and McNelis, 2005; Nakamura, 2005; Almosova and Andresen, 2019). Closer to our approach, Garcia et al. (2017) and Medeiros et al. (2019) forecast Brazilian and US CPI inflation, respectively, using large sets of macroeconomic predictors with various methods, where for the US the Random Forest performs best. Clark et al. (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. In a similar vein, Hauzenberger et al. (2022) provide evidence that non-linear dimension re-

duction techniques with shrinkage priors improve US inflation forecasts in real time, and that non-linear models are particularly useful during recessionary episodes.

Our analysis also relates to a rather small set of studies that have used disaggregated data to forecast aggregate series. Hernández-Murillo and Owyang (2006) and Owyang et al. (2015) use US state-level data to forecast national-level GDP while accounting for spatial interactions between the states, finding forecast gains relative to aggregate predictors. Hendry and Hubrich (2011) show that adding disaggregated sector-level information into forecast models improves forecast accuracy for aggregate US inflation. Aparicio and Bertolotto (2020) use combinations of high-frequency online price item series to forecast CPI one to three months ahead in ten advanced economies; their forecasts outperform benchmark models as well as surveys of forecasters by anticipating changes in official inflation rates. 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 to extract information from the data.

The remainder of the paper is organized 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 to our high-dimensional forecasting setting through Shapley value-based inference. Section 6 concludes.

2 Data

We use the headline CPI index from the UK 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.[1] 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

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

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, labor 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 and Watson, 2002a,b). This data also has the advantage of being readily available over longer sample periods and being continuously monitored by central banks and professional economists. Prior to estimation, the series are transformed to year-on-year log differences to achieve stationarity and are standardised (see Table B1 in Appendix B).[2]

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, are a guide for changes in households' living costs. While the CPI and price inflation are both macroeconomic concepts, they are constructed from the prices of single items over time, i.e. prices observed through local collection in physical shops or online or central collection in case of national prices. That is, item prices connect the disaggregated indices and aggregate inflation, which we exploit in this paper. The UK CPI is constructed by the ONS from an evolving set of representative monthly item indices, weighted according to household expenditure shares. At the lowest level, single item prices, or price quotes, are aggregated into item-level indices.[3] 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. But 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 structure of the data is a challenge 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

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

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

the composition of different goods in aggregate CPI by running estimations over rolling windows of item samples. We choose a window length of 8 years. Hence, we start with an initial balanced sample of items available over the period 2002-2009. We then iterate the sample forward, with items that are being discontinued at the end of the rolling window dropping out, and new items that are fully covered over the rolling 8-year window entering at each iteration step. Items that do not have coverage for that 8 years are dropped from that sample. For each estimation window, we estimate our models on the first 7 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 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 B1 in the appendix depicts the evolving sample size for the 8-year sample window, as well as for 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.[4]

We chain-link the item indices and take year-on-year log differences, which removes stochastic seasonality and smooths extreme observations through the log transform. Item series are mean-variance standardised in line with the expanding window approach of our forecasting setting described in Section 3. Figure B2 in Appendix B plots a selection of the transformed item series for illustration.

2.2 Descriptive statistics

To better understand how the item series dynamics compare to the aggregate CPI, we provide descriptive statistics for the disaggregated data we use. 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

[4]We ran estimations for the window length of 6 years with 5 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 2002-2008 and than 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.

(November 2021; see also (ONS, 2019)).[5] 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).[6] The series included in the balanced panel cover on average 69% of the item indices in each division. The left panel of the table shows the mean and standard deviation of yearly changes of our chained-linked index series. The mean across items for most CPI divisions is comparable to 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 the amount of heterogeneity in the disaggregated data.

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 until 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 of items entering and exiting the CPI basket over time. The number of items, balanced, refers to those included in our sample since they cover at least the 8 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.

This also becomes evident in Figure 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 CPI inflation (right panel). As previously documented (Klenow and Kryvtsov, 2008; Ozmen and 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.[7] In line with Table 1, the median of item index growth

[5] A set of zero-weight indices not in the CPI have been added to Housing & Fuel (440249, 410201, 410701, 410703, 410801, 440202, 610307, 610308).

[6] 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.

[7] A slight difference here to other studies is that we look at chained index series and not individual

rates is close to average headline inflation, both on average over the entire sample as well as over time. The fit of the median across item indices to aggregate inflation improves over time (right panel), in line with the improved coverage of item series through our rolling windows. On the other hand, there is a large amount of heterogeneity in 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).

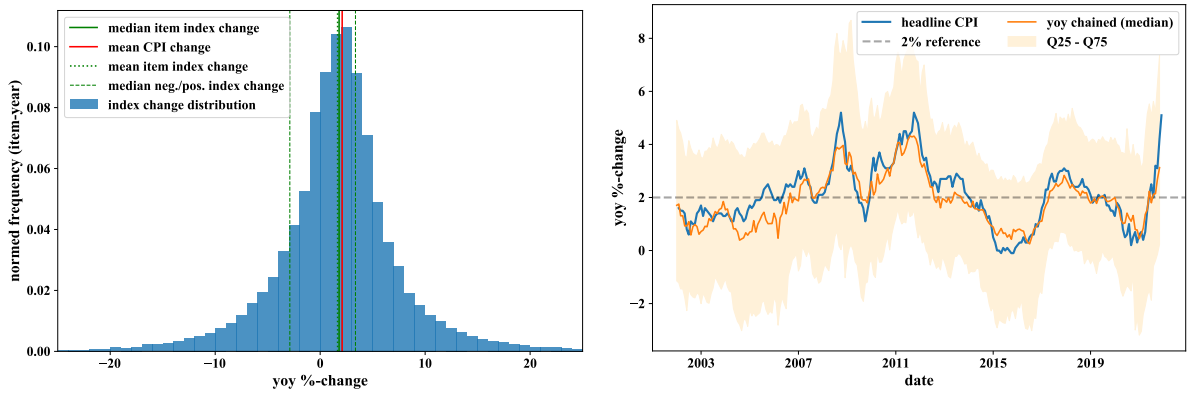

Figure 1: Distribution and moments of year-on-year growth rates in item-level CPI indices. Notes: Statistics are computed over chain-linked items that are included in our rolling window estimations. The left panel shows the histogram of CPI item growth rates over the entire sample period (blue bars), 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 inter-quartile range (orange swathe) of year-on-year growth rates of item indices over time, and for comparison headline CPI year-on-year inflation (blue line). Source: ONS and authors' calculation.

3 Methodology

3.1 Forecasting set-up and evaluation

We forecast monthly aggregate year-on-year UK CPI inflation (Headline, Core, or Services) over horizons of $h = 1, \ldots, 12$ month. As discussed in section 2, we start with an initial training sample for the period 2002m1-2008m12 over which we also tune model hyperparameters within the first seven years (see below for details), and we evaluate out-of-sample forecasts over the h subsequent months. We then iterate the training sample forward by one month, also adjusting the composition of CPI items to assure 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

item price quotes, with growth rates in the latter typically being centred around zero and hence more difficult to compare to aggregate price changes.

of the target variable, of the form

$$y_{t+h} = \alpha + \Sigma_{j=1}^{p} \gamma_j y_{t-j+1} + \epsilon_{t+h} \tag{1}$$

where y_t is the target variable, h is the forecast horizon, and the number of lags is set to $p = 2$ based on the Bayesian Information Criterion (BIC).[8]

To pin down the forecast gain from the large set of predictors relative to the $AR(2)$ benchmark, we proceed in two steps. First, we estimate an $AR(2)$ for inflation and we compute the residual $\tilde{y}_t$ to strip inflation off the part that is explained by the autoregressive component. Then, we feed the residual as target to the set of models outlined below. We, thus, look to forecast inflation using the large set of predictors conditional on the autocorrelation in inflation being already accounted for. This two-step approach assures that the signals from the large set of predictors are picked up separately from the signals coming from lagged dynamics, and hence it helps the models to exploit the large data set more efficiently. By contrast, adding the lags together with a large number of disaggregated predictors in a single forecasting step, can result in each predictor individually having a very low impact compared to the two lags, given that inflation follows a highly persistent process. This can blur the different signals and can result in the models laying an overly strong weight on the lagged coefficients. Further, since a number of models we use are nonlinear in nature, having lagged dynamics added to such a nonlinear specification can make accounting for such dynamics more complex. Therefore, removing lagged dynamics, as we do in a first step, places all models on a equal footing in terms of evaluating their predictive performance.

We evaluate the average precision of the forecasts against the $AR(2)$ benchmark based on relative root mean squared errors (RMSE). We test for statistical difference in forecast accuracy using the Diebold and Mariano (1995) test with Harvey's correction for short samples (Harvey and Newbold, 2000). First, we run the forecast evaluations over the full test period 2009m1 to 2021m11. Additionally, we are interested in whether our models with disaggregated CPI items might perform differently in periods where the level of inflation or the momentum in inflation are high (low) compared to the overall sample period. We therefore also evaluate the out-of-sample forecasts over sub-sets of months where the outturn of the aggregate inflation series that we forecast fell into certain segments.

3.2 Overview of forecasting methods

All the forecasting methods we present here have the advantage that they can deal with large datasets and thus wider information sets. We are given a dataset of a large number of predictors $x_t = (x_{1t}, \ldots, x_{Nt})\prime$, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. We are interested in

[8]The value $p = 2$ is very stable across time, such that we use this value throughout.

forecasting the residual $\tilde{y}_t$ of headline CPI inflation, after stripping it from the impact of its own lags in a AR(2) regression, in period $t+h$, based on a set of predictors x_t[9]

$$\tilde{y}_{t+h} = \alpha + \Sigma_{i=1}^{N} \beta x_t + \Sigma_{j=1}^{p} \gamma_j y_{t-j+1} + u_{t+h}. \quad (2)$$

Due to the large number of predictors, estimating (2) directly with each predictor included individually would lead to high estimation uncertainty and a lack of degrees of freedom, where the dimension of β might be larger than T. Such a model thus suffers from over-parametrization, and some form of dimensionality reduction is required. The models we employ take different approaches to deal with this, either by reducing the dimensionality of the input space directly or via explicit or implicit weighting (shrinkage). The main ideas of the models are outlined below, with details presented in Appendix A.

Dimensionality reduction techniques

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 summarized effectively in a small set of common components. This substantially reduces the number of parameters in the model, addressing over-parametrisation and degrees of freedom issues in rather short samples. In particular, a vector of N indicator series x_t is summarized by a vector $r \times 1$ of finite latent components f_t. The two models have in common that they use information densely, i.e. information from a wide range of available predictors is drawn upon by summarising them through common components. PCA summarises the joint variability of predictors x_t into a static factor which is added into a prediction regression as in equation 2. 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 y_{t+h} is maximised.

Shrinkage methods

The goal of shrinkage methods is, using different penalisation schemes, to reduce the dimension of the matrix of indicator series x_t. This produces linear combinations of the original regressors, where those coefficients that do not carry any predictive power 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 shrinks the coefficients of predictors that contribute little to the predictive ability of the model towards zero, albeit they never become exactly zero—it is therefore a dense model which draws on

[9]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.

all available information albeit to different degree. In the case of no shrinkage, i.e. $\lambda = 0$, Ridge regression becomes equivalent to a linear OLS regression. LASSO regression, on the other hand, penalises the sum of squared residuals according to the sum of absolute coefficients which results in some of the coefficients being shrunk to exactly zero. It is thus a sparse model which performs shrinkage through variable selection. The Elastic Net is a hybrid approach which combines these two types of shrinkage: in a first step, it finds Ridge regression coefficients and, in a second-step, Lasso-type shrinkage i.e. variable selection is applied. A correction factor is applied to account for increased bias through double shrinkage.

Non-Linear Machine Learning Models

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

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

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

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

Artificial Neural Networks (ANN) consist of an input layer, at least one hidden layer, and an output layer. Layers are connected via the network weights W representing the model parameters and pass through non-linear activation functions at each hidden layer. Note that, without hidden layer, an ANN becomes a linear function and is similar to solving the least squares problem. We use multilayer perceptrons (MLP), a form of feed-forward network, as ANN architecture. The activation function $g(z_t, W)$ acts as a gate for signals and introduce non-linearity into the model. Its functional form is subject to hyperparameter tuning. The variables z_t in the input layer are multiplied by weight matrices W at each layer, then transformed by an activation function in the hidden layers

and passed on through the network until the linear output layer is reached resulting in a prediction $\hat{y}_{t+h}$. Deeper networks are generally more accurate but also require more data to train them due to the larger number of parameters in the weight matrices. The number of hidden layers, i.e. the depth of the network, and the number of neurons in each layer as well as appropriate weight penalisation in our ANN are hyper-parameters, and are determined by cross-validation as discussed below.

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

3.3 Tuning of hyperparameters

All of our models require some form of hyperparameter selection prior to estimation. In the case of dimensional reduction techniques, we use a form of information criteria (e.g. AIC, BIC) to choose the lag length or the number of common components. In cases where the derivation of information criteria is not feasible, such as the shrinkage methods and machine learning tools, we use cross-validation procedures.[10] The main difference between information criteria and cross-validation methods is that the latter depends on out-of-sample performance, whereas information criteria are "in-sample" statistics.

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

The hyper-parameters selected through cross-validation include the penalty imposed on shrinkage methods but also the maximum depth of trees for the Random Forest, the architecture of the ANN and the choice of the kernel function for SVM.[11] Given that the estimation is done over rolling window, the selected hyperparameters can change over windows, as well as over forecast horizons and specifications. Overall, cross-validation

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

[11] We choose the following grid sets: For Ridge, LASSO and Elastic Net $\alpha \in \{1e-05, 0.0001, 0.001, 0.01, 0.1, 1.0\}$, for Elastic Net L1-ratio $\in \{0.1, .5, .9, .95, 1\}$, for Forest max. depth $\in \{1, 2, 3, 5, 6, 7, 8, 9, 10\}$, for ANN hidden layer dimension $\in \{(2, 3), 10, 2), (20, 2), (2, 3), (20, 3), (5, 5)\}$ and activation function tanh or ReLU, for SVM $C \in \{100, 10, 1000\}$ and $\epsilon \in \{0.01, 0.1, 0.5, 0.9\}$, the kernel is chosen to be RBF.

favours quite similar parameters and model architectures across specifications that use different sets of predictors and also across horizons, although they evolve somewhat over time, i.e. over rolling window test sets.[12] 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 item and macroeconomic time series, the procedure selects deeper versions of the network and larger tree structures of the Random Forest. This suggests an increase in complexity as more data are involved in training the model. Regularisation plays an important role both for linear and non-linear models. Notably, the Ridge regression always imposes a heavy penalty which might explain the overall strong performance of this model.

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. The comparison with the AR model provides information on the marginal forecast gain through the inclusion of a large set of predictors, and it indicates a ranking between models in terms of the extent to which they outperform or lose out against the AR.

We start with results over the entire sample period. We then present forecasting results over sub-periods with certain inflation characteristics, i.e. periods of rising and falling inflation or high and low inflation. This allows us to assess whether the use of disaggregated data and models such as non-linear machine learning tools might be particularly useful during periods where inflation experiences turning points or tail outcomes, and where a simple AR model might have more difficulties forecasting inflation compared to more normal times.

4.1 Forecast results over the entire sample period

Table 2 shows relative RMSE against the benchmark for the specification with CPI item series as predictors over the entire sample period. Forecasts for the three different inflation measures are shown in the three panels of the table for selected forecast horizons. Results indicate that it is difficult for the models to significantly outperform the $AR(2)$ benchmark. The shrinkage methods perform best for forecasting headline CPI and Core CPI inflation. The Ridge model provides a substantial improvement against the AR at the 6-month horizon, with a relative RMSE of 0.86, and a smaller and insignificant improvement at the 12-month horizon. The LASSO model provides a significant improvement at the

[12] Results for selected hyperparamerters are not shown for space constraints and are available upon request.

3-month horizon for Core CPI inflation, and also some insignificant improvements for headline CPI. For Core Service inflation, which is a rather stable component that seems to be captured well by an $AR(2)$, none of the models outperforms the benchmark.

Table 2: Forecasting exercise results, CPI items series predictors.

Predictors: CPI item indices

	Target: headline CPI				*Target: Core CPI*				*Target: Service CPI*		
horizon	*3*	*6*	*12*	*horizon*	*3*	*6*	*12*	*horizon*	*3*	*6*	*12*
PCA	1.0	0.98	1.04	PCA	0.96	0.97	1	PCA	1.03	1.06	1.08
PLS	1.09	1.19	1.29	PLS	1.1	1.08	1.11*	PLS	0.98	1.06	1.11
Ridge	1.0	**0.84***	0.96	Ridge	1.05	0.89	1.02	Ridge	1.01	1.01	1.01
Lasso	0.98	1.01	0.96	Lasso	**0.97***	1.05	0.98	Lasso	1.04	1.09**	1.06
Elastic	1.0	1	0.98	Elastic	0.99	0.99	1	Elastic	0.99	1.01	1.03
SVM	1.0	1.02	1.14	SVM	1.02	1.03	1.08	SVM	1.04	1.05	1.03
Forest	0.99	1.07	1.21	Forest	1.01	1.09	1.19	Forest	1.06*	1.11**	1.14**
NN	1.19**	1.24*	1.16	NN	1.12	1.35	1.11	NN	1.4	1.26	1.32**

Notes: Root mean squared errors, relative to $AR(2)$ model. Forecasts of headline CPI inflation (left panel), Core inflation (middle) and Service inflation (right) using CPI item series as predictors. Rolling window samples over sample period 2002-2021 with 7 years of training sample and out-of-sample forecasts at horizons of 3, 6, and 12 months. 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 not significant and are not presented for space constraints.

The above findings are reflected in the model predictions of headline inflation shown in Figure 2 for the specification with CPI items only. At a forecast horizon of 3 months, the AR benchmark forecasts capture actual inflation outturns closely, albeit with slight delay. All of our models' forecasts are quite close to the AR benchmark, but most have some excess volatility which explains their weaker performance. At the horizon of 6 months, all models somewhat lag behind the target but get most of the dynamics right. The good performance of the Ridge regression (middle panel) relative to the benchmark becomes evident particularly during periods where inflation experiences turning points. For instance, the rise in inflation during 2011 is captured precisely and without delay by the Ridge regression. The subsequent decline in inflation during the years 2012-2015 is also captured by the Ridge regression, albeit incompletely and with some excess fluctuations. During the periods of rising inflation over the years 2016-2017 and most recently in 2021, the Ridge regression is the model that was closest to the actual outcome. Whereas many models added volatility to the forecasts during these periods, the strong shrinkage underlying the Ridge model likely helped muting volatile signals from individual predictors. All models lag behind the actual outcome at a horizon of 12 month and only capture part of the dynamics. The models combined with CPI items show more fluctuations than the AR, which can explain their worse performance, but which may render them useful during certain periods of more volatile inflation. These observations suggest that the Ridge regression, and other models combined with CPI items, might be useful during periods where the inflation momentum is changing or where the level of inflation is particularly high or low, compared to the overall sample. We will investigate this further

in the next sub-section by evaluating our forecasts over periods where inflation outcomes have particular characteristics.

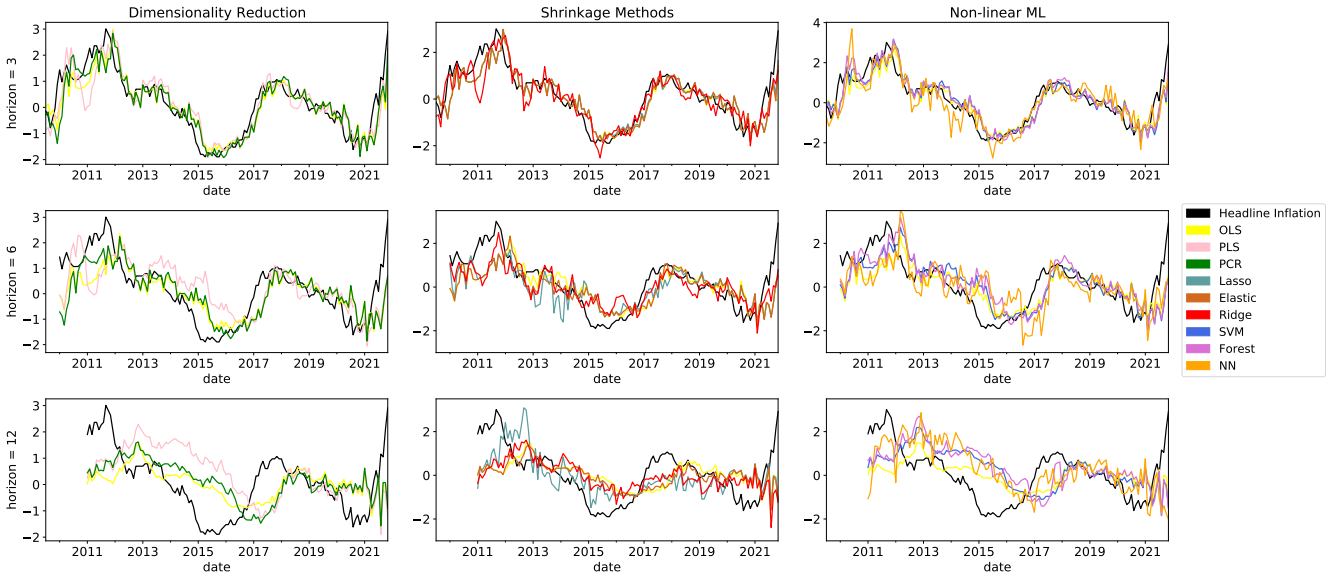

Figure 2: Predicted values for headline CPI inflation.
Notes: Forecasts of CPI headline inflation (standardised), from different types of forecasting models (columns, coloured lines) using CPI items as predictors, over horizons $h = 3, 6, 12$ (rows). Out-of-sample predictions for rolling samples from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

Table 3 shows results for headline CPI inflation forecasts using macroeconomic indicators either in combination with CPI item series (left panel) or on their own (right panel). This allows to assess whether the disaggregated CPI item series are relevant predictors beyond the predictive power from macroeconomic dynamics, and whether certain models perform better when fed with different types of predictors. The additional gain from macroeconomic indicators is small overall.[13] When using them in combination with item series, a couple of models show small improvements against the $AR(2)$, but those are insignificant. Hence, including even more series and combining series of different nature does not seem to be helpful and rather results in a loss of degrees of freedom. When using macroeconomic indicators on their own, LASSO performs better than in the specifications with item data and reaches significant improvements against the $AR(2)$ at the 3-month and 12-month horizon. On the other hand, the Ridge regression performs much worse with macroeconomic data than with CPI item data, and now significantly loses out against the benchmark. Overall, this shows that different models are preferred with different types of data. When using the highly disaggregated CPI data, the Ridge regression works particularly well, i.e. a dense model that shrinks many of the parameters towards but not exactly zero, and does not do variable selection. When using macroeconomic data, LASSO which performs shrinkage akin to variable selection works better for forecasting

[13]For Core inflation and Service Core inflation forecasts, there were no improvements from adding macroeconomic indicators.

inflation. This is different from existing evidence for GDP forecasts using macroeconomic data, where previous studies found dense models such as the Ridge regression to perform best (Giannone et al., 2017).

Table 3: Forecasting exercise results, Macroeconomic indicators as predictors.

Predictors: CPI items & Macro ind.				**Predictors: Macro ind. only**			
	Target: headline CPI				*Target: headline CPI*		
horizon	*3*	*6*	*12*	*horizon*	*3*	*6*	*12*
PCA	1.04	1.01	1	PCA	0.98	0.99	1.1
PLS	1.3	1.2*	1.09	PLS	0.98	1.01	1.81
Ridge	0.97	0.9	0.99	Ridge	0.99	1.22**	1.72**
Lasso	0.94	1.05	0.98	Lasso	**0.92****	0.97	**0.99***
Elastic	0.98	0.95	1	Elastic	1	0.98	1.02
SVM	1.14	1.01	0.99	SVM	0.96	1.18	1.52**
Forest	1.2	1.07	0.96	Forest	0.97	0.99	1.18
NN	1.84**	1.25*	1.28**	NN	0.95	1.16*	1.44**

Notes: Root mean squared errors, relative to AR(2) model. Forecasts of headline CPI inflation using CPI item series and Macroeconomic indicators (left panel) or Macroeconomic indicators only (right panel) as predictors. Also see notes for Figure 2.

4.2 Forecast results for specific inflation outcomes

We have seen that it is difficult to beat the $AR(2)$ model significantly on average over the entire sample period. However, the comparison of predicted values indicated that there might be differences in the relative performance of the models during sub-periods where inflation was for instance high or rising. We therefore evaluate the forecast errors separately over months where aggregate inflation displayed certain characteristics. For this purpose, we define sub-periods during which the outturn of the aggregate inflation series that we forecast 1) is *high* (above 3%, i.e. in its upper quartile since the year 2008), 2) *low* (below 1.5%, i.e. in its lower quartile), 3) within its *interquartile range (IQR)* (between 1.5% and 3%), 4) has positive momentum meaning that inflation is *rising* for an extended period (defined as the 3-month moving average of percentage point change in inflation being positive for at least five consecutive months), 5) negative momentum or in other words inflation is *falling* for an extended period (3-month moving average of percentage point change in inflation being positive for at least five consecutive months), 6) or else when inflation is *stable*.

Table 4 shows the results for headline inflation forecasts, over different horizons (columns) and using different sets of predictors (panels). For readability, the table only shows the sub-periods and models for which we observed significant improvements against the benchmark, and types of models are coloured differently (dimension reduction techniques in blue, shrinkage methods in red, machine learning tools in green). We find a wide

range of significant improvements with RMSE relative to the benchmark of 0.75-0.95 certain inflation regimes. Which models beat the benchmark depends on the horizon and also on the set of predictors that they are combined with.

Table 4: Forecast performance over sub-periods with different inflation characteristics (Headline inflation).

Headline CPI inflation forecast — CPI items only								
Horizon	1		3		6		12	
CPI inflation falling	-		PCA	0.92**	Elastic	0.91**	Elastic	0.93***
					Ridge	0.74***		
CPI inflation rising	-		Forest	0.88**	Lasso	0.92*	Lasso	0.89***
					Ridge	0.75***	Elastic	0.99*
CPI inflation stable	-		-		-		Ridge	0.86*
CPI inflation low	PCA	0.94*	PCA	0.95*	-		Ridge	0.84*
CPI inflation high	-		-		Ridge	0.8***	PCA	0.87**
CPI inflation in IQR	-		-		-		Elastic	0.95***
Headline CPI inflation forecast — CPI items & Macro indicators								
Horizon	1		3		6		12	
CPI inflation falling	-		PCA	0.93**	Elastic	0.9**	Elastic	0.93***
					Ridge	0.78***		
CPI inflation rising	Ridge	0.78*	Forest	0.81**	Ridge	0.82***	Lasso	0.88***
							Elastic	0.99*
CPI inflation stable	-		-		SVM	0.89*	-	
CPI inflation low	-		PCA	0.94*	-		Ridge	0.84*
CPI inflation high	-		-		Ridge	0.88***	PCA	0.87**
CPI inflation in IQR	-		-		-		Elastic	0.94***
Headline CPI inflation forecast — Macro idicators only								
Horizon	1		3		6		12	
CPI inflation rising	Lasso	0.9**	Forest	0.88**	-		-	
	NN	0.87*	SVM	0.9*	-		-	
CPI inflation high	PCA	0.84*	-		-		PCA	0.79*

Notes: Relative RMSE computed over sub-periods with different inflation characteristics. These are periods during which the actual headline CPI inflation outcome at a given horizon fell into a certain category (falling, rising or stable inflation, or inflation in the upper/lower quartile of the inflation distribution or in its interquartile range). Only models with significant gains in terms of relative RMSE against the AR(2) benchmark over a sub-period and horizon are listed ("-" indicates that there were no significant gains for a sub-period/horizon; sub-periods where no models showed significant gains for a specification are not listed). Significance of forecast accuracy is assessed via Diebold and Mariano (1995) test statistics with Harvey's adjustment. Significance levels: ***:1%, **:5%, *:10%.

When using CPI item series (either alone or in combination with macroeconomic indicators), shrinkage methods perform strongly at higher horizons of 6 and 12 months. At the 12-month horizon, there are significant improvements compared to the benchmark from at least one model for most sub-periods, whereas at the 6-month horizon, improvements are achieved for the periods where inflation is falling, rising or high. The Elastic Net performs particularly well when headline inflation is rising, while the LASSO performs well when it is falling. The Ridge regression does well for a wider range of sub-periods, particularly at the 6 month horizon, in line with its strong performance over the entire sample period presented above. Here, Ridge regression significantly improves against the benchmark with a relative RMSE of 0.75 during periods of falling inflation and rising

inflation, respectively, and 0.8 during periods of high inflation. It also shows a significant improvement with relative RMSE of 0.85 during periods of stable or low inflation at the 12-month horizon but does not perform well anymore for the other sub-periods. On the other hand, PCA and two of the machine learning tools, Random Forest and SVM, achieve improvements against the benchmark mostly at lower to medium horizons, although PCA also forecasts well at the 12 month horizon when inflation is high.

There is little gain overall from adding macroeconomic data compared to exploiting CPI items alone. However, the Random Forest does gain from the joint set of macroeconomic indicators and item series predictors and reaches a relative RMSE of 0.81 during periods when inflation is rising. Using macroeconomic indicators alone also helps at low horizons when CPI inflation is rising or high. Machine learning tools in particular are able to achieve forecasting gains from exploiting the macroeconomic indicators in the short run when inflation is rising, with relative RMSE of 0.87-0.9 at the 1-month and 3-month horizons. This can indicate the presence of a potentially non-linear Phillips Curve relationship that machine learning tools are able to exploit to achieve forecasting gains in the short run, while other linear models hardly draw any signals from macroeconomic data for inflation.

Table 5 shows the corresponding results over sub-periods for the specifications for core and service CPI inflation. Whereas over the entire sample we did not observe any significant forecast gains against the benchmark, there are substantial improvements for some of the sub-periods. For core inflation forecasts, similarly to headline CPI forecasts, shrinkage methods combined with CPI items yield improvements at higher horizons. The Elastic Net yields a significant improvement for falling core inflation, and the LASSO for rising core inflation at the 12-month horizon, while the Ridge regression improves against the benchmark during periods of rising core inflation at the 6-month horizon. The PCA also shows gains when core inflation is high at shorter horizons and when core inflation is falling at the 12-month horizon. Again, machine learning tools work best when exploiting macroeconomic indicators over sub-periods. The Random Forest and Artificial Neural Net achieve significant relative RMSE of 0.84-0.91 at the 1-month horizon during periods of rising or high core inflation, but also at higher horizons with the SVM at 6-months and with the Random Forest at 12-months. Shrinkage methods, in particular the Elastic Net and LASSO also show significant improvements at the 1-month horizon, but these are weaker compared to the machine learning tools. Turning to service inflation, on the other hand, shrinkage methods and PCA dominate. Here, substantial improvements in RMSE of up to 40% are achieved at horizons of 3 to 12 months when service inflation is stable, falling, or low. The strongest improvements are achieved when CPI items are combined with macroeconomic data. However, no improvements for service inflation forecasts are achieved from using macroeconomic indicators alone, or for any set of predictors during periods when service inflation is high or rising.

Table 5: Forecast performance over sub-periods with different inflation characteristics (Core and Service inflation).

Core CPI inflation forecast — CPI items only								
Horizon	1		3		6		12	
Core inflation falling	-		Lasso	0.84*	-		Elastic	0.86*
							PCA	0.89***
Core inflation rising	-		-		Ridge	0.82*	Lasso	0.89**
Core inflation high	PCA	0.89*	PCA	0.91*	-		-	
Core CPI forecast - CPI items & Macro indicators								
Horizon	**1**		**3**		**6**		**12**	
Core inflation falling	-		-		-		PCA	0.89***
Core inflation rising	-		-		-		Lasso	0.88***
Core inflation high	-		-		-		SVM	0.94***
Core CPI forecast - Macro indicators only								
Horizon	**1**		**3**		**6**		**12**	
Core inflation rising	Elastic	0.99*	-		Ridge	0.82*	-	
	Lasso	0.98*			SVM	0.95*		
	Forest	0.89**						
	NN	0.84*						
Core inflation high	Elastic	0.99*	-		-		Forest	0.89**
	Lasso	0.98*						
	Forest	0.91*						
	NN	0.87*						
Service CPI forecast — CPI items only								
Horizon	1		3		6		12	
Service inflation falling	-		Ridge	0.84**	Ridge	0.71**	Elastic	0.94**
							Lasso	0.86*
Service inflation stable	-		Elastic	0.97**	PCA	0.86**	-	
Service inflation low	-		Ridge	0.87*	-		Elastic	0.97**
							Ridge	0.89***
Service CPI forecast - CPI items & Macro indicators								
Horizon	**1**		**3**		**6**		**12**	
Service inflation falling	-		Ridge	0.68***	Ridge	0.68**	Lasso	0.87***
Service inflation stable	-		Elastic	0.59**	PCA	0.83**	-	
			PCA	0.89*				
Service inflation low	-		Ridge	0.75**	-		Ridge	0.88***

Notes: Relative RMSE computed over sub-periods with different inflation characteristics. These are periods during which the actual inflation outcome (Core CPI or Service CPI) at a given horizon fell into a certain category (falling, rising or stable inflation, or inflation in the upper/lower quartile of the inflation distribution or in its interquartile range). Also see Notes to Table 4.

All in all, various findings emerge from the evaluation of forecasts over sub-periods. First, we obtain a wide range of improvements against the benchmark for periods when inflation is rising, falling, high or low. This indicates that it is important to consider non-linearities over time when forecasting UK inflation. The consideration of the entire sample period masks differences across sub-periods, and while an AR is hard to beat during periods of stable inflation or on average across different periods, exploiting a large set of predictors does help forecast inflation during turning points and when inflation

dynamics are more extraordinary. Second, there is not one single model that performs best across sub-periods and horizons, such that it is advisable to consider a wide range of models. When inflation is rising or falling, shrinkage methods perform best at horizons of 6-12 months when combined with CPI items, whereas machine learning methods tend to be stronger when also fed with macroeconomic indicators and for shorter horizons of 1-3 months when inflation is rising. Third, with regard to the current environment of rising and high inflation, the results over sub-periods imply that our approach can be particularly useful for forecasting headline and core CPI inflation. Service inflation, on the other hand, can be difficult to forecast in the current environment given that our approach typically yields improvements when service inflation is falling, low or stable, but could become useful again to detect turning points towards falling inflation.

5 Opening the forecasting "black boxes"

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

For the interpretation of results, three questions are of interest. First, what is the contribution of a predictor to the forecast? Second, are certain sub-groups of predictors (i.e. CPI items that belong to a certain item group, e.g core items, energy items, or goods and service items) more relevant than others? And if they, does this old across time, predictors and targets? Answers to these questions can be informative for identifying relevant predictors for lower-dimensional forecasting frameworks, or for communicating forecasts and to inform policy decisions. We address this through a model-agnostic approach that is based on three steps: model decomposition, partial re-aggregation or grouping, and the investigation of group properties. We describe this approach in more detail in the following before presenting results.

5.1 Shapley values to explain statistical models

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

analogy between variables in a model and players in a cooperative game. The Shapley value framework has a set of appealing analytical properties while being applicable to any model.[14] It consists in calculating the 'payoff' for including a specific predictor in the model, conditional on other predictors being present. Each prediction (i.e. a predictive value at time t and horizon h) from a model is decomposed into the sum of contributions from the individual input variables, the so-called *Shapley values*.

Let the total number of predictors be $M = |\mathcal{M}|$ from the set $\mathcal{M}$ of price items and macroeconomic series as described in Section 3. A predicted value, $\tilde{y}_{t+h}$, of the residual at time t for the forecast at horizon h can be decomposed into its Shapley components ϕ_{tj}^h for the j^{th} variable. That is,

$$\tilde{y}_{t+h} = \sum_{j=0}^{M} \phi_{tj}^h \,, \qquad \text{(decomposition)} \tag{4}$$

The $j = 0$ component is set to the mean predicted value in the training set and can be interpreted as an intercept.[15] For a non-linear forecasting model, computation of (4) requires deriving the marginal contribution of predictor j by running sequential forecasts of all possible coalitions of predictors, with and without j. Thus, the Shapley value for predictor j (ignoring time subscript and forecast horizon superscript for the moment) is computed as

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

Here, the payoff of a coalition $\mathcal{S} \subseteq \mathcal{M}$ is $f(\mathcal{S})$, the payoff of this coalition combined with predictor j is $f(\mathcal{S} \cup \{j\})$, and their difference measures the marginal contribution of j to that coalition. The intercept ϕ_0 corresponds to $f(\varnothing)$, i.e. with no variables in the model. After summing these marginal contributions over all coalitions, we get an estimate of the contribution of variable j to a single model prediction. Comparing all possible combinations of predictors with $M \approx 400$ is computationally infeasible.[16] We therefore focus our analysis on models where an exact solution exists, namely linear models and the Random Forest. For a linear regression model, the Shapley value of predictor j is simply the product of its regression coefficient w_j and the difference between the predictor value and its mean, i.e. $\phi_j = w_j(x_j - \mathbb{E}_t[x_j])$ with the expectation taken over the training dataset. For the Random Forest, or tree-based models more generally, variable coalitions correspond to paths down the branches of the model where these variables lie on the same

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

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

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

branch. These can generally be enumerated easily, reducing the complexity of the sum in Eq. 5 (see Lundberg et al. (2018) for details). For other models, coalitions can be sampled with a readjustment of the weights in (5).

The second step is **context-specific grouping and re-aggregation**. Predictors in our case are mostly disaggregated price indices. Their values have a clear interpretation as the relative price of a narrowly defined product at a given point in time. However, single item series can be volatile and difficult to keep track of, as is the case with the corresponding Shapley values. We therefore group the M model components ϕ_{tj}^h into $K << M$ higher-level sub-groups denoted $\mathcal{G}_k$, which represent an an aggregation level between the disaggregated input and aggregate inflation to be forecast,

$$\hat{y}_{t+h} = \sum_{k=0}^{K} \psi_{tk}^h \quad \text{with} \quad \psi_{tk}^h = \sum_{j \in \mathcal{G}_k} \phi_{tj}^h \qquad \text{(sub-group aggregation)}\,, \tag{6}$$

with $\psi_{t0}^h = \phi_{t0}^h$ being the same intercept. We opt for grouping CPI items according to categories conventionally used in decision making situations, such as sub-categories of CPI typically monitored by central banks. These are goods and services, with the former two additionally divided into less volatile core goods (services) and more volatile components such as energy and food & beverages. We also aggregate all macroeconomic indicators into a sub-group. This gives us six mutually exclusive sub-groups of predictors: core goods, food & beverages, energy, core services, volatile services, and macroeconomic indicators.[17]

In the last step, we look at two simple metrics based on sub-group Shapley values, namely the **absolute and relative weights** given to a group for forecasts at a particular horizon,

$$\bar{\Phi}_k^h = \frac{|\Phi|_k^h}{\sum_{k=1}^{K} |\Phi|_k^h} \quad \text{with} \quad |\Phi|_k^h = \sum_{t \in \mathcal{T}} \sum_{j \in \mathcal{G}_k} |\phi_{tj}^h| \quad \text{(absolute weight)} \tag{7}$$

$$\tilde{\Phi}_k^h = \frac{\bar{\Phi}_k^h}{M_k/M} \qquad \text{(relative weight)}\,, \tag{8}$$

where $\mathcal{T}$ collects time indices, such as the whole test period or those during particular macroeconomic inflation regimes, like falling or rising aggregate prices, over which these metrics are averaged. Absolute sub-group Shapley weights are shares and sum to one over all K groups. Relative weights penalise absolute weights according to the share of indices entering a model averaged over the evaluated test sample. These have an expected value of one and measure how much a model relies on a certain sub-group compared to a uniform representation of all variables in terms of Shapley attributions.[18] If the relative

[17] The detailed groupings can be provided upon request.

[18] Sub-groups should not be too small compared to the overall number if inputs, as this can potentially

Shapley weight of a sub-group is greater than one, a model puts over-proportional weight on that group compared to other groups, and vice versa for relative weights below one.

5.2 Results based on Shapley values

We first look at the absolute Shapley weights of sub-groups for different targets. We focus on the specifications including macroeconomic variables as our analysis allows for a direct comparison between groups, be it different item groups or types of variables.

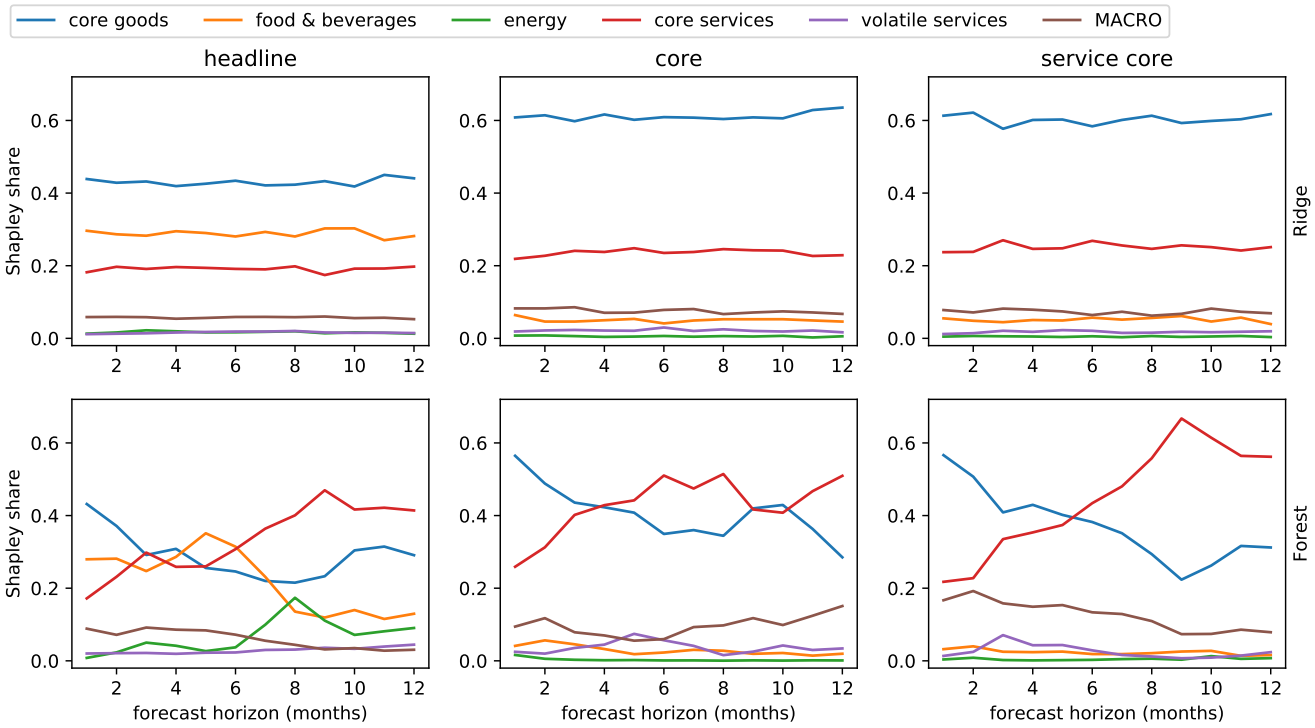

Figure 3: Absolute Shapley weights of sub-groups (colour) by forecast horizon.
Notes: Rows show the Ridge regression (upper) and the Random Forest (lower). Columns different macroeconomic inflation targets: headline (left), core (middle), and services (right).

The absolute weights for different groups as a function of the forecasting horizon are shown in Figure 3 for the different models (rows) and targets (columns). Core goods items are attributed the largest shares across models and horizons for the Ridge. This is in line with these also being the largest groups by number and weight in the CPI basket. For headline inflation, food & beverage items also play a sizeable role for both models, whereas their share drops for the other targets, where this items are less prevalent. An important difference between the Ridge regression (upper row) and the Random Forest (lower row) for all targets are the different paths of Shapley value shares over forecast horizons. The Ridge regression gives stable weights to all groups across forecast horizons.

cause divergence and associated instability caused by the denominator of Eq. 8. We do not find this to be of concern in our case, while this may increase the volatility of relative Shapley weights in some of our results, such as for the energy group of items.

The Random Forest shows a similar overall ordering of group shares, but Shapley value shares fluctuate much more across horizons. As such, the Random Forest increasingly draws on core services—a less volatile component—for higher horizons, and less on goods and food & beverages. The Random Forest also draws on energy items for predicting headline inflation over medium horizons, whereas Ridge regression relies very little on these items. The Shapley value shares of macroeconomic indicators are comparatively low for both models, reflecting the result from the forecasting exercise of little added contribution from macroeconomic indicators once CPI items are included.

Overall, this suggests that the two models learn from different signals in the data, giving different weights to different groups across horizons. This is in line with the two models having very different underlying optimisation algorithms, with the Ridge being a dense linear model and the Random Forest a hierarchical nonlinear universal function approximator. Given our modelling environment with relatively small T compared to a large M, we do not expect both models to have the same generalisation properties, i.e. having learned from the same signals. On the one hand, the stable Shapley value share attribution of the Ridge regression suggests an advantage in terms of robustness of models interpretations. On the other hand, the changes in the comparative importance of different groups over horizons, as suggested by the Random Forest, provides relevant information in itself and reflects the higher complexity and non-linearity of the model. Furthermore, these contributions may differ in their informativeness for different inflation regimes. For instance, the higher variance of shares of the Random Forest may make it useful to provide information about turning points, such as rising or falling inflation.

To assess this, and to also check whether the two models give different *relative* importance to certain sub-groups, we next look at relative Shapley weights for rising and falling headline inflation. This is shown in Figure 4. Again, the two models are shown along rows with different inflation regimes (all, rising, falling) along the columns.

Naïvely one may expect that the sub-groups to be attributed model weights proportional to their fraction of the input space assuming that the items in each group have similar information content. However, this is not the case as we can see in Figure 4. Both model over-proportionally rely on more volatile sub-groups food & beverages, non-core services and energy.[19] While the relative sub-group attributions of the Ridge are more volatile than its absolute shares across horizons, this effect is much stronger still for the Forest, where energy and volatile-service items can receive multiple times their input weights in terms of Shapley value shares. This is again in line with the Forest being the more flexible models. While potentially delivering useful signals, high relative shares should also be compared to the high absolute shares, which do provide the main forecasting contributions from Figure 3. For instance, energy and volatile services make up about

[19]It is reassuring that the importance of energy drops for core inflation measures given that these measure do not include those.

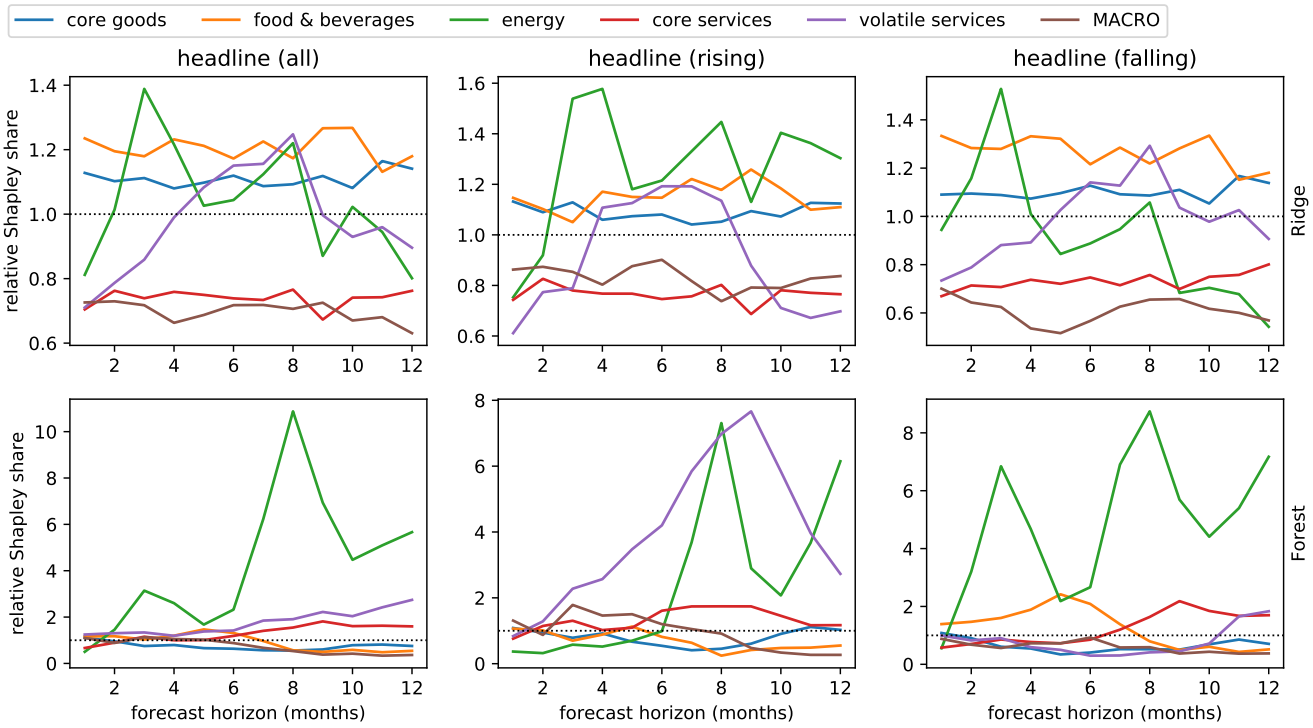

Figure 4: Relative Shapley weights of sub-groups (colour) by forecast horizon for headline inflation.
Notes: Rows show the Ridge regression (upper) and the Random Forest (lower). Columns different macroeconomic inflation regimes: all (left), rising (middle), and falling (right).

1.6% of items each, albeit energy with an over-proportional weight in the consumption basket. Whereas food & beverages is a major sub-group, especially for headline inflation, where it constitutes about 24% of items on average, albeit with an underproportial consumption weight.

More generally, we do not find an evident systematic relation between relative Shapley weights and forecast performance, suggesting that idiosyncratic factors are important, i.e. the current point in time and the particular drivers of changes in certain consumer price groups. The Shapley value decomposition and analyses presented here can form the basis for such assessments.

6 Conclusion

We have conducted a forecasting exercise aimed at predicting UK inflation using a unique and granular set of monthly CPI item series, as well as a more standard 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: dimension reduction techniques, shrinkage methods, and non-linear machine learning tools.

We have shown that, while it is difficult to beat the AR benchmark over the entire sample period, a wide range of forecast gains are achieved when evaluating forecasts over

sub-periods where inflation was rising, falling, high or low. Exploiting a large set of predictors substantially helps to forecast inflation during turning points when inflation dynamics are changing or inflation outturns fall into tails. In this, there is not one single model that performs best across segments and horizons, and so it is advisable to consider a wide range of models. When inflation is rising or falling, shrinkage methods perform best at horizons of 6-12 months when combined with CPI items, whereas machine learning methods tend to be stronger when also fed with macroeconomic indicators and for shorter horizons when inflation is rising. For the current environment of rising and high inflation, the results over segments imply that our approach can be particularly useful for forecasting headline and core CPI inflation with shrinkage or machine learning methods, whereas service inflation can be difficult to forecast when it is rising. These findings are in line with recent literature on forecasting inflation using machine learning and non-parametric methods, according to which a range of models can provide forecast gains, but in general non-linearities help forecasting inflation during business cycle turning points or rare events (Clark et al., 2022; Hauzenberger et al., 2022). Our analysis shows the potential of combining such non-parametric and non-linear methods with disaggregated price data to forecast aggregate inflation, and that this provides gains in particular in periods of changing inflation momentum and periods of high or low inflation.

Beyond the gains in forecast accuracy, forecasts derived from item series can help researchers and policy makers to interpret and communicate adjustments to forecasts based on dynamics observed across sub-groups of items and economic sectors. Item-level series connect individual prices at the product level with the macroeconomic CPI inflation concepts policy makers and economists are ultimately interested in. The large dimension of the input space, the volatility of individual CPI items, and the opacity of some of the models that can deal with such large data also pose challenges for the interpretation of forecasting results. We have addressed these through the Shapley value framework which derives linear contributions to the forecast from groups of items along CPI divisions. It thus represents a universal way to understand and communicate forecast results within economic policy settings. Here we find differences between more conventional linear models, like the Ridge regression, and more flexible machine learning models, like the Random Forest. The former tends to allocate stable shares of model output weights to different item sub-groups across specifications, while the latter shows considerable variance. Both give over-proportional weight to items belonging to more volatile groups.

References

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

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

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

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

Buckmann, M. and A. Joseph (2022). An interpretable machine learning workflow with an application to economic forecasting. *International Journal of Central Banking* (forthcoming).

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

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

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

Clark, T. E., F. Huber, G. Koop, and M. Marcellino (2022). Forecasting us inflation using bayesian nonparametric models. *arXiv preprint arXiv:2202.13793*.

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

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

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

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

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

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

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

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

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

Hauzenberger, N., F. Huber, and K. Klieber (2022). Real-time inflation forecasting using non-linear dimension reduction techniques. *International Journal of Forecasting*.

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

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

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

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

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

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

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

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

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

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

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

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

Odendahl, F., B. Rossi, and T. Sekhposyan (2022). Evaluating forecast performance with state dependence. *Journal of Econometrics*.

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

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

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

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

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

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

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

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

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

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

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

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

Strumbelj, E. and I. Kononenko (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 (Methodological)*, 267–288.

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

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

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

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

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

A Forecasting Models

Dimensionality reduction techniques

Principal Component Analysis (PCA) and regression is widely used in the forecasting literature, having been introduced by Stock and Watson (2002c). The indicator series x_t are summarized by a joint principal component, or factor, f_t. In a first step, the first principal component from the set of indicator series. In a second step, the factor is included in forecasting regression as follows

$$\hat{y}_{t+h} = \hat{\alpha} + \Sigma_{j=1}^{p} \hat{\beta} f_t + \Sigma_{j=1}^{p} \hat{\gamma}_j y_{t-j+1}. \tag{9}$$

The key idea is that a small number of principal components suffices to explain most of the variability in the data, and that these components also hold the bulk of predictive power for the target variable y_{t+h}. We set the number of principal components to $r = 5$ in the specifications where we include CPI item series and to $r = 3$ when we include macroeconomic predictors only.[A1]

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

Shrinkage methods

Ridge Regression. is a shrinkage method that penalises the residual sum of squares with the sum of squared coefficients (L2-norm). This shrinks the coefficients of those predictors with a minor contribution in terms of predictive ability of the model towards zero, albeit they never become exactly zero. As such, the Ridge regression is a dense modelling technique—it uses the full range of predictors, although assuming that the

[A1] The Bai and Ng (2002) selection criteria suggested a high numbers of principal components with a high explained variance share. Since this does not correspond to the goal of dimension reduction, we instead select the number of factors equal to the lowest number of factors which explains 50% of the variance in the data.

contribution of many of them might be small. Under our framework, the optimisation problem can be written as:

$$\hat{\beta}^{Ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i}^{T} (y_i - \alpha - \sum_{j}^{N} \beta z_{ij})^2 + \lambda \sum_{j}^{N} \beta_j^2 \right\} \tag{10}$$

for given values of α and $\lambda \geq 0$. It is common practice to centre the values of predictors around the mean first, and not to include the constant term.[A2] The parameter λ stands for the penalty imposed on coefficients and controls its overall magnitude. We have $\hat{\beta}^{Ridge} \to \hat{\beta}^{OLS}$ as $\lambda \to 0$ which is the no penalty case, and $\hat{\beta}^{Ridge} \to 0$ as $\lambda \to \infty$. Selecting a good value for the tuning parameter λ is crucial and is done via cross-validation.

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

$$\hat{\beta}^{Lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i}^{T} (y_i - a - \sum_{j}^{N} \beta z_{ij})^2 + \lambda \sum_{j}^{N} |\beta_j| \right\} \tag{11}$$

As such, the Lasso is a sparse modelling technique which performs shrinkage in terms of variable selection; it, thus, tends to give more parsimonious models compared to the Ridge. Again, the values of the parameters are centred, the constant term is excluded, and cross-validation is employed for the selection of the tuning parameter λ.[A3]

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

[A2] The reason for this is that the Ridge regression coefficients estimates can substantially change when multiplying a given predictor by a constant, due to the sum of squared coefficients term in the penalty part of the objective function.

[A3] The L1-Lasso-penalty makes the solutions nonlinear in the y_i's, and there is no closed form, unlike for the Ridge regression. However, there are efficient algorithms for computing the entire path of solutions as λ varies. For example, Least Angle Regression (LARs, Efron et al. (2004)) provides an efficient algorithm for computing the Lasso estimates.

$$\hat{\beta}^{n-EN} = \min_{\beta}\{\sum_{i}^{T}(y_i - a - \sum_{j}^{N} z_{ij}\beta)^2 + \lambda_1 \sum_{j}^{N} \beta_j^2 + \lambda_2 \sum_{j}^{N} |\beta_j|\} \tag{12}$$

The naïve version of Elastic Net method finds an estimator in a two-stage procedure: First, for each fixed λ_2 it finds the ridge regression coefficients, and then a Lasso-type shrinkage is applied. This kind of estimation incurs a double amount of shrinkage which leads to increased bias and poor predictions. However, using the correction factor $1 + \lambda_2$ the prediction performance is improved and the elastic net estimators are given by $\hat{\beta}^{EN} = (1 + \lambda_2)\hat{\beta}^{n-EN}$.

Non-Linear Machine Learning Models

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

$$y_{t+h} = \sum_{m=1}^{N_{leaf}} \hat{\beta}_m I(z_t \in Z_m) + \varepsilon_t, \quad \text{with} \quad \hat{\beta}_m = 1/|Z_m| \sum_{y^{tr} \in Z_m} y_{t+h}^{tr},\, m \in \{1, \dots, N_{leaf}\}\,. \tag{13}$$

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

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

Artificial neural networks (ANN) are similar to linear and non-linear least squares regressions and can be viewed as an alternative statistical approach to solving the least squares problem. A standard architecture of ANNs are multilayer perceptrons (MLP), a

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

$$y_{t+h} = g(z_t, W) + \varepsilon_t = \sum_{k=0}^{N_h} \hat{\beta}_k \, g_L(g_{L-1}(g_{L-2}(\ldots g_1(z_t, W_1), \ldots, W_{L-2}), \beta_{L-1}), W_L)_k + \varepsilon_t \tag{14}$$

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

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

$$y_{t+h} = \hat{\alpha}_0 + \sum_{i=1}^{m} \hat{\alpha}_i \mathcal{K}\big(z_i^{tr}, z_t\big) + \varepsilon_t \,, \tag{15}$$

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

B Additional Tables and Figures

Table B1: Macroeconomic Series used as Predictors

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

Notes: Sources are the Office for National Statistics (ONS), the Bank of England database (BOE). Transformation codes: LD = log year-on-year difference, L = levels, D = year-on-year difference. Category (Cat) codes: real = real activity, hp = house prices, fin = financial, price = prices.

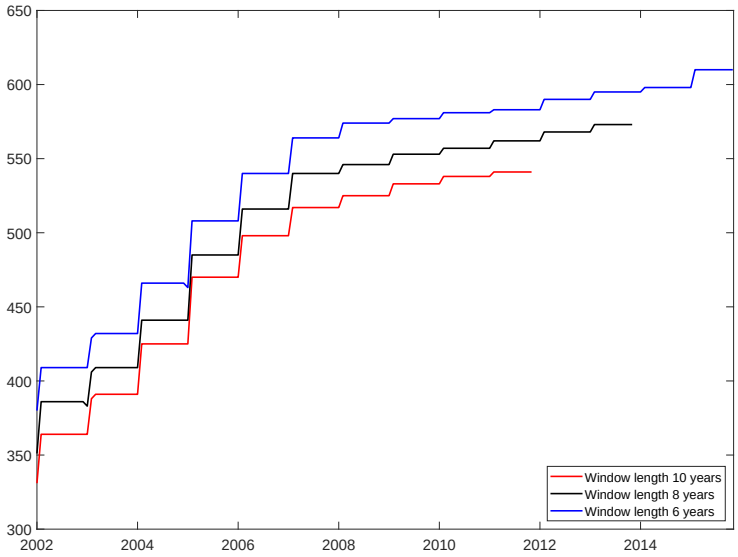

Figure B1: Evolving sample of CPI items for different rolling window sizes. Notes: Items are included if they fully cover a 6-, 8-, or 10-year rolling windows at each point in time. A shorter window allows for the inclusion of a larger set of items, but gives a shorter training period for estimation.

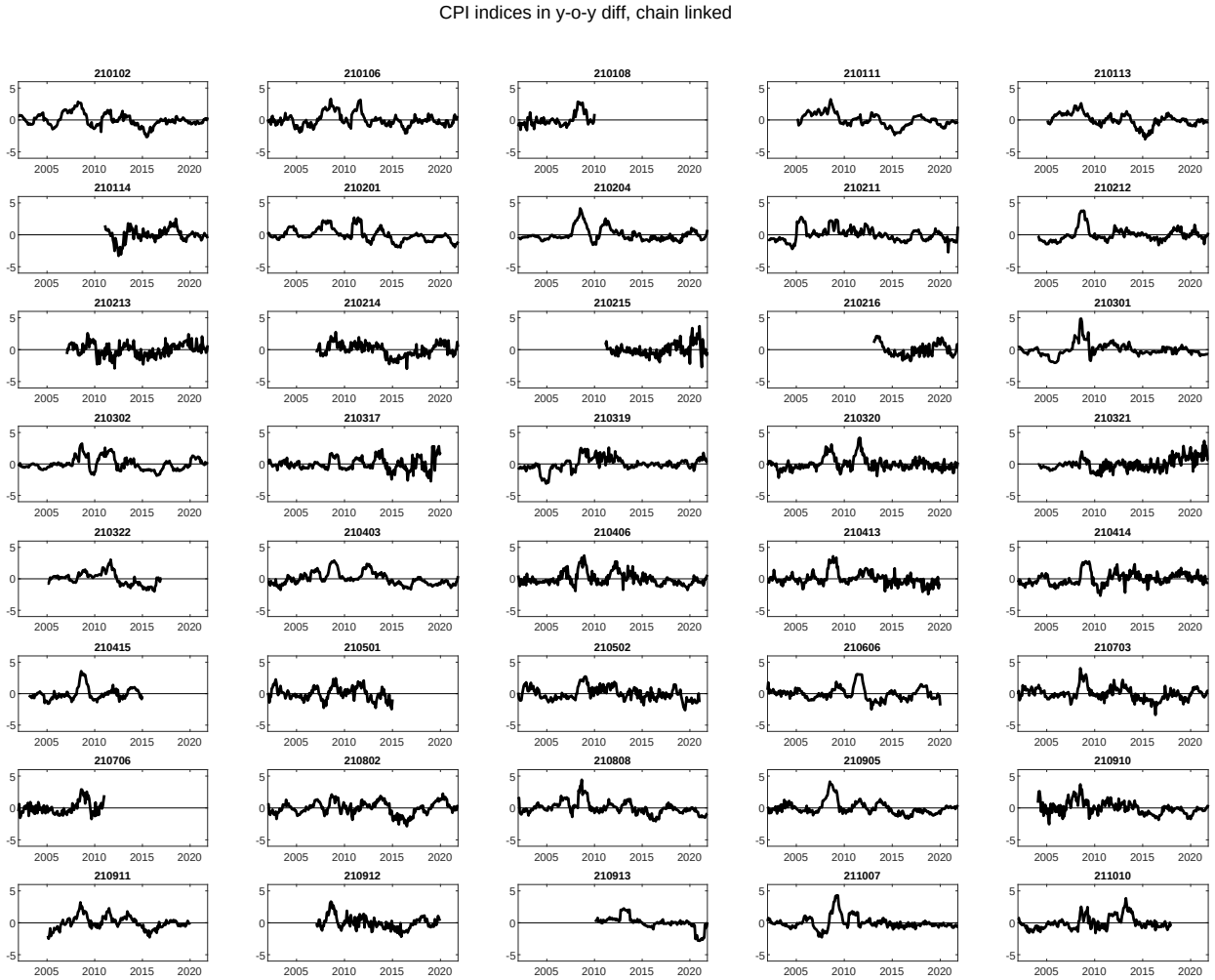

Figure B2: Selected item series. Notes: Data in year-on-year growth rates, standardised. Item identifiers No. 210102 to No. 211010. Series that do not cover at least the estimation window length of 8 years are dropped from the sample. Discontinued series only enter the estimation for the estimation windows that they cover in full.